Nº 12-001-X au catalogue

Techniques d'enquête

Juin 2008

Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 1-800-263-1136).

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.ca. Vous pouvez également communiquer avec nous par courriel à infostats@statcan.ca ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :	
Service de renseignements	1-800-263-1136
Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
Télécopieur	1-877-287-4369
Appels locaux ou internationaux :	
Service de renseignements	1-613-951-8116
Télécopieur	1-613-951-0581

Programme des services de dépôt

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

Comment accéder à ce produit ou le commander

Le produit nº 12-001-XIF au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique « Publications » > « Publications Internet gratuites ».

Ce produit nº 12-001-XPF au catalogue est aussi disponible en version imprimée standard au prix de 30 $CAN l'exemplaire et de 58 $CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 $CAN	12 $CAN
Autres pays	10 $CAN	20 $CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.ca
- Poste Statistique Canada
Finances
Immeuble R.-H.-Coats, 6e étage
150, promenade Tunney's Pasture
Ottawa (Ontario) K1A 0T6
- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

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.ca sous « À propos de nous » > « Offrir des services aux Canadiens ».

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.
∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.
∞

Statistique Canada
Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

Juin 2008

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2008

Tous droits réservés. Le contenu de la présente publication électronique peut être reproduit en tout ou en partie, et par quelque moyen que ce soit, sans autre permission de Statistique Canada, sous réserve que la reproduction soit effectuée uniquement à des fins d'étude privée, de recherche, de critique, de compte rendu ou en vue d'en préparer un résumé destiné aux journaux et/ou à des fins non commerciales. Statistique Canada doit être cité comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, il est interdit de reproduire le contenu de la présente publication, ou de l'emmagasiner dans un système d'extraction, ou de le transmettre sous quelque forme ou par quelque moyen que ce soit, reproduction électronique, mécanique, photographique, pour quelque fin que ce soit, sans l'autorisation écrite préalable des Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Juin 2008

N° 12-001-XIF au catalogue
ISSN 1712-5685

N° 12-001-XPB au catalogue
ISSN 0714-0045

Périodicité : semestriel

Ottawa

This publication is available in English upon request (catalogue no. 12-001-X)

Note de reconnaissance

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

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 34, numéro 1, juin 2008

Table des matières

Dans ce numéro 1

Articles Réguliers

Mary E. Thompson et Changbao Wu
Échantillonnage PPT systématique randomisé basé sur la simulation en cas de substitution d'unités 3

Mahmoud Torabi et J.N.K. Rao
Estimation pour petits domaines sous un modèle à deux niveaux 13

Yong You
Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions infraprovinciales au Canada 21

Junyuan Wang, Wayne A. Fuller et Yongming Qu
Estimation pour petits domaines sous une contrainte 33

Balgobin Nandram et Jai Won Choi
Une répartition bayésienne des électeurs indécis 41

Radu Lazar, Glen Meeden et David Nelson
Une approche bayésienne non informative de l'échantillonnage d'une population finie en utilisant des variables auxiliaires 55

Alan M. Zaslavsky, Hui Zheng et John Adams
Répartition optimale de l'échantillon pour une régression convergente sous le plan dans une enquête sur les services aux cancéreux quand les variables du plan d'échantillonnage sont connues pour des agrégats 71

Yan Li
Estimateurs par la régression généralisée d'un total de population finie utilisant la méthode de Box-Cox 87

Cédric Béguin et Beat Hulliger
L'algorithme BACON-EEM pour la détection d'observations aberrantes multivariées dans des données d'enquête incomplètes 99

Annette Jäckle et Peter Lynn
Offre de primes d'encouragement aux répondants dans une enquête par panel multimodes : effets cumulatifs sur la non-réponse et le biais 115

Leyla Mohadjer et Lester R. Curtin
Trouver l'équilibre entre les divers objectifs du plan d'échantillonnage de la National Health and Nutrition Examination Survey 131

Techniques d'enquête, juin 2008
Vol. 34, N° 1, pp. 1-2
Statistique Canada, N° 12-001-X au catalogue

Dans ce numéro

Ce numéro de *Techniques d'enquête* comprend des articles traitant de divers sujets de méthodologie.

Dans le premier article, Thompson et Wu considèrent le problème de l'obtention des probabilités d'inclusion pour le calcul des poids d'échantillonnage quand des modifications ou des compromis ont été apportés au plan d'échantillonnage original à cause de contraintes ou de limites pratiques. L'étude du problème a été motivée par l'enquête réalisée par la Chine dans le cadre du projet International Tobacco Control Policy Evaluation Survey en utilisant un plan d'échantillonnage à plusieurs degrés avec probabilités inégales pour la sélection de fumeurs et de non-fumeurs adultes dans sept villes. Parce que certains districts ont refusé de participer à l'enquête, des unités de substitution ont dû être sélectionnées après le tirage de l'échantillon original. Cette substitution a rendu difficile le calcul des probabilités d'inclusion de premier ordre et virtuellement impossible celui des probabilités de deuxième ordre. Dans l'article, les auteurs démontrent, théoriquement et empiriquement, que les probabilités d'inclusion de premier et de deuxième ordres peuvent être estimées avec précision au moyen de simulations Monte-Carlo.

Torabi et Rao calculent l'erreur quadratique moyenne (EQM) d'un nouvel estimateur par la régression généralisée (GREG) d'une moyenne de petit domaine sous un modèle à deux niveaux et comparent, théoriquement et empiriquement, l'efficacité relative du nouvel estimateur GREG proposé et de l'estimateur fondé sur le meilleur prédicteur linéaire sans biais.

Dans son article, You discute de l'utilisation de divers modèles transversaux et chronologiques pour petits domaines en vue d'estimer le taux de chômage dans des régions infraprovinciales au Canada. En particulier, il considère un modèle non linéaire intégré à effets mixtes sous un cadre hiérarchique bayésien (HB). Il utilise une approche HB avec la méthode d'échantillonnage de Gibbs pour obtenir des estimations des moyennes et des variances a posteriori des taux de chômage dans les petits domaines. Le modèle HB proposé produit des estimations fondées sur le modèle fiables si l'on s'en tient à la réduction du coefficient de variation. You analyse aussi la qualité de l'ajustement du modèle proposé et compare les estimations fondées sur le modèle aux estimations directes.

L'article de Wang, Fuller et Qu traite de l'estimation pour petits domaines sous une contrainte. Les auteurs étudient l'effet de divers modèles augmentés en ce qui concerne l'EQM et le prédicteur EBLUP. Ils considèrent des modèles pour petits domaines augmentés d'une variable explicative supplémentaire pour lesquels les prédicteurs pour petits domaines ordinaires présentent la propriété d'autocalage. En outre, ils considèrent des modèles pour petits domaines augmentés d'une variable auxiliaire supplémentaire qui est fonction de la taille du domaine afin de réduire le biais lorsqu'un modèle incorrect est utilisé aux fins de prédictions.

Nandram et Choi exposent une approche intéressante de répartition des électeurs indécis dans les sondages réalisés avant une élection. Les données des sondages électoraux sont habituellement présentées dans des tableaux de contingence à double entrée et de nombreux sondages sont réalisés avant l'élection proprement dite. Les auteurs décrivent la construction et l'analyse d'un modèle de non-réponse non ignorable variant en fonction du temps par des méthodes bayésiennes. Ils comparent leurs modèles à des versions élargies (afin d'inclure la variable de temps) des modèles de non-réponse ignorable et non ignorable introduits par Nandram, Cox et Choi (*Techniques d'enquête*, 2005). Ils construisent aussi un nouveau paramètre pour faciliter la prédiction du gagnant. Ils illustrent leur approche à l'aide de données de sondage recueillies durant la campagne pour l'élection du gouverneur de l'Ohio de 1998.

Dans leur article, Lazar, Meeden et Nelson élaborent une approche bayésienne de l'échantillonnage d'une population finie qui s'appuie sur l'utilisation d'une loi a posteriori de Pólya, quand l'information a priori est disponible sous la forme de données partielles sur une variable auxiliaire. Ils présentent l'estimateur de Pólya contraint et montrent qu'il possède des similarités avec l'estimateur par la régression généralisée sous échantillonnage aléatoire simple. Toutefois, leur estimateur ne requiert pas la spécification d'un modèle linéaire. Il est également apparenté aux méthodes fondées sur la vraisemblance empirique. Des exemples servent à illustrer la théorie.

Zaslavsky, Zheng et Adams considèrent les taux d'échantillonnage optimaux dans des plans d'échantillonnage par élément, quand l'analyse prévue est la régression linéaire pondérée par les poids de sondage et que les quantités à estimer sont des combinaisons linéaires des coefficients de régression provenant d'un ou de plusieurs modèles. Ils commencent par élaborer des méthodes en supposant que des renseignements exacts sur les variables du plan existent dans la base de sondage, puis ils les généralisent à des situations où l'information pour certaines variables du plan n'est disponible que sous forme d'agrégat pour des groupes de sujets éventuels ou provient de données inexactes ou périmées. Les applications éventuelles comprennent l'estimation des moyennes pour plusieurs ensembles de domaines chevauchants, l'estimation pour des sous-populations telles que les races minoritaires par échantillonnage non proportionnel des régions géographiques, ainsi que les études où les caractéristiques pour lesquelles des données sont disponibles dans les bases de sondage sont mesurées avec une erreur.

L'article de Li porte sur le problème de l'estimation d'un total de population finie à l'aide d'un estimateur par la régression généralisée non linéaire. L'auteur utilise la méthode de transformation de Box-Cox ainsi que l'estimation par la méthode du pseudo maximum de vraisemblance pour obtenir des prédictions dictées par les données. Il montre que l'estimateur par la régression résultant est convergent sous le plan et évalue également les propriétés de cet estimateur au moyen d'une étude par simulation.

Béguin et Hulliger étendent l'algorithme BACON au traitement de données d'enquête incomplètes. L'algorithme BACON a été élaboré pour repérer les observations aberrantes multivariées en utilisant la distance de Mahalanobis. Quand des valeurs manquent, l'algorithme EM peut être utilisé pour estimer la matrice de covariance à chaque étape d'itération de l'algorithme BACON. Afin de pouvoir traiter l'échantillonnage en population finie, les auteurs modifient l'algorithme EM et obtiennent une version qu'ils nomment algorithme EEM (espérance estimée/maximisation) qu'ils combinent à l'algorithme BACON. Ils obtiennent ainsi l'algorithme BACON-EEM qu'ils appliquent à deux ensembles de données et comparent à d'autres méthodes.

L'article de Jäckle et Lynn offre une évaluation empirique des effets des paiements incitatifs continus sur l'attrition, le biais de non-réponse et la non-réponse partielle, ainsi que de la variation éventuelle de ces effets d'une vague à l'autre d'une enquête par panel à mode d'interview mixte réalisée auprès des jeunes au Royaume-Uni. Ils testent plusieurs hypothèses quant aux effets des primes d'encouragement. Ils concluent que l'offre de ces primes est un moyen efficace de maintenir la taille d'échantillon d'un panel, donc de s'assurer de sa valeur en ce qui concerne l'efficacité, surtout pour les analyses de sous-groupes. Cependant, ils constatent aussi que les primes d'encouragement n'ont aucun effet sur le biais dû à l'attrition.

Enfin, Mohadjer et Curtin discutent des défis associés à la conception et à la mise en œuvre d'un processus d'échantillonnage permettant d'atteindre les objectifs de la National Health and Nutrition Examination Survey (NHANES). Ils décrivent comment le plan d'échantillonnage de la NHANES doit établir un juste équilibre entre les exigences liées à l'obtention d'échantillons de sous-domaines efficaces, d'une part, et d'une charge de travail pouvant être bien gérée par le personnel chargé des interviews et des examens dans les centres d'examen mobiles (CEM), d'autre part, tout en maintenant les taux de réponse aussi élevés et les coûts aussi faibles que possible. L'article décrit en détails plusieurs caractéristiques uniques du plan de sondage de la NHANES et se termine par un bref résumé des réalisations et de certains défis que posera dans l'avenir l'élaboration du plan de sondage de la NHANES.

Harold Mantel, Rédacteur en chef délégué

Techniques d'enquête, juin 2008
Vol. 34, N° 1, pp. 3-11
Statistique Canada, N° 12-001-X au catalogue

Échantillonnage PPT systématique randomisé basé sur la simulation en cas de substitution d'unités

Mary E. Thompson et Changbao Wu [1]

Résumé

L'enquête réalisée par la Chine dans le cadre du projet *International Tobacco Control* (*ITC*) *Policy Evaluation Survey* comprend un plan d'échantillonnage à plusieurs degrés avec probabilités inégales où les grappes du niveau supérieur sont sélectionnées par la méthode d'échantillonnage PPT systématique randomisé. Durant l'exécution de l'enquête, il faut résoudre le problème suivant : plusieurs grappes de niveau supérieur échantillonnées refusent de participer à l'enquête et doivent être remplacées par des unités de substitution sélectionnées parmi les unités non incluses dans l'échantillon initial, de nouveau par la méthode d'échantillonnage PPT systématique randomisé. Dans de telles conditions, les probabilités d'inclusion de premier ordre des unités finales sélectionnées sont très difficiles à calculer et la détermination des probabilités d'inclusion de deuxième ordre devient pratiquement impossible. Dans le présent article, nous élaborons une méthode fondée sur la simulation pour calculer les probabilités d'inclusion de premier et de deuxième ordre lorsque le calcul direct est prohibitif ou impossible. Nous démontrons l'efficacité et la faisabilité de la méthode que nous proposons en nous appuyant sur des considérations théoriques et des exemples numériques. Nous incluons plusieurs fonctions et codes R/S-PLUS pour la procédure proposée. La méthode peut être étendue à des situations de refus/substitution plus complexes susceptibles de survenir en pratique.

Mots clés : Probabilité d'inclusion; estimateur d'Horvitz-Thompson; méthode de Rao-Sampford; biais relatif; échantillonnage avec probabilités inégales sans remise.

1. Introduction

Dans l'analyse de données d'enquête complexe, la construction des poids de sondage est la première étape critique. Elle débute par le calcul des probabilités d'inclusion de premier ordre, qui est souvent simple si le plan d'échantillonnage original est bien exécuté, sans aucune altération ni modification. Par exemple, si les unités d'échantillonnage sont sélectionnées avec probabilité d'inclusion (π) proportionnelle à la taille (PPT ou πpt), les probabilités d'inclusion s'obtiennent facilement par un simple rééchelonnage de la variable de taille. Parmi les méthodes d'échantillonnage PPT avec probabilités inégales sans remise utilisables lorsque la taille d'échantillon est fixée arbitrairement, la méthode d'échantillonnage PPT systématique randomisé est la plus simple à appliquer. Elle a été décrite pour la première fois dans Goodman et Kish (1950) à titre de méthode de sélection contrôlée, et a été perfectionnée par Hartley et Rao (1962) qui ont étudié le problème important, mais difficile, du calcul des probabilités d'inclusion de deuxième ordre. Soit $x_i, i = 1, 2, \ldots, N$ les valeurs connues de la variable de taille, où N est le nombre total d'unités dans la population. Soit $z_i = x_i / X,$ où $X = \sum_{i=1}^{N} x_i,$ et supposons que $nz_i < 1$ pour tout i. La méthode d'échantillonnage PPT systématique randomisé procède comme suit. Arranger les N unités de la population dans un ordre aléatoire et poser que $A_0 = 0$ et $A_j = \sum_{i=1}^{j} (nz_i)$ sont les totaux cumulatifs des nz_i dans l'ordre en question, de sorte que $0 = A_0 < A_1 < \ldots < A_N = n.$ Soit u un nombre aléatoire uniforme dans l'intervalle [0, 1]. Les n unités qu'il faut inclure dans l'échantillon sont celles dont l'indice j satisfait $A_{j-1} \le u + k < A_j$ pour $k = 0, 1, \ldots, n-1.$ Soit s l'ensemble des n unités échantillonnées et $\pi_i = P(i \in s)$ les probabilités d'inclusion de premier ordre. La méthode d'échantillonnage PPT systématique randomisé satisfait la condition

$$\pi_i = nz_i, \ i = 1, 2, \ldots, N. \tag{1.1}$$

Plusieurs autres méthodes d'échantillonnage sans remise qui satisfont (1.1) pour une taille d'échantillon fixe arbitraire n ont également été proposées dans la littérature, y compris la méthode d'échantillonnage avec probabilités inégales bien connue de Rao-Sampford (Rao 1965; Sampford 1967) et celles de Chao (1982), Chen, Dempster et Liu (1994), Tillé (1996), ainsi que Deville et Tillé (1998), entre autres.

Les importants travaux de recherche sur les méthodes d'échantillonnage PPT ont été stimulés en grande partie par l'utilisation de l'estimateur d'Horvitz-Thompson (HT) $\hat{T} = \sum_{i \in s} y_i / \pi_i$ pour le total de population $T = \sum_{i=1}^{N} y_i$ d'une variable d'intérêt y. L'estimateur HT est très efficace quand y est fortement corrélé à la variable de taille x et que la méthode d'échantillonnage satisfait (1.1). Il s'agit de l'unique estimateur sans biais par rapport au plan parmi la classe d'estimateurs linéaires $\sum_{i \in s} w_i y_i$ pour T si les poids w_i dépendent uniquement de i.

1. Mary E. Thompson, Département de statistique et de science actuarielle, Université de Waterloo. Courriel : methomps@uwaterloo.ca; Changbao Wu, Département de statistique et de science actuarielle, Université de Waterloo. Courriel : cbwu@uwaterloo.ca.

Bien que le choix d'une méthode d'échantillonnage PPT puisse être souhaitable d'un point de vue théorique, l'exécution est souvent compliquée et parfois même impossible, en raison de contraintes et de limites pratiques. Certains compromis et modifications sont nécessaires. Toutefois, le plan modifié ne satisfait plus la condition (1.1). Le calcul direct des probabilités finales d'inclusion devient souvent difficile, voire impossible. Parmi les problèmes courants qui, dans la pratique des sondages, nécessitent l'altération du plan d'échantillonnage original, les plus fréquents sont les unités qui refusent de participer et la substitution d'unités. L'exemple qui suit illustre cette situation.

L'enquête réalisée par la Chine dans le cadre du projet *International Tobacco Control* (*ITC*) *Policy Evaluation Survey* (Enquête ITC Chine) repose sur un plan d'échantillonnage à plusieurs degrés avec probabilités inégales pour la sélection de fumeurs et de non-fumeurs adultes dans sept villes. Chaque ville possède une structure administrative hiérarchique naturelle

Ville → District de voirie → Îlot résidentiel → Ménage → Individu

qui a été intégrée dans le plan d'échantillonnage. Aux niveaux supérieurs, la méthode d'échantillonnage PPT systématique randomisé est appliquée pour sélectionner dix districts de voirie dans chaque ville, avec probabilité proportionnelle à la taille de la population du district, puis deux îlots résidentiels sont tirés dans chaque district sélectionné, de nouveau par échantillonnage PPT systématique randomisé avec probabilité proportionnelle à la taille de la population de l'îlot. Les ménages et les individus dans les ménages sont ensuite sélectionnés en utilisant une méthode d'échantillonnage aléatoire simple modifié. Le plan original consistait à sélectionner 40 fumeurs adultes et 10 non-fumeurs adultes dans chacun des 20 îlots résidentiels, afin d'obtenir un échantillon final de 800 fumeurs et de 200 non-fumeurs dans chaque ville.

Cependant, un problème s'est posé durant l'exécution de l'enquête : plusieurs grappes de niveau supérieur sélectionnées (d'abord les districts de voirie, puis les îlots résidentiels) ont refusé de participer à l'enquête, à cause de conflits avec d'autres activités ou de la non-disponibilité des ressources humaines. Ces grappes qui refusent doivent être remplacées par des unités de substitution sélectionnées parmi les unités non incluses dans l'échantillon initial; une possibilité consiste à recourir de nouveau à l'échantillonnage PPT systématique randomisé pour obtenir la taille d'échantillon global cible.

Sous les plans d'échantillonnage à plusieurs degrés tels que celui utilisé pour l'enquête ITC Chine, les probabilités d'inclusion de premier ordre des individus sélectionnés dans l'échantillon final peuvent être calculées en multipliant les probabilités d'inclusion des unités aux diverses étapes. Lorsque la méthode d'échantillonnage PPT systématique randomisé est modifiée à cause de la substitution d'unités à une certaine étape, la condition (1.1) n'est plus vérifiée pour l'échantillon final à cette étape. Le calcul des probabilités d'inclusion de premier ordre devient alors très difficile et celui des probabilités d'inclusion de deuxième ordre, virtuellement impossible. À l'annexe A, nous donnons une méthode de calcul direct (5.2) des π_i quand l'échantillon initial et l'échantillon de substitution sont sélectionnés tous deux par échantillonnage PPT systématique randomisé, en supposant que le refus de participer est aléatoire dans l'échantillon initial et qu'aucun refus n'a lieu dans l'échantillon de substitution. L'expression est valide conditionnellement au nombre de refus et à l'ordre de population utilisé (après randomisation) pour la sélection de l'échantillon initial. Il est évident que, même sous des conditions et hypothèses aussi contraignantes, le traitement de l'expression proprement dite devient difficile pour une taille d'échantillon qui n'est pas tellement grande.

Dans le présent article, nous démontrons, en nous appuyant sur des arguments théoriques et des exemples numériques, que les probabilités d'inclusion de premier et de deuxième ordre peuvent être estimées exactement par des simulations de Monte Carlo lorsque l'on dispose de renseignements complets sur le plan de sondage. Nos exemples numériques sont motivés par l'enquête ITC Chine pour laquelle l'échantillonnage PPT systématique randomisé sert de méthode de référence, mais nos résultats théoriques et la méthodologie générale s'appliquent aussi à d'autres méthodes d'échantillonnage avec probabilités inégales sans remise. À la section 2, nous présentons les résultats relatifs à l'exactitude des méthodes fondées sur la simulation. À la section 3, nous donnons des exemples numériques et des comparaisons. À l'annexe C, nous présentons plusieurs fonctions et codes R/S-PLUS pour la procédure proposée, qui ont été élaborés au départ pour l'enquête ITC Chine. À la section 4, nous formulons certaines remarques supplémentaires.

2. Propriétés des méthodes basées sur la simulation

Si le calcul des probabilités d'inclusion exactes est impossible ou prohibitif, mais que l'on dispose de renseignements complets sur le plan de sondage, il est facile d'utiliser des méthodes de simulation de Monte Carlo pour estimer les probabilités d'inclusion. Désignons le plan d'échantillonnage probabiliste entièrement spécifié par p. La méthode basée sur la simulation est simple, à savoir tirer K échantillons indépendants, tous selon le même plan

d'échantillonnage *p*; soit M_i le nombre d'échantillons qui contiennent l'unité *i*. Alors, la probabilité d'inclusion de premier ordre $\pi_i = P(i \in s)$ peut être estimée par $\pi_i^* = M_i / K$. Pour un *i* particulier, les M_i suivent une loi binomiale et les π_i^* satisfont $E(\pi_i^*) = \pi_i$ et $\mathrm{Var}(\pi_i^*) \le (4K)^{-1}$. Si, par exemple, nous pouvons nous permettre de prendre *K* aussi grand que 25×10^6, alors, $P(|\pi_i^* - \pi_i| < 0{,}001) \ge 0{,}99$ pour toute probabilité donnée π_i.

Une mesure plus pertinente de l'exactitude des méthodes basées sur la simulation est la performance de l'estimateur d'Horvitz-Thompson lorsqu'on utilise les probabilités d'inclussion simulées. Soit $\hat{T} = \sum_{i \in s} y_i / \pi_i$ et $\tilde{T} = \sum_{i \in s} y_i / \pi_i^*$. Pour un échantillon donné, le biais relatif dû à l'utilisation de $\tilde{T}$ au lieu de $\hat{T}$ est défini par $(\hat{T} - \tilde{T}) / \hat{T}$. Sans perte de généralité, nous supposons que $y_i \ge 0$ pour tout *i*. Nous montrons à l'annexe B que, pour tout $\varepsilon > 0$ et l'échantillon donné *s*,

$$P\left(\frac{|\hat{T} - \tilde{T}|}{\hat{T}} \le \varepsilon\right) \ge 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\sum_{i \in s} \frac{1}{\pi_i} - n\right). \qquad (2.1)$$

Comme $\sum_{i \in s} (1/\pi_i)$ est l'estimateur d'Horvitz-Thompson de la population de taille *N*, une borne inférieure pratique pour $P(|\hat{T} - \tilde{T}| / \hat{T} \le \varepsilon)$ avec un petit ε est donnée par

$$\Delta = 1 - \frac{2(N-n)}{K\varepsilon^2}. \qquad (2.2)$$

Si l'on veut que $\varepsilon = 0{,}01$ et $\Delta = 0{,}98$, alors pour $N - n = 100$ le nombre (théorique) d'échantillons indépendants requis pour la simulation est $K = 10^8$. Puisque la borne inférieure donnée par (2.1) est prudente, et valide pour toute variable de réponse, on s'attendrait à ce qu'un *K* plus petit ayant une valeur autour de 10^7, voire même 10^6, donne de bons résultats pour la plupart des scénarios pratiques où $N - n \le 100$, ce que corroborent les exemples numériques présentés à la section 3.

L'estimation des probabilités d'inclusion de deuxième ordre $\pi_{ij} = P(i, j \in s)$ n'impose aucune difficulté supplémentaire, excepté que le nombre total d'échantillons simulés, *K*, requis pour obtenir le même niveau d'exactitude relative que pour les probabilités de premier ordre est plus élevé. Soit M_{ij} le nombre d'échantillons simulés parmi les *K* échantillons indépendants qui contiennent à la fois *i* et *j*. Soit $\pi_{ij}^* = M_{ij} / K$ l'estimation pour π_{ij}. Supposons que le but soit d'estimer une grandeur de population quadratique

$$Q = \sum_{i=1}^{N} \sum_{j=1}^{N} q(y_i, y_j).$$

Les estimateurs de type Horvitz-Thompson de *Q* en utilisant π_{ij} ou π_{ij}^* sont donnés, respectivement, par

$$\hat{Q} = \sum_{i \in s} \sum_{j \in s} \frac{q(y_i, y_j)}{\pi_{ij}} \text{ et } \tilde{Q} = \sum_{i \in s} \sum_{j \in s} \frac{q(y_i, y_j)}{\pi_{ij}^*}.$$

En suivant le même argument que celui menant à (2.1), nous pouvons montrer que

$$P\left(\frac{|\hat{Q} - \tilde{Q}|}{\hat{Q}} \le \varepsilon\right) \ge 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\sum_{i \in s} \sum_{j \in s} \frac{1}{\pi_{ij}} - n^2\right). \qquad (2.3)$$

Comme $\sum_{i \in s} \sum_{j \in s} (1/\pi_{ij})$ est un estimateur de N^2 sans biais sous le plan, une borne inférieure pratique pour $P(|\hat{Q} - \tilde{Q}| / \hat{\mathrm{Q}} \le \varepsilon)$ est donnée par $1 - 2(N+n)(N-n)/(K\varepsilon^2)$. Si nous la comparons à Δ donnée par (2.2), il est évident que nous avons besoin d'un *K* beaucoup plus grand pour obtenir la même borne inférieure, quoique, dans les deux cas, la borne inférieure est prudente et la valeur de *K* effectivement requise peut être plus faible. Par ailleurs, les probabilités d'inclusion de deuxième ordre sont utilisées pour estimer les paramètres de deuxième ordre, tels que la variance de population ou la variance d'un estimateur linéaire. L'exactitude souhaitée de l'estimation est moins cruciale que dans le cas des paramètres de premier ordre, comme le total ou la moyenne de population et, par conséquent, un nombre compris entre 10^6 et 10^7 devrait être acceptable pour *K* dans de nombreuses situations pratiques.

L'aspect le plus critique pour les méthodes fondées sur la simulation est manifestement la faisabilité des calculs. Entre autres, elle dépend en grande partie de la valeur de *K* choisie, de la complexité du plan d'échantillonnage et de la puissance de calcul disponible. Si $K = 10^6$ et que l'on souhaite obtenir les résultats basés sur la simulation en dix heures, il faut tirer 28 échantillons simulés par seconde. L'échantillonnage PPT systématique randomisé est la méthode d'échantillonnage avec probabilités inégales sans remise la plus efficace en ce qui concerne l'exécution des calculs. Elle ne comporte qu'un classement aléatoire simple et la sélection d'un point de départ aléatoire. La plupart des méthodes concurrentes sont des méthodes d'échantillonnage réjectif ou des sélections séquentielles compliquées pour lesquelles le tirage d'échantillons simulés est beaucoup plus long. Nous comparons à la section 4 les temps de processeur requis pour calculer les π_i simulées pour l'échantillonnage PPT systématique randomisé et pour l'échantillonnage avec probabilités inégales de Rao-Sampford.

3. Exemples numériques

L'information sur le plan de sondage utilisée à la présente section est adaptée d'après l'enquête ITC Chine. Le nombre de districts de voirie (grappes de niveau supérieur) dans chacune des sept villes faisant partie du champ

d'observation de l'enquête varie de $N = 20$ à $N = 120$. Dans chaque ville, $n = 10$ districts sont sélectionnés par la méthode d'échantillonnage PPT systématique randomisé. En cas de refus, des districts de substitution sont sélectionnés parmi ceux non inclus dans l'échantillon initial, de nouveau par échantillonnage PPT systématique randomisé. Pour les besoins de l'illustration, nous utilisons l'information sur le plan de sondage provenant de la plus petite ville (c'est-à-dire $N = 20$). Des commentaires supplémentaires sur les cas pour lesquels N est grand sont formulés à la section 4.

3.1 Probabilités d'inclusion de premier ordre

Nous démontrons d'abord l'exactitude des valeurs simulées de π_i lorsque les valeurs exactes sont connues. Puis, nous examinons l'effet des substitutions d'unités sur la valeur finale de π_i et les propriétés de l'estimateur d'Horvitz-Thompson pour un total de population en utilisant les valeurs simulées de π_i. Les probabilités d'inclusion simulées en présence de substitution d'unités sont comparées à celles obtenues en supposant que le plan modifié est encore un échantillonnage PPT.

Exemple 1. π_i^* basées sur la simulation en l'absence de refus. Dans ce cas, les valeurs exactes de π_i sont données par $\pi_i = nz_i$.

i) Valeurs exactes de π_i :

0,5840 0,5547 0,6702 0,5331 0,3085 0,2652 0,3930 0,4180 0,6952 0,3471
0,5993 0,5393 0,8240 0,6868 0,4469 0,2191 0,4237 0,4180 0,7567 0,3163

ii) π_i^* simulées, $K = 10^5$:

0,5828 0,5545 0,6656 0,5339 0,3071 0,2656 0,3929 0,4205 0,6969 0,3474
0,6009 0,5429 0,8227 0,6865 0,4446 0,2186 0,4215 0,4179 0,7569 0,3194

iii) π_i^* simulées, $K = 10^6$:

0,5836 0,5558 0,6701 0,5336 0,3081 0,2654 0,3931 0,4180 0,6950 0,3469
0,5994 0,5394 0,8242 0,6864 0,4469 0,2186 0,4237 0,4172 0,7569 0,3166

Les π_i^* simulées concordent avec les π_i jusqu'à la deuxième décimale pour $K = 10^5$ et jusqu'à la troisième pour $K = 10^6$ dans la plupart des cas.

Exemple 2. Afin d'évaluer les propriétés de l'estimateur d'Horvitz-Thompson (HT) pour un total de population en utilisant les valeurs réelles de π_i et les π_i^* simulées de l'exemple 1, nous avons généré la variable de réponse d'après le modèle $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, N,$ où x_i est la variable de taille et les ε_i sont indépendants et suivent la même loi normale de moyenne 0 et de variance σ^2. Nous avons considéré trois populations (trois valeurs de σ^2) pour lesquelles les coefficients de corrélation de population entre x et y sont, respectivement, 0,3, 0,5 et 0,8. Pour chacune des trois populations, nous avons sélectionné $B = 2000$ échantillons répétés de taille $n = 10$ par échantillonnage PPT systématique randomisé, et pour chaque échantillon, nous avons calculé trois estimateurs HT en utilisant les valeurs réelles de π_i, les π_i^* simulées avec $K = 10^5$ et les π_i^* simulées avec $K = 10^6$, respectivement. Les résultats, qui ne sont pas présentés ici faute d'espace, montrent que le biais relatif de chacun des trois estimateurs HT est inférieur à 0,04 % et que leurs erreurs quadratiques moyennes sont presque identiques.

Exemple 3. Lorsque des refus de participation se produisent dans l'échantillon PPT initial et que des unités de substitution sont sélectionnées parmi les unités non incluses dans l'échantillon initial en utilisant la même méthode d'échantillonnage PPT, deux questions se posent : 1) comment calculer les probabilités d'inclusion π_i pour l'échantillon final et 2) dans quelle mesure la méthode de substitution a-t-elle altéré le plan d'échantillonnage PPT original? Nous pouvons calculer les π_i^* simulées et les comparer aux $\tilde{\pi}_i$ obtenues en supposant que l'échantillonnage demeure PPT après que l'on ait retiré de la base de sondage les unités ayant exprimé un refus. En simulant les π_i^*, nous supposons pour simplifier qu'aucun refus n'est possible parmi les unités non comprises dans l'échantillon initial et, par conséquent, qu'il n'existe aucun refus parmi les unités de substitution. Le nombre de répétitions K est fixé à 10^6 pour la simulation. Nous envisageons deux scénarios dans les conditions où la population contient trois unités qui refusent de participer et que toutes ces unités sont sélectionnées dans l'échantillon initial de taille $n = 10$.

i) Refus de trois grandes unités : π_i^* simulées (deux premières lignes) contre $\tilde{\pi}_i$ (deux dernières lignes) sous l'hypothèse d'un échantillonnage PPT.

0,7231 0,6981 0,7947 0,6773 0,4354 0,3811 0,5339 0,5619 0,0000 0,4815
0,7363 0,6826 0,0000 0,8070 0,5919 0,3210 0,5678 0,5615 0,0000 0,4441

0,7560 0,7182 0,8677 0,6901 0,3994 0,3434 0,5088 0,5412 0,0000 0,4494
0,7759 0,6983 0,0000 0,8892 0,5786 0,2837 0,5486 0,5412 0,0000 0,4096

ii) Refus de trois petites unités : π_i^* simulées (deux premières lignes) contre $\tilde{\pi}_i$ (deux dernières lignes) sous l'hypothèse d'un échantillonnage PPT.

0,6326 0,6049 0,7167 0,5829 0,0000 0,0000 0,4415 0,4668 0,7406 0,3937
0,6482 0,5901 0,8558 0,7330 0,4965 0,0000 0,4728 0,4664 0,7976 0,3590

0,6343 0,6025 0,7280 0,5790 0,0000 0,0000 0,4268 0,4540 0,7550 0,3770
0,6510 0,5858 0,8949 0,7459 0,4854 0,0000 0,4602 0,4540 0,8218 0,3436

Il est évident que la taille des unités qui refusent a un effet très important sur la distribution des probabilités d'inclusion finales. Si l'on omet de tenir compte de l'altération du plan d'échantillonnage résultant de la substitution des unités et que l'on traite ce plan comme s'il s'agissait encore d'un échantillonnage PPT, les probabilités d'inclusion pour les grandes unités sont exagérées et le rôle des petites unités est atténué. Cette tendance est plus prononcée lorsque de grandes unités figurent parmi celles qui refusent de participer, c'est-à-dire le cas (i) où $\pi_{14}^* = 0,8070$ comparativement à $\tilde{\pi}_{14} = 0,8897$ et $\pi_{16}^* = 0,3210$ comparativement à $\tilde{\pi}_{16} = 0,2837$.

3.2 Probabilités d'inclusion de deuxième ordre

L'échantillonnage PPT systématique randomisé a suscité de nombreux travaux de recherche, principalement en vue d'obtenir les probabilités d'inclusion de deuxième ordre π_{ij} et les estimateurs de variance. Hartley et Rao (1962) ont établi des formules exactes pour les π_{ij} quand $n = 2$ et $N = 3$ ou $N = 4$; Connor (1966) a étendu les résultats et dérivé la formule exacte pour les cas généraux *n* et *N*, et la méthode de calcul connexe a été implémentée plus tard dans le langage Fortran par Hidiroglou et Gray (1980). La procédure est assez lourde, comme en témoignent les 165 lignes de code Fortran.

Le résultat le plus curieux est probablement l'approximation asymptotique des π_{ij} calculées par Hartley et Rao (1962). Dans un papier récent, Kott (2005) a montré que l'estimateur de variance d'un estimateur d'Horvitz-Thompson fondé sur l'approximation de Hartley-Rao non seulement donne de bons résultats sous le cadre fondé sur le plan de sondage, mais a aussi de bonnes propriétés sous le modèle. L'approximation de Hartley-Rao a été dérivée au départ en partant de l'hypothèse que *n* est fixe et que *N* est grand et correct jusqu'à l'ordre $O(N^{-4})$ (Hartley et Rao 1962 : Équation (5.15) à la page 369). Lors d'une conversation privée avec J.N.K. Rao durant le 23e Symposium international sur les questions de méthodologie de Statistique Canada, celui-ci a fait remarquer que l'approximation est encore valide, même si n est grand, à condition que n/N soit faible. Dans les cas où *N* n'est pas grand et(ou) que n/N n'est pas faible, comme dans l'exemple de l'enquête ITC Chine considéré ici, aucune donnée n'a été publiée au sujet de la qualité de l'approximation de Hartley-Rao.

Quand la méthode d'échantillonnage PPT systématique randomisé est altérée par la substitution d'unités, il est virtuellement impossible de calculer les probabilités d'inclusion de deuxième ordre ou toute approximation. En revanche, selon l'approche basée sur la simulation, l'obtention d'estimations très fiables des π_{ij} à l'aide d'un grand nombre d'échantillons simulés reste simple, à condition que la procédure d'échantillonnage altérée soit entièrement spécifiée. Nous examinons dans la suite les propriétés des estimateurs de variance en utilisant les π_{ij}^* simulées lorsque la méthode d'échantillonnage PPT systématique randomisé n'est pas altérée. Dans ces conditions, $\pi_i = nz_i$ et l'approximation $\tilde{\pi}_{ij}$ de Hartley-Rao pour π_{ij} peut aussi être incluse dans la comparaison.

Exemple 4. Nous commençons par comparer π_{ij}^* à $\tilde{\pi}_{ij}$ pour chacune des entrées individuelles. Faute d'espace, nous présentons uniquement les résultats pour $i = 1, \ldots, 5$ et $j = 1, \ldots, 10$, qui suffisent à dépeindre le tableau général. L'approximation de Hartley-Rao $\tilde{\pi}_{ij}$ est très proche des π_{ij}^* simulées, la concordance allant jusqu'à la deuxième décimale pour la majorité des entrées. Il s'agit manifestement d'une observation intéressante, sachant que $N = 20$ et $n = 10$.

i) π_{ij}^* simulées, $K = 10^6$:

```
0,0000 0,3121 0,3821 0,2975 0,1669 0,1442 0,2116 0,2249 0,3975 0,1873
0,3121 0,0000 0,3623 0,2816 0,1590 0,1372 0,2025 0,2141 0,3766 0,1784
0,3821 0,3623 0,0000 0,3469 0,1899 0,1640 0,2483 0,2659 0,4586 0,2153
0,2975 0,2816 0,3469 0,0000 0,1523 0,1312 0,1938 0,2061 0,3606 0,1717
0,1669 0,1590 0,1899 0,1523 0,0000 0,0742 0,1124 0,1197 0,1968 0,0988
```

ii) Approximation de Hartley-Rao $\tilde{\pi}_{ij}$:

```
0,0000 0,3079 0,3769 0,2952 0,1668 0,1427 0,2143 0,2286 0,3921 0,1884
0,3079 0,0000 0,3569 0,2795 0,1579 0,1351 0,2029 0,2164 0,3712 0,1784
0,3769 0,3569 0,0000 0,3421 0,1932 0,1654 0,2484 0,2649 0,4544 0,2183
0,2952 0,2795 0,3421 0,0000 0,1514 0,1296 0,1946 0,2075 0,3559 0,1710
0,1668 0,1579 0,1932 0,1514 0,0000 0,0732 0,1099 0,1172 0,2010 0,0966
```

Exemple 5. Dans le cas des probabilités d'inclusion de deuxième ordre, l'aspect principal est l'estimation de la variance. Pour une taille d'échantillon fixe, un estimateur sans biais de la variance de l'estimateur d'Horvitz-Thompson $\hat{Y}_{\mathrm{HT}} = \sum_{i \in s} y_i / \pi_i$ est donné par la forme bien connue de Yates-Grundy,

$$v(\hat{Y}_{\mathrm{HT}}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 . \qquad (3.1)$$

Considérons les trois populations synthétiques décrites dans l'*exemple 2*. Nous obtenons la variance réelle $V = \mathrm{Var}(\hat{Y}_{\mathrm{HT}})$ par simulation en utilisant $B = 10^5$ échantillons simulés et en la calculant sous la forme $B^{-1} \sum_{b=1}^{B} (\hat{Y}_b - Y)^2$, où *Y* est le total de population réel et $\hat{Y}_b$ est l'estimateur d'Horvitz-Thompson de *Y* calculé d'après le *b*e échantillon simulé. Nous examinons trois estimateurs de la variance de la forme (3.1), désignés respectivement par v_1, v_2 et v_3, en remplaçant les π_{ij} dans (3.1) respectivement par l'approximation de Hartley-Rao $\tilde{\pi}_{ij}$, les π_{ij}^* simulées pour $K = 10^5$ et les π_{ij}^* pour $K = 10^6$. Nous évaluons la performance de ces estimateurs en nous appuyant sur le biais relatif simulé $\mathrm{BR} = B^{-1} \sum_{b=1}^{B} (v^{(b)} - V)/V$ et l'instabilité simulée $\mathrm{INST} = \{B^{-1} \sum_{b=1}^{B} (v^{(b)} - V)^2\}^{1/2} / V$, où $v^{(b)}$ est l'estimation de la variance calculée d'après le *b*e échantillon, en utilisant un autre ensemble de $B = 10^5$ échantillons indépendants. Les résultats sont résumés au tableau 1 ci-après. Les trois populations sont indiquées par le coefficient de corrélation ρ entre *y* et *x*.

Tableau 1 Biais relatif et instabilité des estimateurs de variance

	BR (%)			INST		
Population	v_1	v_2	v_3	v_1	v_2	v_3
$\rho = 0,30$	6,1 %	1,4 %	-0,3 %	0,66	0,65	0,65
$\rho = 0,50$	4,3 %	2,5 %	-1,1 %	0,42	0,44	0,42
$\rho = 0,80$	2,6 %	1,2 %	-0,2 %	0,61	0,60	0,60

En ce qui concerne le biais relatif, les trois estimateurs de variance sont acceptables, celui (v_1) basé sur

l'approximation de Hartley-Rao $\tilde{\pi}_{ij}$ ayant le biais le plus important. Dans le cas des estimateurs de variance utilisant les π^*_{ij} simulées, l'accroissement de la valeur de K de 10^5 (c'est-à-dire v_2) à 10^6 (c'est-à-dire v_3) rend le biais négligeable, quoique celui obtenu pour $K = 10^5$ soit clairement acceptable en pratique. Pour ce qui est de l'instabilité, les mesures sont comparables pour les trois versions de l'estimateur de la variance.

4. Quelques remarques supplémentaires

En théorie, la méthode de calcul des probabilités d'inclusion basée sur la simulation s'applique à n'importe quel plan d'échantillonnage, à condition que l'on dispose de renseignements complets sur ce dernier. Elle convient bien au traitement de scénarios de substitution plus complexes ou d'autres types de modification du plan original. Dans le cas de l'enquête ITC Chine, l'une des unités ayant refusé de participer à dû être remplacée par une unité provenant d'une région particulière de la ville, à cause de contraintes de charge de travail et de restrictions relatives au travail sur le terrain. Dans celui d'une enquête nationale auprès des jeunes réalisée au Canada, certaines unités ont refusé de participer au deuxième, ainsi qu'au troisième degré (écoles) et des unités de substitution ont donc été sélectionnées avant d'atteindre la taille d'échantillon cible. Comme l'a fait remarquer un rédacteur associé, une situation comparable a également été mentionnée lors du 57e cycle de l'enquête de la National Sample Survey Organization du gouvernement de l'Inde (www.mospi.gov.in), où une modification avait été apportée à l'échantillonnage systématique circulaire avec probabilité proportionnelle à la taille afin de sélectionner deux sous-échantillons distincts. Gray (1973) a décrit une méthode d'accroissement de la taille de l'échantillon (nombre d'UPE) quand l'échantillon initial a été sélectionné par la méthode d'échantillonnage PPT systématique randomisé. Le calcul des probabilités d'inclussion de deuxième ordre sous les méthodes proposées est difficile, même si l'échantillon est de très petite taille. Dans tous ces cas, les solutions analytiques pour les probabilités d'inclussion sont difficiles à utiliser ou n'existent pas, mais l'approche fondée sur la simulation peut être appliquée sans trop de difficulté.

Dans un article récent, Fattorini (2006) discute de l'utilisation de la méthode basée sur la simulation dans le cas de l'échantillonnage spatial où les unités sont sélectionnées séquentiellement. Quand un plan d'échantillonnage PPT est altéré par une ou plusieurs séries de substitutions d'unités, le plan modifié peut également être considéré comme séquentiel. Toutefois, nos résultats théoriques sur l'exactitude des méthodes basées sur la simulation diffèrent de ceux de Fattorini. Nous avons utilisé un argument conditionnel et proposé d'évaluer les propriétés de l'estimateur en utilisant les probabilités d'inclusion simulées pour un échantillon donné présentant un intérêt en pratique.

La question fondamentale que soulèvent les méthodes basées sur la simulation est celle de la faisabilité des calculs. L'échantillonnage PPT systématique randomisé présente un énorme avantage pour ce qui est de l'efficacité des calculs. La méthode d'échantillonnage avec probabilités inégales de Rao-Sampford (Rao 1965; Sampford 1967), par exemple, est une autre méthode d'échantillonnage PPT d'usage répandu. Elle possède plusieurs attributs désirables, comme des expressions explicites pour les probabilités d'inclusion de deuxième ordre, et est plus efficace que l'échantillonnage PPT systématique randomisé (Asok et Sukhatme 1976). Suit une comparaison des temps de processeur sous échantillonnage PPT systématique randomisé et sous-échantillonnage PPT de Rao-Sampford pour la simulation des probabilités d'inclusion de premier ordre. La taille d'échantillon est fixée à $n = 10$ et le nombre d'échantillons simulés est $K = 10^6$. Les résultats sont obtenus en utilisant R sur une machine unix à processeur double.

N	PPT systématique	PPT de Rao-Sampford
200	4,7 heures	7,5 heures
100	2,5 heures	5,0 heures
50	1,6 heure	4,4 heures
20	1,2 heure	8,9 heures

Il convient de souligner que, si la méthode de Rao-Sampford demande généralement plus de temps pour obtenir les résultats, elle en prend nettement plus dans le cas où $N = 20$. Cela tient au fait que la méthode de Rao-Sampford s'appuie sur une procédure réjective et qu'il faut habituellement un grand nombre de rejets pour arriver à un échantillon final lorsque la fraction d'échantillonnage n/N est grande. L'échantillonnage PPT systématique randomisé, quant à lui, n'est pas affecté par cette situation, et la méthode basée sur la simulation peut fournir rapidement des résultats ayant l'exactitude souhaitée pour $N = 400$, voire une valeur plus élevée. Plusieurs fonctions R/S-PLUS et les principaux codes pour l'approche proposée sont présentés à l'annexe C et sont applicables à d'autres scénarios de substitution moyennant de légères modifications.

L'une des raisons de l'utilisation de l'échantillonnage PPT systématique randomisé pour sélectionner les grappes de niveau supérieur dans l'enquête ITC Chine est que le plan final est autopondéré. Un problème se pose en cas de refus de participer. Comment faut-il sélectionner les unités de substitution de façon que le plan d'échantillonnage modifié final soit encore (approximativement) autopondéré? Dans d'autres circonstances, comme l'échantillonnage avec renouvellement, cela est réalisable; consulter, par exemple, Fellegi (1963). La façon d'atteindre cet objectif dans le cas du plan de l'enquête ITC Chine est à l'étude.

Remerciements

Cette étude a été financée en partie par des subventions du Conseil de recherches en sciences naturelles et en génie du Canada. Les auteurs remercient aussi de leur appui les responsables du projet *International Tobacco Control* (*ITC*) *Policy Evaluation* et du projet de l'enquête ITC Chine. Le projet ITC est financé en partie par des subventions du Roswell Park Transdisciplinary Tobacco Use Research Center du National Cancer Institute des États-Unis (P50 CA11236) et des Instituts de recherche en santé du Canada (57897). Le financement du projet ITC Chine est assuré par le ministère de la Santé et le ministère des Finances de la Chine.

Annexe A

Un calcul direct sous refus aléatoire

Sous le plan d'échantillonnage PPT systématique randomisé et en supposant que le refus est aléatoire, il est possible, en principe, de calculer directement les probabilités d'inclusion sous une règle de substitution. Le point de départ consiste à énumérer tous les échantillons initiaux possibles et leurs probabilités en se basant sur l'ordre de population particulier utilisé pour sélectionner l'échantillon initial.

Rappelons que $A_0 = 0, A_j = \sum_{i=1}^{j} (nz_i)$ et $A_N = n$. Pour une valeur de départ uniforme choisie $u \in [0, 1]$, l'unité *j* est sélectionnée si

$$A_{j-1} \le u + k < A_j \qquad (5.1)$$

pour une valeur de $k = 0, 1, \ldots, n-1$. Soit k_j le plus grand nombre entier inférieur à A_j, et soit le reste e_j donné par $e_j = A_j - k_j$. Soit $0 < e_{(1)} \le e_{(2)} \le \ldots \le e_{(N)}$ les statistiques d'ordre des restes, et soit $k_{(1)}, \ldots, k_{(N)}$ les valeurs correspondantes de k_j. Notons que $e_{(N)} = 1$. Nous pourrions générer *N* échantillons possibles $s_1, \ldots, s_N$ avec les probabilités respectives

$$e_{(1)}, e_{(2)} - e_{(1)}, \ldots, e_{(N)} - e_{(N-1)},$$

dont certaines pourraient être égales à 0. Nous commençons par générer s_1. Partant de chaque unité $j = 1, \ldots, N$, plaçons j dans s_1 si $A_{j-1} \le k < A_j$ pour une valeur $k = 0, 1, \ldots, n-1$, c'est-à-dire que s_1 est sélectionné en utilisant $u = 0$ dans (5.1). À mesure que nous faisons varier u de 0 à 1, nous pouvons identifier séquentiellement différents échantillons possibles. Maintenant, sachant $s_1, \ldots, s_m$, supposons que s_{m+1} est identique à s_m excepté que le $(k_{(m)}+1)^{\text{e}}$ élément est avancé d'une valeur 1. Par exemple, supposons que $n = 4$ et $s_m = \{1, 3, 6, 9\}$, et supposons que $k_{(m)} = 0$, alors $s_{m+1} = \{2, 3, 6, 9\}$. Par ailleurs, si $k_{(m)} = 2$, alors $s_{m+1} = \{1, 3, 7, 9\}$. L'échantillon s_{m+1} aura la probabilité $e_{(m+1)} - e_{(m)}$.

Par construction, $\pi_i = nz_i$ pour $i = 1, \ldots, N$. Si l'on souhaite calculer les probabilités d'inclusion de premier et de deuxième ordre seulement, on peut utiliser un algorithme similaire mais plus simple pour calculer directement les probabilités d'inclusion de deuxième ordre, sachant l'ordre initial. Cependant, dans les applications où les probabilités de tous les échantillons sont nécessaires, l'algorithme de génération d'échantillons peut être exécuté. Ainsi, pour les populations de petite taille, il est alors également possible de calculer les probabilités d'inclusion de premier ordre en cas de refus et de substitution. Supposons que nous sélectionnions pour commencer un échantillon de taille *n* par échantillonnage PPT systématique randomisé. Supposons que n_1 unités de cet échantillon consentent à répondre et qu'un nombre $n_2 = n - n_1$ d'unités supplémentaires soient sélectionnées, de nouveau par échantillonnage PPT systématique randomisé à partir des unités n'ayant pas été échantillonnées la première fois. Supposons pour simplifier que les cas de refus dans le premier échantillon surviennent au hasard et qu'il n'y en a aucun dans l'échantillon de substitution. Notons que cette hypothèse diffère de celle utilisée dans l'exemple 3, où l'ensemble des cas de refus est considéré comme étant non aléatoire. La probabilité d'inclusion de l'unité *i*, sachant l'ordre de population initial supposé, est

$$nz_i \times \frac{n_1}{n} + \sum_{m: i \notin s_m} p_1(s_m) \frac{n_2 z_i}{\sum_{j: j \notin s_m} z_j}. \qquad (5.2)$$

La somme externe est calculée sur l'ensemble des échantillons s_m de taille *n*, générés conformément au processus décrit plus haut, mais sans avoir d'unité *i*, avec les probabilités $p_1(s_m) = e_{(m)} - e_{(m-1)}$. La somme interne qui figure au dénominateur est calculée sur l'ensemble des unités *j* non incluses dans les échantillons s_m de la somme externe. Les probabilités d'inclusion inconditionnelles peuvent être obtenues par calcul de la moyenne approprié sur l'ensemble des ordres de population qui produisent des valeurs distinctes. Cela n'est manifestement faisable que si la population est petite ou que *z* prend un petit nombre de valeurs.

Annexe B

Obtention de l'expression (2.1)

Nous montrons ici que, pour tout $\varepsilon > 0$ et un échantillon donné *s*,

$$P\left(\frac{|\hat{T} - \tilde{T}|}{\hat{T}} \le \varepsilon \right) \ge 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2} \left(\sum_{i \in s} \frac{1}{\pi_i} - n \right),$$

où $\hat{T} = \sum_{i \in s} y_i / \pi_i, \tilde{T} = \sum_{i \in s} y_i / \pi_i^*$, et les π_i^* sont les probabilités d'inclusion de premier ordre simulées à partir de K échantillons indépendants. En notant que $E(\pi_i^*) = \pi_i$ et $\mathrm{Var}(\pi_i^*) = \pi_i(1-\pi_i)/K$, en vertu de l'inégalité de Chebyshev, nous avons $P(|\pi_i^* - \pi_i| > c) \le \pi_i(1-\pi_i)/(Kc^2)$ pour tout $c > 0$. Il s'ensuit que :

$$
\begin{aligned}
&P\left(\frac{|\pi_i^* - \pi_i|}{\pi_i^*} > \varepsilon\right) \\
&= P(\pi_i^* - \pi_i > \pi_i^* \varepsilon) + P(\pi_i^* - \pi_i < -\pi_i^* \varepsilon) \\
&= P(\pi_i^* - \pi_i > \varepsilon \pi_i/(1-\varepsilon)) + P(\pi_i^* - \pi_i < -\varepsilon \pi_i/(1+\varepsilon)) \\
&\le P(|\pi_i^* - \pi_i| > \varepsilon \pi_i/(1-\varepsilon)) + P(|\pi_i^* - \pi_i| > \varepsilon \pi_i/(1+\varepsilon)) \\
&\le \frac{(1-\varepsilon)^2 \pi_i(1-\pi_i)}{K\varepsilon^2 \pi_i^2} + \frac{(1+\varepsilon)^2 \pi_i(1-\pi_i)}{K\varepsilon^2 \pi_i^2} \\
&= \frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\frac{1}{\pi_i} - 1\right).
\end{aligned}
$$

Si $y_i \ge 0$ pour tout i, alors

$$|\hat{T} - \tilde{T}| \le \sum_{i \in s} \frac{y_i}{\pi_i} \frac{|\pi_i^* - \pi_i|}{\pi_i^*} \le \max_{i \in s}\left\{\frac{|\pi_i^* - \pi_i|}{\pi_i^*}\right\} \hat{T}.$$

Pour tout $\varepsilon > 0$ et l'échantillon donné s,

$$
\begin{aligned}
P\left(\frac{|\hat{T} - \tilde{T}|}{\hat{T}} \le \varepsilon\right) &\ge P\left(\max_{i \in s}\left\{\frac{|\pi_i^* - \pi_i|}{\pi_i^*}\right\} \le \varepsilon\right) \\
&\ge 1 - \sum_{i \in s} P\left(\frac{|\pi_i^* - \pi_i|}{\pi_i^*} > \varepsilon\right) \\
&\ge 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2}\left(\sum_{i \in s} \frac{1}{\pi_i} - n\right).
\end{aligned}
$$

Annexe C

Mise en œuvre en R/S-PLUS

C1. Une fonction R pour l'échantillonnage PPT systématique randomisé.

Les variables d'entrée de la fonction sont `x` : le vecteur de population de variables de taille et `n` : la taille de l'échantillon. La fonction `syspps` donne l'ensemble de n unités sélectionnées.

```
syspps<-function(x,n){
N<-length(x)
U<-sample(N,N)
xx<-x[U]
z<-rep(0,N)
for(i in 1:N) z[i]<-n*sum(xx[1:i])/sum(x)
r<-runif(1)
s<-numeric()
for(i in 1:N){
if(z[i]>=r){
s<-c(s,U[i])
r<-r+1
 } }
return(s[order(s)])
}
```

C2. Une fonction R pour la simulation des probabilités d'inclusion de deuxième ordre.

Les variables d'entrée de la fonction sont `x`: le vecteur de population des variables de taille et `s`: l'ensemble d'étiquettes des unités dans l'échantillon. La méthode d'échantillonnage par défaut est l'échantillonnage PPT systématique randomisé et le nombre d'échantillons répétés est $K = 10^6$. La fonction `piij` donne une matrice de dimensions $n \times n$ où la $(ij)^{\mathrm{e}}$ entrée est la probabilité simulée $\pi_{ij}^*, i, j \in s$.

```
piij<-function(x,s){
N<-length(x)
n<-length(s)
p<-matrix(0,n,n)
for(k in 1:1000000){
ss<-syspps(x,n)
for(i in 1:(n-1)){
for(j in (i+1):n){
if(min(abs(ss-s[i]))+min(abs(ss-s[j]))==0)
    p[i,j]<-p[i,j]+1
  } } }
p<-(p+t(p))/1000000
return(p)
  }
```

C3. Une fonction R pour l'échantillonnage PPT sous substitution d'unités.

```
sysppssub<-function(x,n,refus){
s<-syspps(x,n)
sub<-numeric()
for (i in 1:n){
if(min(abs(s[i]-refus))==0) sub<-c(sub,i)
}
m<-length(sub)
if(m>0){
s<-s[-sub]
U1<-(1:length(x))[-c(refus,s)]
x1<-x[-c(refus,s)]
s1<-syspps(x1,m)
s<-c(s,U1[s1])
}
return(s[order(s)])
}
```

La méthode par défaut pour la sélection de l'échantillon initial et de l'échantillon de substitution est l'échantillonnage PPT systématique randomisé. La fonction R `sysppssub` est utilisée pour simuler les probabilités d'inclusion sous substitution d'unités. Les variables d'entrée sont `x` : le vecteur de population des variables de taille, `n` : la taille de l'échantillon et `refus` : l'ensemble d'unités de l'échantillon initial qui refusent de répondre. La fonction donne un ensemble d'unités pour l'échantillon final.

C4. Codes R pour la simulation des π_i sous substitution d'unités.

```
pi<-rep(0,N)
for(i in 1:1000000){
s<-sysppssub(x,n,refus)
for(j in 1:N){
if(min(abs(s-j))==0) pi[j]<-pi[j]+1
 } }
pi<-pi/1000000
```

Bibliographie

Asok, C., et Sukhatme, B.V. (1976). On Sampford's procedure of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 71, 912-918.

Chao, M.T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69, 653-656.

Chen, X.H., Dempster, A.P. et Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81, 457-469.

Connor, W.S. (1966). An exact formula for the probability that two specified sampling units occur in a sample drawn with unequal probability and without replacement. *Journal of the American Statistical Association*, 61, 384-390.

Deville, J.C., et Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85, 89-101.

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93, 269-278.

Fellegi, I.P. (1963). Sampling with varying probabilities without replacement: Rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.

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

Gray, G.B. (1973). On increasing the sample size (number of psu's). Note technique, Statistique Canada.

Hartley, H.O., et Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

Hidiroglou, M.A., et Gray, G.B. (1980). Construction of joint probability of selection for systematic P.P.S. sampling. *Applied Statistics*, 29, 107-112.

Kott, P.S. (2005). A note on the Hartley-Rao variance estimator. *Journal of Official Statistics*, 21, 433-439.

Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal Indian Statist. Association*, 3, 173-180.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

Tillé, Y. (1996). An elimination procedure for unequal probability sampling without replacement. *Biometrika*, 83, 238-241.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2008
Vol. 34, N° 1, pp. 13-20
Statistique Canada, N° 12-001-X au catalogue

Estimation pour petits domaines sous un modèle à deux niveaux

Mahmoud Torabi et J.N.K. Rao [1]

Résumé

Lehtonen et Veijanen (1999) ont proposé un nouvel estimateur par la régression généralisée (GREG) assisté par modèle d'une moyenne de petit domaine sous un modèle à deux niveaux. Ils ont montré que l'estimateur proposé donne de meilleurs résultats que l'estimateur GREG habituel en ce qui concerne le biais relatif absolu moyen et l'erreur relative absolue médiane moyenne. Nous calculons l'erreur quadratique moyenne (EQM) du nouvel estimateur GREG sous le modèle à deux niveaux et nous la comparons à celle de l'estimateur fondé sur le meilleur prédicteur linéaire sans biais (BLUP). Nous présentons aussi des résultats empiriques concernant l'efficacité relative des estimateurs. Nous montrons que le nouvel estimateur GREG a de meilleures propriétés que l'estimateur GREG habituel en ce qui concerne l'EQM moyenne et l'erreur relative absolue moyenne. Nous montrons aussi que, parce qu'il emprunte de l'information aux petits domaines apparentés, l'estimateur EBLUP donne des résultats nettement meilleurs que l'estimateur GREG habituel et que le nouvel estimateur GREG. Nous fournissons les résultats de simulation sous un modèle, ainsi qu'en population finie réelle.

Mots clés : Estimateur BLUP; estimateur GREG; erreur quadratique moyenne; effets aléatoires; moyennes de petit domaine.

1. Introduction

L'Estimation pour petits domaines a suscité beaucoup d'intérêt ces dernières années à cause de la demande croissante de statistiques fiables pour des petits domaines. Les estimateurs par domaine directs classiques ne sont pas suffisamment précis, parce que, pour les petits domaines, les tailles d'échantillons sont souvent insuffisantes. Il faut donc recourir à des estimateurs indirects renforcés par l'emprunt d'information à des domaines apparentés, en particulier des estimateurs indirects fondés sur un modèle. En estimation sur petits domaines, des modèles à effets aléatoires au niveau de l'unité, dont les modèles de régression linéaire à erreurs emboîtées et les modèles à deux niveaux, sont fréquemment adoptés pour obtenir de bons estimateurs fondés sur un modèle des moyennes de petit domaine. Rao (2003) donne une revue complète de l'Estimation pour petits domaines fondée sur un modèle.

Un modèle à deux niveaux est donné par

$$y_{ij} = x'_{ij}\beta_i + e_{ij};$$
$$\beta_i = Z_i\beta + \nu_i,\ j = 1, ..., N_i; i = 1, ..., m \qquad (1)$$

où N_i est le nombre d'unités dans le i^{e} domaine ($i = 1, ..., m$), y_{ij} est la réponse et x_{ij} est un vecteur de dimension $p \times 1$ de covariables au niveau de l'unité attaché à la j^{e} unité dans le i^{e} domaine. En outre, Z_i est une matrice de dimensions $p \times q$ de covariables au niveau du domaine, β est un vecteur de dimension $q \times 1$ de paramètres de régression, les ν_i sont des vecteurs aléatoires indépendants de moyenne nulle et de covariance $\sum_\nu$, et les e_{ij} sont des variables aléatoires indépendantes de moyenne nulle et de variance σ_e^2 qui sont indépendantes des ν_i. Nous pouvons exprimer la moyenne $\bar{Y}_i$ du i^{e} domaine sous la forme

$$\bar{Y}_i \approx \mu_i = \bar{X}'_i(Z_i\beta + \nu_i),$$

en supposant que N_i est grand, où $\bar{X}_i$ est la moyenne de population connue de x_{ij} dans le i^{e} domaine. Nous supposons que les valeurs d'échantillon $\{(y_{ij}, x_{ij}), j = 1, ..., n_i; i = 1, ..., m\}$ obéissent au modèle (1), c'est-à-dire qu'il n'y a pas de biais de sélection d'échantillon. Pour l'échantillon, le modèle est donné par

$$y_{ij} = x'_{ij}(Z_i\beta + \nu_i) + e_{ij},\ j = 1, ..., n_i;\ i = 1, ..., m. \qquad (2)$$

En notation matricielle, (2) peut s'écrire

$$y_i = X_i(Z_i\beta + \nu_i) + e_i,\ i = 1, ..., m$$

avec $\text{Var}(y_i) = V_i = X_i\Sigma_\nu X'_i + \sigma_e^2 I_{n_i}$, où y_i est un vecteur de dimension $n_i \times 1$ et X_i est une matrice de dimensions $n_i \times p$. Le modèle à deux niveaux (2) a été proposé pour la première fois par Moura et Holt (1999) dans le contexte de l'estimation sur petits domaines. Il intègre effectivement l'utilisation de covariables aux niveaux de l'unité et du domaine dans un seul modèle, par modélisation des pentes aléatoires β_i dans (1) en fonction de covariables au niveau du domaine Z_i.

Lehtonen et Veijanen (1999) ont proposé un nouvel estimateur par la régression généralisée (GREG) assisté par modèle d'une moyenne de petit domaine sous le modèle à deux niveaux. Ils ont montré que le nouvel estimateur GREG basé sur le modèle (1) donne de meilleurs résultats que l'estimateur GREG habituel fondé sur un modèle avec $\beta_i = Z_i\beta$ fixes. Moura et Holt (1999) ont obtenu

1. Mahmoud Torabi, Department of Pediatrics, University of Alberta, Edmonton (Alberta) T6G 2J3; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa (Ontario) K1S 5B6.

l'estimateur fondé sur le meilleur prédicteur linéaire sans biais (BLUP) de la moyenne de petit domaine μ_i et son EQM sous le modèle à deux niveaux (2); voir la section 2. Lehtonen, Särndal et Veijanen (2003) ont étudié l'effet du choix du modèle sur divers types d'estimateurs (synthétique, GREG et composite) de moyenne de petit domaine.

À la section 3, nous commençons par calculer l'erreur quadratique moyenne (EQM) du nouvel estimateur GREG et de l'estimateur GREG habituel (section 2) sous le modèle à deux niveaux (2), en supposant que les paramètres du modèle sont connus. Puis, nous comparons les EQM de l'estimateur GREG habituel, du nouvel estimateur GREG et de l'estimateur BLUP, et nous obtenons une expression explicite de l'accroissement de l'EQM du nouvel estimateur GREG relativement à l'EQM de l'estimateur BLUP. À la section 4, nous présentons des résultats empiriques concernant l'efficacité relative des estimateurs quand les paramètres du modèle sont estimés. Nous réalisons l'étude par simulation sous un cadre fondé sur un modèle, ainsi qu'en population finie réelle. Enfin, à la section 5, nous présentons certaines conclusions.

2. Estimateurs BLUP et GREG

Le modèle à deux niveaux (2) est un cas particulier d'un modèle linéaire mixte général avec une structure diagonale par blocs de la matrice de covariance. Par conséquent, en supposant que les paramètres du modèle sont connus, nous pouvons calculer l'estimateur BLUP de μ_i sous la forme

$$\tilde{\mu}_i^B = \bar{X}_i'(Z_i\beta + \tilde{\nu}_i), \tag{3}$$

où

$$\tilde{\nu}_i = \Sigma_v X_i' V_i^{-1}(y_i - X_i Z_i \beta),$$

et l'indice supérieur B sur $\tilde{\mu}_i$ désigne l'estimateur BLUP (Rao 2003, page 107). De même, l'estimateur BLUP d'une observation non échantillonnée j dans le i^{e} domaine peut s'écrire

$$\tilde{y}_{ij} = x_{ij}'(Z_i\beta + \tilde{\nu}_i). \tag{4}$$

Par ailleurs, un estimateur GREG assisté par modèle de μ_i (ou de $\bar{Y}_i$) est donné par

$$\tilde{\mu}_i^G = \frac{1}{N_i}\left[\sum_{j=1}^{N_i}\tilde{y}_{ij} + \sum_{j=1}^{n_i} w_{ij}(y_{ij} - \tilde{y}_{ij})\right],\ i = 1, \ldots, m \tag{5}$$

où $\tilde{y}_{ij}$ est le prédicteur de y_{ij} sous le modèle postulé, et w_{ij} est le poids de sondage qui est égal à N_i/n_i dans le cas de l'échantillonnage aléatoire simple (EAS) dans les domaines. Dans le présent exposé, nous nous intéressons essentiellement à l'EAS dans les domaines.

Partant de (5) avec $\tilde{y}_{ij} = x_{ij}' Z_i \beta$ pour prédicteur de y_{ij} sous le modèle (1) avec $\beta_i = Z_i\beta$ fixes, nous pouvons écrire l'estimateur GREG habituel sous la forme

$$\tilde{\mu}_i^G = \bar{y}_i + (\bar{X}_i - \bar{x}_i)' Z_i \beta, \tag{6}$$

où l'indice supérieur G sur $\tilde{\mu}_i$ signifie GREG (Särdnal, Swensson et Wretman 1992, page 225), $\bar{y}_i$ est la moyenne d'échantillon de y_{ij} dans le i^{e} domaine, et $\bar{x}_i$ est la moyenne d'échantillon de x_{ij} dans le i^{e} domaine, respectivement. En utilisant dans (5) le prédicteur (4) fondé sur le modèle à deux niveaux (1), nous obtenons un nouvel estimateur GREG de μ_i (ou de $\bar{Y}_i$) de la forme

$$\begin{aligned}\tilde{\mu}_i^{\text{LV}} &= [\bar{X}_i'(Z_i\beta + \tilde{\nu}_i) + (\bar{y}_i - \bar{x}_i'(Z_i\beta + \tilde{\nu}_i))] \\ &= \bar{y}_i + (\bar{X}_i - \bar{x}_i)'(Z_i\beta + \tilde{\nu}_i),\end{aligned} \tag{7}$$

où l'indice supérieur LV sur $\tilde{\mu}_i$ indique qu'il a été proposé pour la première fois par Lehtonen et Veijanen (1999). Les estimateurs $\tilde{\mu}_i^B$, $\tilde{\mu}_i^G$ et $\tilde{\mu}_i^{\text{LV}}$ sont linéaires en les y_{ij} et sans biais sous le modèle à deux niveaux (1). En pratique, nous remplaçons les paramètres β, Σ_v et σ_e^2 dans (3), (6) et (7) par des estimateurs appropriés. Les estimateurs résultants sont désignés par $\hat{\mu}_i^B$, $\hat{\mu}_i^G$ et $\hat{\mu}_i^{\text{LV}}$, respectivement, où $\hat{\mu}_i^B$ est l'estimateur BLUP empirique (EBLUP). Sous l'hypothèse de normalité, $\hat{\mu}_i^B$ est le meilleur estimateur empirique (EB pour *empirical best*). L'estimateur EBLUP de $\bar{Y}_i$ est donné à la section 4.2.2. Notons que $\hat{\mu}_i^G$ et $\hat{\mu}_i^{\text{LV}}$ sont valides en tant qu'estimateurs de $\bar{Y}_i$.

3. Erreur quadratique moyenne

L'erreur quadratique moyenne (EQM) de l'estimateur GREG habituel $\tilde{\mu}_i^G$ sous le modèle à deux niveaux peut s'écrire

$$\begin{aligned}\text{EQM}(\tilde{\mu}_i^G) &= E(\tilde{\mu}_i^G - \mu_i)^2 \\ &= E[\bar{y}_i + (\bar{X}_i - \bar{x}_i)'Z_i\beta - \bar{X}_i'(Z_i\beta + \nu_i)]^2 \\ &= E[(\bar{x}_i - \bar{X}_i)'\nu_i + \bar{e}_i]^2 \\ &= (\bar{x}_i - \bar{X}_i)'\Sigma_v(\bar{x}_i - \bar{X}_i) + \frac{\sigma_e^2}{n_i},\end{aligned}$$

comme il est énoncé dans le théorème 1.

Théorème 1. *L'EQM de l'estimateur GREG* (6) *est donnée par*

$$\text{EQM}(\tilde{\mu}_i^G) = (\bar{x}_i - \bar{X}_i)'\Sigma_v(\bar{x}_i - \bar{X}_i) + \frac{\sigma_e^2}{n_i}. \tag{8}$$

En outre, nous pouvons écrire l'EQM de l'estimateur BLUP $\tilde{\mu}_i^B$ comme il suit :

$$\begin{aligned}\mathrm{EQM}(\tilde{\mu}_i^B) &= E(\tilde{\mu}_i^B - \mu_i)^2 \\ &= E[\bar{X}_i'(\tilde{\nu}_i - \nu_i)]^2 \\ &= \bar{X}_i'\left(\Sigma_v - \Sigma_v X_i' V_i^{-1} X_i \Sigma_v\right)\bar{X}_i,\end{aligned}$$

comme il est énoncé dans le théorème 2.

Théorème 2. *L'EQM de l'estimateur BLUP* (3) *est donnée par*

$$\mathrm{EQM}(\tilde{\mu}_i^B) = \bar{X}_i'\left(\Sigma_v - \Sigma_v X_i' V_i^{-1} X_i \Sigma_v\right)\bar{X}_i. \qquad (9)$$

Le théorème 3 donne l'EQM du nouvel estimateur GREG $\tilde{\mu}_i^{\mathrm{LV}}$.

Théorème 3. *L'EQM du nouvel estimateur GREG* (7) *est donnée par*

$$\begin{aligned}\mathrm{EQM}(\tilde{\mu}_i^{\mathrm{LV}}) = \mathrm{EQM}(\tilde{\mu}_i^B) \\ + \left\{\bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{x}_i - \bar{x}_i' \Sigma_v \bar{x}_i + \frac{\sigma_e^2}{n_i}\right\}. \qquad (10)\end{aligned}$$

La preuve du théorème 3 est donnée en annexe.

Par définition, nous avons $\mathrm{EQM}(\tilde{\mu}_i^B) \leq \mathrm{EQM}(\tilde{\mu}_i^{\mathrm{LV}})$ et (10) donne une expression explicite de l'accroissement de l'EQM de $\tilde{\mu}_i^{\mathrm{LV}}$ comparativement à l'EQM de l'estimateur BLUP $\tilde{\mu}_i^B$.

4. Résultats empiriques

4.1 Comparaison empirique des valeurs de l'EQM

Afin d'étudier l'efficacité du nouvel estimateur GREG, nous avons utilisé des données tirées de Moura et Holt (1999) recueillies auprès de 38 740 ménages dans les districts de recensement (petits domaines) d'un comté du Brésil. Le revenu du chef de ménage a été traité comme la variable de réponse y. Deux variables indépendantes ont été définies au niveau de l'unité, à savoir le niveau d'études du chef du ménage (échelle ordinale de 0 à 5) et le nombre de pièces dans le logement du ménage (1 à 11+). Nous avons supposé que le modèle à deux niveaux pour ces données était de la forme :

$$y_{ij} = \beta_{i0} + \beta_{i1} x_{1ij} + \beta_{i2} x_{2ij} + e_{ij},$$
$$j = 1, \ldots, N_i;\ i = 1, \ldots, m \qquad (11)$$

avec

$$\beta_{i0} = \beta_0 + \nu_{i0};\ \beta_{i1} = \beta_1 + \nu_{i1};\ \beta_{i2} = \beta_2 + \nu_{i2}, \quad (12)$$

où

$$\nu = (\nu_{i0}, \nu_{i1}, \nu_{i2})' \overset{\text{iid}}{\sim} N_3(0, \Sigma_v),\ e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$$

et x_1 et x_2 représentent, respectivement, le nombre de pièces et le niveau d'études du chef du ménage (centrées chacune autour de leur moyenne de population). Notons que le modèle (12) pour les coefficients β_i aléatoires ne contient pas de covariables Z au niveau du domaine.

Moura et Holt (1999) ont également étudié un autre modèle avec une covariable Z au niveau du domaine pour la modélisation des β_i dans (12). Pour ces données, le nombre moyen d'automobiles par ménage dans chaque domaine a été utilisé comme covariable z pour la modélisation des coefficients aléatoires β_{i1} et β_{i2} correspondant aux variables x_1 et x_2, mais non pour le terme constant aléatoire, β_{i0}. Le modèle à deux niveaux (11) avec la covariable z au niveau du domaine est donné par

$$y_{ij} = \beta_{i0} + \beta_{i1} x_{1ij} + \beta_{i2} x_{2ij} + \varepsilon_{ij},$$
$$j = 1, \ldots, N_i;\ i = 1, \ldots, m \qquad (13)$$

avec

$$\beta_{i0} = \beta_0 + \nu_{i0};$$
$$\beta_{i1} = \beta_1 + \alpha_1 z_i + \nu_{i1};\ \beta_{i2} = \beta_2 + \alpha_2 z_i + \nu_{i2}. \quad (14)$$

Moura et Holt (1999) ont ajusté les modèles (11)-(12) et (13)-(14) à l'ensemble de données complet susmentionné. Nous résumons leurs résultats au tableau 1.

Tableau 1
Estimations des paramètres fondées sur l'ensemble de données de Moura et Holt (1999), où σ_0^2, σ_1^2 et σ_2^2 sont les éléments diagonaux et σ_{01}, σ_{02} et σ_{12} sont les éléments non diagonaux de la matrice de covariance Σ_ν

Paramètre	Covariance diagonale : modèle (14) avec z	Covariance diagonale : modèle (12) sans z	Covariance générale : modèle (12) sans z
β_0	8,442	8,688	8,456
β_1	0,451	1,321	1,223
β_2	0,744	2,636	2,596
α_1	3,779	-	-
α_2	1,659	-	-
σ_0^2	0,745	0,637	1,385
σ_1^2	0,237	0,471	0,234
σ_2^2	0,700	1,472	0,926
σ_{01}	-	-	0,354
σ_{02}	-	-	0,492
σ_{12}	-	-	0,333
σ_e^2	44,00	44,01	47,74

Les moyennes de domaine de x_1 et x_2 sont calculées d'après l'ensemble complet de données et traitées comme étant les moyennes de population $\bar{X}_{1i}$ et $\bar{X}_{2i}$. Nous avons sélectionné un sous-échantillon aléatoire de 10 % des enregistrements dans chaque petit domaine. La taille globale de l'échantillon est $n = 3\,876$ et le nombre de petits domaines est $m = 140$. En utilisant les valeurs d'échantillon de x_1, x_2 et z, ainsi que les moyennes de population $\bar{X}_{1i}$ et $\bar{X}_{2i}$, nous avons calculé $\mathrm{EQM}(\tilde{\mu}_i^G)$, $\mathrm{EQM}(\tilde{\mu}_i^B)$ et $\mathrm{EQM}(\tilde{\mu}_i^{\mathrm{LV}})$ en utilisant (8), (9) et (10), respectivement, et en traitant les estimations des paramètres de régression, Σ_v et σ_e^2 du tableau 1 pour les données complètes comme étant les valeurs réelles. Ensuite, nous avons calculé les valeurs moyennes de l'EQM sur les domaines :

$$\overline{\mathrm{EQM}}_G = \frac{1}{m}\sum_{i=1}^{m} \mathrm{EQM}(\tilde{\mu}_i^G),$$

$$\overline{\mathrm{EQM}}_B = \frac{1}{m}\sum_{i=1}^{m} \mathrm{EQM}(\tilde{\mu}_i^B)$$

et

$$\overline{\mathrm{EQM}}_{\mathrm{LV}} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{EQM}(\tilde{\mu}_i^{\mathrm{LV}}).$$

Nous définissions l'efficacité relative de $\tilde{\mu}^B$ par rapport à $\tilde{\mu}^G$ comme étant EFF_B et l'efficacité relative de $\tilde{\mu}^{\mathrm{LV}}$ par rapport à $\tilde{\mu}^G$ comme étant $\mathrm{EFF}_{\mathrm{LV}}$, où

$$\mathrm{EFF}_B = \frac{\overline{\mathrm{EQM}}_G}{\overline{\mathrm{EQM}}_B}; \qquad \mathrm{EFF}_{\mathrm{LV}} = \frac{\overline{\mathrm{EQM}}_G}{\overline{\mathrm{EQM}}_{\mathrm{LV}}}.$$

Nous résumons les résultats aux tableaux 2 et 3. Ces derniers révèlent que le nouvel estimateur GREG est un peu plus efficace que l'estimateur GREG habituel en ce qui concerne l'EQM moyenne : $\mathrm{EFF}_{\mathrm{LV}} \le 112$ %. Toutefois, il est considérablement moins efficace que l'estimateur BLUP sous le modèle à deux niveaux postulé. Par exemple, pour le modèle avec z et la matrice de covariance diagonale (tableau 2), $\mathrm{EFF}_B = 292$ % comparativement à $\mathrm{EFF}_{\mathrm{LV}} =$ 106 %, et $\overline{\mathrm{EQM}}_B = 0{,}62$ comparativement à $\overline{\mathrm{EQM}}_{\mathrm{LV}} =$ 1,72.

Tableau 2
Comparaison des estimateurs sur petits domaines : efficacité relative (EFF) et EQM moyenne (EQM) pour le cas de la matrice de covariance diagonale basé sur l'ensemble de données de Moura et Holt (1999)

Mesure de la qualité	Modèle sans z			Modèle avec z		
	GREG	Nouveau GREG	BLUP	GREG	Nouveau GREG	BLUP
EFF	100 %	112 %	306 %	100 %	106 %	292 %
$\overline{\mathrm{EQM}}$	1,92	1,71	0,62	1,83	1,72	0,62

Tableau 3
Comparaison des estimateurs sur petits domaines : efficacité relative (EFF) et EQM moyenne ($\overline{\mathrm{EQM}}$) pour le cas d'une matrice de covariance générale basé sur l'ensemble de données de Moura et Holt (1999)

Mesure de la qualité	Modèle sans z		
	GREG	Nouveau GREG	BLUP
EFF	100 %	108 %	253 %
$\overline{\mathrm{EQM}}$	2,02	1,87	0,80

4.2 Étude par simulation

4.2.1 Étude par simulation sous un cadre fondé sur un modèle

Afin d'étudier l'efficacité du nouvel estimateur GREG lorsque les paramètres du modèle sont estimés, nous avons entrepris une petite étude par simulation fondée sur les modèles à deux niveaux (11)-(12) et (13)-(14). Nous avons considéré uniquement la structure de covariance diagonale Σ_v avec les éléments diagonaux σ_0^2, σ_1^2 et σ_2^2. Nous avons de nouveau utilisé les données tirées de Moura et Holt (1999). Les estimations de $\beta_0, \beta_1, \beta_2, \alpha_1, \alpha_2, \sigma_0^2, \sigma_1^2, \sigma_2^2$ et σ_e^2 présentées au tableau 1 sont traitées comme les valeurs réelles.

Dans notre étude par simulation, nous avons pris (x_{1ij}, x_{2ij}, z_i) provenant de Moura et Holt (1999), puis nous avons généré y_{ij} en nous fondant sur les modèles (11)-(12) et (13)-(14). En utilisant les échantillons ainsi produits $(y_{ij}^{(b)}, x_{1ij}, x_{2ij}, z_i)$, $b = 1, ..., B = 1\,000$, nous avons calculé $\hat{\beta}^{(b)}$ par les moindres carrés généralisés pour la nouvelle méthode GREG ainsi que pour la méthode BLUP. Pour la méthode GREG, nous avons utilisé les moindres carrés ordinaires pour estimer β par $\hat{\beta}_{\mathrm{ols}}^{(b)}$. En outre, nous avons calculé $\hat{\Sigma}_v^{(b)}$ et $\hat{\sigma}_e^{2(b)}$ par la méthode du maximum de vraisemblance restreint (REML). Pour chaque échantillon produit, nous avons calculé

$$\mu_i^{(b)} = \bar{X}'_i(Z_i\beta + \nu_i^{(b)}),\ i = 1, ..., m;\ b = 1, ..., B.$$

Nous avons calculé le nouvel estimateur GREG de $\mu_i^{(b)}$ selon $\hat{\mu}_i^{\mathrm{LV}(b)} = \bar{y}_i^{(b)} + (\bar{X}_i - \bar{x}_i)'(Z_i\hat{\beta}^{(b)} + \hat{\nu}_i^{(b)})$, l'estimateur GREG de $\mu_i^{(b)}$ selon $\hat{\mu}_i^{G(b)} = \bar{y}_i^{(b)} + (\bar{X}_i - \bar{x}_i)' Z_i \hat{\beta}_{\mathrm{ols}}^{(b)}$ et l'estimateur BLUP empirique (EBLUP) de $\mu_i^{(b)}$ selon $\hat{\mu}_i^{B(b)} = \bar{X}'_i(Z_i\hat{\beta}^{(b)} + \hat{\nu}_i^{(b)})$, où $\hat{\nu}_i^{(b)} = \hat{\Sigma}_v^{(b)} X'_i \hat{V}_i^{-1(b)}(y_i^{(b)} - X_i Z_i \hat{\beta}^{(b)})$.

Ensuite, nous avons calculé la moyenne des erreurs quadratiques moyennes ($\overline{\mathrm{EQM}}_1$) et la moyenne des erreurs relatives absolues ($\overline{\mathrm{ERA}}_1$)

$$\overline{\mathrm{EQM}}_1 = \frac{1}{m}\sum_i^m \mathrm{EQM}_{1i} \text{ où } \mathrm{EQM}_{1i} = B^{-1}\sum_{b=1}^{B}(\hat{\mu}_i^{(b)} - \mu_i^{(b)})^2,$$

$$\overline{\mathrm{ERA}}_1 = \frac{1}{m}\sum_i^m \mathrm{ERA}_{1i} \text{ où } \mathrm{ERA}_{1i} = B^{-1}\sum_{b=1}^{B}|\hat{\mu}_i^{(b)} - \mu_i^{(b)}|/\mu_i^{(b)},$$

où $\hat{\mu}_i^{(b)}$ dénote $\hat{\mu}_i^{\mathrm{LV}(b)}$, $\hat{\mu}_i^{G(b)}$ ou $\hat{\mu}_i^{B(b)}$. Nous présentons les résultats au tableau 4. Les deux modèles avec covariable z au niveau du domaine et sans covariable z produisent des valeurs un peu plus faibles de $\overline{\mathrm{EQM}}_1$ et de $\overline{\mathrm{ERA}}_1$ pour le nouvel estimateur GREG que pour l'estimateur GREG habituel. Cependant, $\overline{\mathrm{EQM}}_1$ et $\overline{\mathrm{ERA}}_1$ sont significativement plus faibles pour l'estimateur EBLUP, à cause de l'emprunt d'information aux domaines apparentés. En outre, la comparaison des tableaux 2 et 4 révèle que les valeurs de $\overline{\mathrm{EQM}}_1$ dans le tableau 4 sont un peu plus grandes que les valeurs correspondantes dans le tableau 2, à cause de l'estimation des paramètres du modèle.

Tableau 4
Comparaison des estimateurs sur petits domaines : EQM moyenne ($\overline{\mathrm{EQM}}_1$) et erreur relative absolue moyenne ($\overline{\mathrm{ERA}}_1$) sous un cadre fondé sur un modèle

Mesure de la qualité	Modèle sans z			Modèle avec z		
	GREG	Nouveau GREG	EBLUP	GREG	Nouveau GREG	EBLUP
$\overline{\mathrm{EQM}}_1$	1,93	1,73	0,67	1,84	1,73	0,73
$\overline{\mathrm{ERA}}_1$	0,14	0,13	0,08	0,13	0,12	0,08

4.2.2 Étude par simulation sous un cadre de population finie

Pour étudier les propriétés des estimateurs sous un cadre de population finie, nous avons créé une population finie synthétique d'après les données brésiliennes constituées de $n = 3\,876$ valeurs d'échantillon $(y_{ij}, x_{1ij}, x_{2ij}, z_i)$. En reproduisant les valeurs d'échantillon $(y_{ij}, x_{1ij}, x_{2ij}, z_i)$ cinq fois, nous avons obtenu le nouvel ensemble de données (y, x_1, x_2, z) de taille 19 380 que nous avons traité comme notre population réelle.

Nous avons produit 500 échantillons indépendants ($B = 500$), chacun de taille $n = 700$ et $n = 1\,400$, en tirant des échantillons aléatoires simples de taille $n_i = 5$ et $n_i = 10$ dans chaque domaine $i = 1, ..., 140$. Comme auparavant, pour chaque échantillon, nous avons calculé $\hat{\beta}^{(b)}$ pour la nouvelle méthode GREG et pour la méthode BLUP, et $\hat{\beta}_{\mathrm{ols}}^{(b)}$ pour la méthode GREG. En outre, nous avons calculé $\hat{\Sigma}_v^{(b)}$ et $\hat{\sigma}_e^{2(b)}$ par la méthode du REML. Nous avons également calculé la moyenne de population de y_{ij} pour chaque domaine i par

$$\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij}/N_i,\ i = 1, ..., 140,$$

où N_i est la taille de la population dans le i^{e} domaine. En outre, pour chaque échantillon $b = 1, ..., B$, nous avons calculé la nouvelle estimation GREG de la moyenne du i^{e} domaine selon $\hat{\bar{Y}}_i^{\mathrm{LV}(b)} = \bar{y}_i^{(b)} + (\bar{X}_i - \bar{x}_i^{(b)})'(Z_i\hat{\beta}^{(b)} + \hat{v}_i^{(b)})$, l'estimation GREG selon $\hat{\bar{Y}}_i^{G(b)} = \bar{y}_i^{(b)} + (\bar{X}_i - \bar{x}_i^{(b)})'Z_i\hat{\beta}_{\mathrm{ols}}^{(b)}$ et l'estimation EBLUP selon $\hat{\bar{Y}}_i^{B(b)} = f_i\,\bar{y}_i + (1-f_i)[\bar{X}_i^{*'}(Z_i\hat{\beta}^{(b)} + \hat{v}_i^{(b)})]$, où

$$f_i = n_i/N_i,\ \bar{X}_i^* = \frac{N_i\bar{X}_i - n_i\bar{x}_i}{N_i - n_i},$$

et $\hat{v}_i^{(b)} = \hat{\Sigma}_v^{(b)}X_i'\,\hat{V}_i^{-1(b)}(y_i^{(b)} - X_i\,Z_i\,\hat{\beta}^{(b)})$.

L'estimateur EBLUP tient compte des corrections pour population finie f_i.

Nous avons calculé la moyenne des erreurs quadratiques moyennes ($\overline{\mathrm{EQM}}_2$) et la moyenne des erreurs relatives absolues ($\overline{\mathrm{ERA}}_2$) comme étant

$$\overline{\mathrm{EQM}}_2 = \frac{1}{m}\sum_i^m \mathrm{EQM}_{2i},\quad \overline{\mathrm{ERA}}_2 = \frac{1}{m}\sum_i^m \mathrm{ERA}_{2i},$$

où

$$\mathrm{EQM}_{2i} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\bar{Y}}_i^{(b)} - \bar{Y}_i)^2,\quad \mathrm{ERA}_{2i} = \frac{1}{B}\sum_{b=1}^{B}|\hat{\bar{Y}}_i^{(b)} - \bar{Y}_i|\ /\bar{Y}_i,$$

et $\hat{\bar{Y}}_i^{(b)}$ dénote $\hat{\bar{Y}}_i^{\mathrm{LV}(b)}$, $\hat{\bar{Y}}_i^{G(b)}$ ou $\hat{\bar{Y}}_i^{B(b)}$. Nous présentons les résultats aux tableaux 5 et 6 pour $n_i = 5$ et $n_i = 10$, respectivement. Nous envisageons les modèles avec covariable au niveau du domaine z et sans la covariable z.

Tableau 5
Comparaison des estimateurs sur petits domaines : EQM moyenne ($\overline{\mathrm{EQM}}_2$) et erreur relative absolue moyenne ($\overline{\mathrm{ERA}}_2$) sous un cadre de population finie (n_i = 5)

Mesure de la qualité	Modèles sans z			Modèle avec z		
	GREG	Nouveau GREG	EBLUP	GREG	Nouveau GREG	EBLUP
$\overline{\mathrm{EQM}}_2$	11,03	10,02	6,50	10,76	10,06	7,06
$\overline{\mathrm{ERA}}_2$	0,27	0,24	0,18	0,25	0,23	0,22

Le tableau 5 montre que, pour $n_i = 5$, le nouvel estimateur GREG donne de meilleurs résultats que l'estimateur GREG habituel en ce sens que $\overline{\mathrm{EQM}}_2$ et $\overline{\mathrm{ERA}}_2$ sont plus faibles. Par ailleurs, le tableau 6 révèle que, pour $n_i = 10$, l'estimateur GREG habituel a d'un peu meilleures propriétés que le nouvel estimateur GREG en ce qui concerne $\overline{\mathrm{EQM}}_2$, mais non $\overline{\mathrm{ERA}}_2$. Cependant, l'estimateur EBLUP donne des valeurs considérablement plus faibles de $\overline{\mathrm{EQM}}_2$ et $\overline{\mathrm{ERA}}_2$ que l'estimateur GREG habituel et le nouvel estimateur GREG dans les deux cas, parce qu'il emprunte de l'information aux petits domaines apparentés. Par exemple, pour le modèle sans z et $n_i = 5$, $\overline{\mathrm{EQM}}_2$ = 10,02, 11,03 et 6,50 pour le nouvel estimateur GREG, l'estimateur GREG habituel et l'estimateur EBLUP, respectivement.

Tableau 6
Comparaison des estimateurs sur petits domaines : EQM moyenne ($\overline{\text{EQM}}_2$) et erreur relative absolue moyenne ($\overline{\text{ERA}}_2$) sous un cadre de population finie (n_i = 10)

Mesure de la qualité	Modèle sans z			Modèle avec z		
	GREG	Nouveau GREG	EBLUP	GREG	Nouveau GREG	EBLUP
$\overline{\text{EQM}}_2$	6,53	6,77	4,73	6,75	6,96	5,24
$\overline{\text{ERA}}_2$	0,20	0,18	0,15	0,19	0,18	0,19

5. Sommaire

Dans le présent article, nous avons calculé l'erreur quadratique moyenne (EQM) de modélisation pour un nouvel estimateur GREG assisté par un modèle à deux niveaux d'une moyenne de petit domaine proposé par Lehtonen et Veijanen (1999). En outre, nous avons utilisé un ensemble de données de Moura et Holt (1999) pour démontrer empiriquement que l'estimateur BLUP est considérablement plus efficace que le nouvel estimateur GREG en ce qui concerne l'EQM sous le modèle, tandis que le nouvel estimateur GREG n'est que légèrement plus efficace que l'estimateur GREG habituel fondé sur le modèle de régression $y_i = X_i Z_i \beta + e_i, i = 1, ..., m.$ En outre, à l'aide d'une étude par simulation sous un cadre fondé sur un modèle, nous avons montré que le nouvel estimateur GREG a systématiquement de meilleures propriétés que l'estimateur GREG habituel en ce qui concerne l'EQM moyenne, $\overline{\text{EQM}}$, et l'erreur relative absolue moyenne, $\overline{\text{ERA}}$. Cependant, étant donné l'emprunt d'information à des petits domaines apparentés, l'estimateur EBLUP donne de nettement meilleurs résultats que les estimateurs GREG habituel et nouveau. En outre, nous avons réalisé une étude par simulation sous un cadre de population finie et montré que l'estimateur EBLUP surpasse le nouvel estimateur GREG et l'estimateur GREG habituel en ce qui concerne l'$\overline{\text{EQM}}$ et l'$\overline{\text{ERA}}$.

Remerciements

La présente étude a été financée par une subvention du Conseil de recherches en sciences naturelles et en génie du Canada. Elle est fondée sur un chapitre de la thèse de doctorat de M.Torabi rédigée sous la supervision de J.N.K. Rao. Les auteurs remercient les examinateurs et le rédacteur adjoint de leurs commentaires constructifs concernant la version originale du présent article.

Annexe

Calcul de $\text{EQM}(\tilde{\mu}_i^{\text{LV}})$:

$$\begin{aligned}\text{EQM}(\tilde{\mu}_i^{\text{LV}}) &= E(\tilde{\mu}_i^{\text{LV}} - \mu_i)^2 \\ &= E[\bar{X}_i'(\tilde{\nu}_i - \nu_i)]^2 + E[\bar{y}_i - \bar{x}_i'(Z_i\beta + \tilde{\nu}_i)]^2 \\ &\quad + 2E[(\bar{y}_i - \bar{x}_i'(Z_i\beta + \tilde{\nu}_i))\,\bar{X}_i'(\tilde{\nu}_i - \nu_i)], \qquad \text{(A.1)}\end{aligned}$$

où le premier terme du deuxième membre de (A.1) est l'EQM de l'estimateur BLUP sous le modèle à deux niveaux donnée par (9). De surcroît, nous pouvons écrire

$$\begin{aligned}E[\bar{y}_i - \bar{x}_i'(Z_i\beta + \tilde{\nu}_i)]^2 &= E[\bar{y}_i - \bar{x}_i' Z_i\beta]^2 \\ &\quad + E(\bar{x}_i'\tilde{\nu}_i)^2 - 2E[(\bar{y}_i - \bar{x}_i' Z_i\beta)(\bar{x}_i'\tilde{\nu}_i)], \qquad \text{(A.2)}\end{aligned}$$

où

$$E[\bar{y}_i - \bar{x}_i' Z_i\beta]^2 = \text{Var}(\bar{y}_i) = \bar{x}_i'\Sigma_v\bar{x}_i + \frac{\sigma_e^2}{n_i},$$

et

$$\begin{aligned}E(\bar{x}_i'\tilde{\nu}_i)^2 &= \text{Var}(\bar{x}_i'\tilde{\nu}_i) + [E(\bar{x}_i'\tilde{\nu}_i)]^2 \\ &= \text{Var}\left[\bar{x}_i'\Sigma_v X_i' V_i^{-1}(\text{y}_i - X_i Z_i\beta)\right] \\ &\quad + \left[E\left(\bar{x}_i'\Sigma_v X_i' V_i^{-1}(\text{y}_i - X_i Z_i\beta)\right)\right]^2 \\ &= \bar{x}_i'\Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{x}_i.\end{aligned}$$

En outre,

$$\begin{aligned}E[(\bar{y}_i - \bar{x}_i' Z_i\beta)(\bar{x}_i'\tilde{\nu}_i)] &= E\left[\bar{y}_i\bar{x}_i'\Sigma_v X_i' V_i^{-1}(\text{y}_i - X_i Z_i\beta)\right] \\ &\quad - E\left[\bar{x}_i' Z_i\beta\,\bar{x}_i'\Sigma_v X_i' V_i^{-1}(\text{y}_i - X_i Z_i\beta)\right],\end{aligned}$$

où le deuxième terme est nul. Par conséquent, nous pouvons écrire

$$\begin{aligned}E[(\bar{y}_i - \bar{x}_i' Z_i\beta)(\bar{x}_i'\tilde{\nu}_i)] &= E\left[\bar{y}_i\bar{x}_i'\Sigma_v X_i' V_i^{-1}\text{y}_i\right] \\ &\quad - \bar{x}_i' Z_i\beta\,\bar{x}_i'\Sigma_v X_i' V_i^{-1} X_i Z_i\beta,\end{aligned}$$

où le premier terme peut s'écrire

$$\begin{aligned}&E\left[\frac{1'}{n_i}\text{y}_i\,\bar{x}_i'\Sigma_v X_i' V_i^{-1}\text{y}_i\right] = \\ &\qquad \frac{1'}{n_i}\left(X_i\Sigma_v\bar{x}_i + X_i Z_i\beta\,\bar{x}_i'\Sigma_v X_i' V_i^{-1} X_i Z_i\beta\right),\end{aligned}$$

en utilisant le lemme suivant :

LEMME 1 (Searle 1971). *Si y est un vecteur de dimension* $n \times 1$ *de moyenne* μ *et de matrice de variance-covariance* Σ *et que* b *est un vecteur de dimension* $n \times 1$, *alors* $E(\mathrm{y}b'\mathrm{y}) = \Sigma b + \mu b'\mu$.

Donc,

$$\begin{aligned} E[(\bar{y}_i - \bar{x}_i' Z_i \beta)(\bar{x}_i' \tilde{\nu}_i)] &= \bar{x}_i' \Sigma_v \bar{x}_i + \bar{x}_i' Z_i \beta \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta \\ &\quad - \bar{x}_i' Z_i \beta \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta \\ &= \bar{x}_i' \Sigma_v \bar{x}_i. \end{aligned}$$

Nous pouvons alors écrire (A.2) sous la forme :

$$\begin{aligned} E[\bar{y}_i - \bar{x}_i'(Z_i\beta + \tilde{\nu}_i)]^2 &= \bar{x}_i' \Sigma_v \bar{x}_i + \frac{\sigma_e^2}{n_i} \\ &\quad + \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{x}_i - 2\bar{x}_i' \Sigma_v \bar{x}_i \\ &= \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{x}_i - \bar{x}_i' \Sigma_v \bar{x}_i + \frac{\sigma_e^2}{n_i}. \quad \text{(A.3)} \end{aligned}$$

Enfin, nous devons trouver le terme produit croisé de (A.1). Nous avons

$$\begin{aligned} E[(\bar{y}_i - \bar{x}_i'(Z_i\beta + \tilde{\nu}_i))\bar{X}_i'(\tilde{\nu}_i - \nu_i)] &= E[\bar{y}_i \bar{X}_i'(\tilde{\nu}_i - \nu_i)] \\ &\quad - E[\bar{x}_i'(Z_i\beta + \tilde{\nu}_i)\bar{X}_i'(\tilde{\nu}_i - \nu_i)], \quad \text{(A.4)} \end{aligned}$$

où le premier terme du deuxième membre de (A.4) peut s'écrire

$$\begin{aligned} E[\bar{y}_i \bar{X}_i'(\tilde{\nu}_i - \nu_i)] &= E\left[\bar{y}_i \bar{X}_i'\left(\Sigma_v X_i' V_i^{-1}(\mathrm{y}_i - X_i Z_i \beta) - \nu_i\right)\right] \\ &= E\left[\bar{y}_i \bar{X}_i' \Sigma_v X_i' V_i^{-1} \mathrm{y}_i\right] \\ &\quad - E\left[\bar{y}_i \bar{X}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta\right] \\ &\quad - E[\bar{y}_i \bar{X}_i' \nu_i]. \quad \text{(A.5)} \end{aligned}$$

Le premier terme de (A.5) est

$$\begin{aligned} E\left[\bar{y}_i \bar{X}_i' \Sigma_v X_i' V_i^{-1} \mathrm{y}_i\right] &= E\left[\frac{1'}{n_i} \mathrm{y}_i \bar{X}_i' \Sigma_v X_i' V_i^{-1} \mathrm{y}_i\right] \\ &= \frac{1'}{n_i} X_i\left(\Sigma_v \bar{X}_i + Z_i \beta \bar{X}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta\right), \quad \text{(A.6)} \end{aligned}$$

le deuxième terme de (A.5) peut s'écrire

$$\begin{aligned} E\left[\bar{y}_i \bar{X}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta\right] &= \bar{x}_i' Z_i \beta \bar{X}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta, \quad \text{(A.7)} \end{aligned}$$

et le troisième terme peut être exprimé par

$$\begin{aligned} E[\bar{y}_i \bar{X}_i' \nu_i] &= E[(\bar{x}_i' Z_i \beta + \bar{x}_i' \nu_i + e_i)\bar{X}_i' \nu_i] \\ &= E[\bar{x}_i' \nu_i \bar{X}_i' \nu_i] = \bar{x}_i' \Sigma_v \bar{X}_i. \quad \text{(A.8)} \end{aligned}$$

Par conséquent, en substituant (A.6), (A.7) et (A.8) dans (A.5), nous avons

$$E[\bar{y}_i \bar{X}_i'(\tilde{\nu}_i - \nu_i)] = 0. \quad \text{(A.9)}$$

Nous nous penchons maintenant sur le deuxième terme de (A.4). Nous avons

$$\begin{aligned} E[\bar{x}_i'(Z_i\beta + \tilde{\nu}_i)\bar{X}_i'(\tilde{\nu}_i - \nu_i)] &= E[\bar{x}_i' Z_i \beta \bar{X}_i' \tilde{\nu}_i] + E[\bar{x}_i' \tilde{\nu}_i \bar{X}_i' \tilde{\nu}_i] \\ &\quad - E[\bar{x}_i' Z_i \beta \bar{X}_i' \nu_i] - E[\bar{x}_i' \tilde{\nu}_i \bar{X}_i' \nu_i]. \quad \text{(A.10)} \end{aligned}$$

Alors, nous obtenons l'expression qui suit pour les quatre termes du deuxième membre de (A.10) :

$$\begin{aligned} E[\bar{x}_i' Z_i \beta \bar{X}_i' \tilde{\nu}_i] &= E\left[\bar{x}_i' Z_i \beta \bar{X}_i' \Sigma_v X_i' V_i^{-1}(\mathrm{y}_i - X_i Z_i \beta)\right] = 0, \quad \text{(A.11)} \end{aligned}$$

$$\begin{aligned} E[\bar{x}_i' \tilde{\nu}_i \bar{X}_i' \tilde{\nu}_i] &= \bar{x}_i' \Sigma_{\tilde{\nu}} \bar{X}_i + \bar{x}_i' E(\tilde{\nu}_i)\bar{X}_i' E(\tilde{\nu}_i) \\ &= \bar{x}_i' \mathrm{Var}\left[\Sigma_v X_i' V_i^{-1}(\mathrm{y}_i - X_i Z_i \beta)\right]\bar{X}_i \\ &\quad + \bar{x}_i' E\left(\Sigma_v X_i' V_i^{-1}(\mathrm{y}_i - X_i Z_i \beta)\right) \\ &\quad \times \bar{X}_i' E\left(\Sigma_v X_i' V_i^{-1}(\mathrm{y}_i - X_i Z_i \beta)\right) \\ &= \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{X}_i, \quad \text{(A.12)} \end{aligned}$$

$$E[\bar{x}_i' Z_i \beta \bar{X}_i' \nu_i] = 0 \quad \text{(A.13)}$$

et

$$\begin{aligned} E[\bar{x}_i' \tilde{\nu}_i \bar{X}_i' \nu_i] &= E\left[\bar{x}_i' \Sigma_v X_i' V_i^{-1}(\mathrm{y}_i - X_i Z_i \beta)\bar{X}_i' \nu_i\right] \\ &= E\left[\bar{x}_i' \Sigma_v X_i' V_i^{-1} \mathrm{y}_i \bar{X}_i' \nu_i\right] \\ &\quad - E\left[\bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta \bar{X}_i' \nu_i\right] \\ &= E\left[\bar{x}_i' \Sigma_v X_i' V_i^{-1}(X_i Z_i \beta + X_i \nu_i + e_i)\bar{X}_i' \nu_i\right] \\ &= E\left[\bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \nu_i \bar{X}_i' \nu_i\right] \\ &= \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{X}_i. \quad \text{(A.14)} \end{aligned}$$

Par conséquent, en introduisant (A.11) à (A.14) par substitution dans (A.10), nous obtenons

$$E[\bar{x}_i'(Z_i\beta + \tilde{\nu}_i)\bar{X}_i'(\tilde{\nu}_i - \nu_i)] = 0. \quad \text{(A.15)}$$

Donc, il découle de (A.4), (A.9) et (A.15) que

$$E[(\bar{y}_i - \bar{x}_i'(Z_i\beta + \tilde{\nu}_i))\bar{X}_i'(\tilde{\nu}_i - \nu_i)] = 0. \quad \text{(A.16)}$$

Il découle maintenant de (A.1), (A.3) et (A.16) que

$$\mathrm{EQM}(\tilde{\mu}_i^{\mathrm{LV}}) = \mathrm{EQM}(\tilde{\mu}_i^{B}) + \left\{ \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{x}_i - \bar{x}_i' \Sigma_v \bar{x}_i + \frac{\sigma_e^2}{n_i} \right\},$$

comme il est énoncé au théorème 3.

Bibliographie

Lehtonen, R., et Veijanen, A. (1999). Domain estimation with logistic generalized regression and related estimators. *IASS Satellite Conference on Small Area Estimation*, Riga : Latvian Council of Science, 121- 128.

Lehtonen, R., Särndal, C.-E. et Veijanen, A. (2003). L'effet du choix d'un modèle dans l'estimation par domaine, dont les petits domaines. *Techniques d'enquête*, 29, 37-49.

Moura, F.A.S., et Holt, D. (1999). Production d'estimations régionales à partir de modèles multiniveau. *Techniques d'enquête*, 25, 81-89.

Rao, J.N.K. (2003). *Small Area Estimation*. Hoboken : New York : John Wiley & Sons, Inc.

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

Searle, S.R. (1971). *Linear Models*. New York : John Wiley & Sons, Inc.

Techniques d'enquête, juin 2008
Vol. 34, N° 1, pp. 21-31
Statistique Canada, N° 12-001-X au catalogue

Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions infraprovinciales au Canada

Yong You [1]

Résumé

L'Enquête sur la population active (EPA) réalisée au Canada permet de produire des estimations mensuelles du taux de chômage aux niveaux national et provincial. Le programme de l'EPA diffuse aussi des estimations du chômage pour des régions infraprovinciales, comme les régions métropolitaines de recensement (RMR) et les centres urbains (CU). Cependant, pour certaines de ces régions infraprovinciales, les estimations directes ne sont pas fiables, parce que la taille de l'échantillon est assez petite. Dans le contexte de l'EPA, l'estimation pour de petites régions a trait à l'estimation des taux de chômage pour des régions infraprovinciales telles que les RMR/CU à l'aide de modèles pour petits domaines. Dans le présent article, nous discutons de divers modèles, dont celui de Fay-Herriot et des modèles transversaux ainsi que chronologiques. En particulier, nous proposons un modèle non linéaire intégré à effets mixtes sous un cadre hiérarchique bayésien (HB) pour l'estimation du taux de chômage d'après les données de l'EPA. Nous utilisons les données mensuelles sur les bénéficiaires de l'assurance-emploi (a.-e.) au niveau de la RMR ou du CU comme covariables auxiliaires dans le modèle. Nous appliquons une approche HB ainsi que la méthode d'échantillonnage de Gibbs pour obtenir les estimations des moyennes et des variances a posteriori des taux de chômage au niveau de la RMR ou du CU. Le modèle HB proposé produit des estimations fondées sur un modèle fiables si l'on s'en tient à la réduction du coefficient de variation. Nous présentons dans l'article une analyse d'ajustement du modèle et une comparaison des estimations fondées sur le modèle aux estimations directes.

Mots clés : Effet de plan; cadre hiérarchique bayésien; modèle loglinéaire à effets mixtes; vérification du modèle; variance d'échantillonnage; petit domaine.

1. Introduction

Le taux de chômage est généralement considéré comme un indicateur important des résultats économiques. Au Canada, les estimations du taux de chômage sont produites mensuellement par le programme de l'Enquête sur la population active (EPA) de Statistique Canada. Cette dernière est une enquête mensuelle réalisée auprès de 53 000 ménages sélectionnés conformément à un plan d'échantillonnage stratifié à plusieurs degrés. Chaque mois, un sixième de l'échantillon est remplacé. Donc, cinq sixièmes de l'échantillon sont communs à deux mois consécutifs. Ce chevauchement des échantillons crée des corrélations qui peuvent être exploitées pour produire de meilleures estimations au moyen d'autres méthodes que l'estimation directe, telles que les méthodes fondées sur un modèle, qui renforcent l'estimation à l'aide de données accumulées au fil du temps; cette approche sera examinée plus en détail à la section 2. Pour une description détaillée du plan de sondage de l'EPA, consulter Gambino, Singh, Dufour, Kennedy et Lindeyer (1998). Le programme de l'EPA diffuse des estimations mensuelles du taux de chômage pour de grandes régions, telles que l'ensemble du pays et les provinces, ainsi que pour des régions locales (petites régions ou petits domaines), telles que les régions métropolitaines de recensement (RMR, c'est-à-dire les villes comptant plus de 100 000 habitants) et les autres centres urbains du Canada. Bien que les estimations nationales et provinciales soient celles qui intéressent le plus les médias, les estimations infraprovinciales sont également très importantes. Le programme de l'assurance-emploi (a.-e.) les utilise pour établir les règles à appliquer pour administrer le programme. En outre, les taux de chômage dans les RMR et les CU sont examinés de près au niveau local. Cependant, pour nombre de régions locales, la taille de l'échantillon n'est pas suffisante pour calculer des estimations directes adéquates, car l'EPA est conçue pour produire des estimations adéquates ou fiables aux niveaux national et provincial. Pour les estimations nationales, le coefficient de variation (CV) est de l'ordre de 2 % et pour les estimations provinciales, de l'ordre de 4 % à 7 %. En revanche, pour les RMR et les CU, il varie d'environ 7 % à 50 %. En outre, pour certains CU, il est même supérieur à 50 %. Les estimations directes basées sur les données de l'EPA ne sont pas très fiables pour certaines régions municipales, le CV étant très grand à cause de la petite taille de l'échantillon pour ces régions. Par conséquent, d'autres estimateurs, en particulier des estimateurs fondés sur un modèle, sont envisagés pour améliorer les estimations directes d'après l'EPA pour les petites régions. L'objectif du présent article est d'obtenir un estimateur fondé sur un modèle fiable qui produit de meilleures estimations que l'estimateur direct d'après l'EPA, c'est-à-dire des estimations dont le CV est petit et stable.

1. Yong You, Division des méthodes d'enquête auprès des ménages, Statistique Canada, Ottawa, (Ontario), K1A 0T6. Courriel : yong.you@statcan.ca.

En général, on utilise les estimateurs directs, fondés uniquement sur des données d'échantillon sur un domaine particulier, pour estimer les paramètres pour de grands domaines, tels que le pays et les provinces. Par contre, les tailles d'échantillon pour les petits domaines, particulièrement les petites régions géographiques, sont rarement suffisantes pour fournir des estimations directes fiables. Quand on calcule des estimations pour des petits domaines, il est nécessaire d'emprunter de l'information supplémentaire à des domaines apparentés pour former des estimateurs indirects qui accroissent la taille effective d'échantillon et, donc, la précision. De tels estimateurs indirects sont fondés sur des modèles implicites ou explicites qui établissent un lien avec des petits domaines apparentés grâce à des données supplémentaires, tels que des dénombrements du recensement et des dossiers administratifs. Aujourd'hui il est généralement reconnu que, s'il faut recourir à des estimations indirectes, celles-ci devraient être fondées sur des modèles explicites qui relient les petits domaines d'intérêt grâce à des données supplémentaires; voir Rao (2003). Les estimateurs fondés sur un modèle sont des estimateurs indirects en ce sens qu'ils sont obtenus en utilisant des modèles pour petits domaines, des estimations directes et des variables auxiliaires connexes. Les estimateurs fondés sur un modèle sont établis en vue d'améliorer les estimateurs directs fondés sur le plan de sondage en ce qui concerne la précision et la fiabilité, c'est-à-dire la réduction des coefficients de variation. Habituellement, les estimateurs sur petits domaines sont renforcés par emprunt de données recueillies pour des petits domaines similaires à une période donnée ou pour le même domaine au fil du temps, mais non les deux. Ces dernières années, plusieurs méthodes ont été élaborées en vue d'emprunter simultanément des données de renfort transversales (dimension spatiale) et longitudinales (dimension temporelle). Les estimateurs fondés sur l'approche établie par Rao et Yu (1994), Ghosh, Nangia et Kim (1996), Datta, Lahiri, Maiti et Lu (1999), Datta, Lahiri et Maiti (2002), ainsi que You, Rao et Gambino (2000, 2003), permettent d'exploiter simultanément les deux dimensions pour produire des estimations améliorées ayant des propriétés souhaitables pour les petits domaines. En particulier, You et coll. (2000, 2003) ont étudié l'estimation fondée sur un modèle des taux de chômage pour des régions infraprovinciales locales, comme les RMR et les agglomérations de recensement (AR) au Canada. Ils ont obtenu des estimateurs fondés sur un modèle efficaces et un ajustement adéquat des modèles pour l'estimation des taux de chômage d'après l'EPA. Toutefois, le modèle proposé par You et coll. (2000, 2003) présente certaines limites. Dans le présent article, nous discutons de ces limites et proposons un nouveau modèle intégré pour l'estimation des taux de chômage d'après l'EPA sous un cadre hiérarchique bayésien (HB). L'idée est de modéliser ensemble les paramètres d'intérêt et les variances d'échantillonnage, comme l'ont suggéré You et coll. (2003), ainsi que You et Chapman (2006). Nous appliquons le modèle proposé aux données de l'EPA de 2005 et obtenons les estimations des taux de chômage fondés sur un modèle. Nous donnons aussi une comparaison des estimations HB aux estimations directes d'après l'EPA et l'analyse de l'ajustement du modèle.

Le plan de l'article est le suivant. À la section 2, nous présentons divers modèles pour petits domaines proposés dans la littérature pour l'estimation des taux de chômage et nous en discutons. À la section 3, nous examinons le problème du lissage et de la modélisation de la matrice de covariance d'échantillonnage. À la section 4, nous proposons un modèle non linéaire intégré à effets mixtes dans un cadre hiérarchique bayésien et décrivons l'utilisation de l'échantillonnage de Gibbs pour générer des échantillons à partir de la loi conjointe a posteriori. À la section 5, nous appliquons le modèle proposé aux données de l'EPA et obtenons les estimations HB pour les taux de chômage pour petits domaines. Nous donnons aussi l'analyse et l'évaluation du modèle. Enfin, à la section 6, nous présentons certaines conclusions et discutons de l'orientation de futurs travaux.

2. Modèles pour petits domaines

2.1 Modèle transversal

Les modèles transversaux, ou au niveau du domaine, sont utilisés pour produire des estimations fondées sur un modèle fiable par combinaison de l'information auxiliaire au niveau du domaine et des estimations directes au niveau du domaine. Un modèle transversal de base est celui, bien connu, de Fay-Herriot (Fay et Herriot 1979). Il possède deux composantes, à savoir 1) un modèle d'échantillonnage pour les estimations directes par sondage et 2) un modèle de lien qui relie les paramètres du petit domaine à des variables auxiliaires au niveau du domaine grâce à un modèle de régression linéaire. Pour l'estimation mensuelle du taux de chômage d'après l'EPA, dénotons par θ_{it} le taux de chômage réel pour la i^{e} RMR/CU à un point particulier dans le temps (mois) t, où $i = 1, ..., m$, où m est le nombre de RMR/CU, et soit y_{it} l'estimation directe d'après l'EPA de θ_{it}. Alors, le modèle d'échantillonnage pour y_{it} peut être exprimé sous la forme

$$y_{it} = \theta_{it} + e_{it}, \ \ i = 1, \ ..., \ m, \qquad (1)$$

où e_{it} est l'erreur d'échantillonnage associée à l'estimateur direct y_{it}. Nous supposons que cette erreur suit une loi normale telle que $e_{it} \sim N(0, \sigma_{it}^2)$ où σ_{it}^2 est la variance

d'échantillonnage. Le modèle de lien pour le taux de chômage réel θ_{it} peut s'écrire sous la forme

$$\theta_{it} = x'_{it}\beta + v_i, \quad i = 1, \ldots, m, \qquad (2)$$

où x_{it} est la variable auxiliaire et v_i est l'effet aléatoire particulier au domaine. Pour chaque point dans le temps (chaque mois), nous pouvons utiliser le modèle de Fay-Herriot pour produire les estimations directes mensuelles. Ce modèle combine l'information transversale, mais n'est pas renforcé par emprunt d'information recueillie au cours des périodes antérieures.

2.2 Modèle transversal et chronologique

Étant donné le plan d'échantillonnage du processus de renouvellement de l'échantillon de l'EPA, il existe, dans chaque domaine, un chevauchement important d'échantillons sur des périodes de six mois. Par conséquent, pour un domaine donné i, il est nécessaire de prendre en compte la corrélation entre les erreurs d'échantillonnage e_{it} et e_{is} $(t \neq s)$. You et coll. (2000, 2003) ont proposé un modèle transversal et chronologique pour estimer les taux de chômage d'après l'EPA. Ils n'ont utilisé que les données recueillies au cours des six mois précédents pour prédire le taux du mois courant, puisque le renouvellement de l'échantillon de l'EPA est fondé sur un cycle de six mois. Chaque mois, le sixième de l'échantillon de l'EPA est remplacé. Donc, après six mois, la corrélation entre les estimations est faible (voir la section 2.1 pour les coefficients de corrélation retardés). Soit $y_i = (y_{i1}, \ldots, y_{iT})'$, $\theta_i = (\theta_{i1}, \ldots, \theta_{iT})'$, et $e_i = (e_{i1}, \ldots, e_{iT})'$, où $T = 6$ ici. En supposant que e_i suit une loi normale multivariée de vecteur moyenne 0 et de matrice de covariance d'échantillonnage $\sum_i$, nous avons

$$y_i \sim N\left(\theta_i, \sum\nolimits_i\right), \quad i = 1, \ldots, m.$$

Donc nous supposons que y_i est sans biais par rapport au plan pour θ_i. La matrice de covariance d'échantillonnage $\sum_i$ est inconnue dans le modèle. Des estimations directes des matrices de covariance d'échantillonnage sont disponibles. Dans le contexte de l'estimation sur petits domaines fondée sur un modèle au niveau du domaine, on suppose habituellement que la variance d'échantillonnage est connue (Rao 2003). Par exemple, le modèle classique de Fay-Herriot repose sur l'hypothèse que la variance d'échantillonnage est connue dans le modèle. Habituellement, on utilise un estimateur lissé de la variance d'échantillonnage. Toutefois, des progrès récents dans le domaine de la modélisation de la variance d'échantillonnage offrent une autre approche pour aborder le problème; par exemple, voir Wang et Fuller (2003), You et Dick (2004), ainsi que You et Chapman (2006). En ce qui concerne l'estimation des taux de chômage, nous fournissons des précisions sur le lissage et la modélisation des variances d'échantillonnage à la section 3.

Afin d'emprunter de l'information aux diverses régions et périodes, et en nous inspirant de You et coll. (2000, 2003), nous pouvons modéliser le taux réel de chômage θ_{it} par un modèle de régression linéaire avec effets aléatoires au moyen de variables auxiliaires x_{it}, c'est-à-dire

$$\theta_{it} = x'_{it}\beta + v_i + u_{it}, \quad i = 1, \ldots, m, \; t = 1, \ldots, T, \qquad (3)$$

où v_i est un effet aléatoire de domaine que nous supposons suivre une loi $N(0, \sigma_v^2)$ et u_{it} est une composante temporelle (période) et spatiale (domaine) aléatoire. Nous pouvons en outre supposer que u_{it} suit un processus de marche aléatoire sur les périodes $t = 1, \ldots, T$, autrement dit que

$$u_{it} = u_{i,\,t-1} + \varepsilon_{it}, \qquad (4)$$

où $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$. Alors, $\mathrm{cov}(u_{it}, u_{is}) = \min(t, s)\,\sigma_\varepsilon^2$. Le vecteur de régression β et les composantes de la variance σ_v^2 et σ_ε^2 sont inconnus dans le modèle et doivent être estimés. En combinant les modèles (1), (3) et (4), nous obtenons un modèle mixte linéaire avec composantes temporelles de la forme

$$y_{it} = x'_{it}\beta + v_i + u_{it} + e_{it}, \quad i = 1, \ldots, m, \; t = 1, \ldots, T. \qquad (5)$$

You et coll. (2003) ont montré que le modèle transversal et chronologique (5) est supérieur au modèle de Fay-Herriot en ce qui concerne le lissage des estimations directes et la réduction des coefficients de variation des estimations directes pour l'estimation du taux de chômage d'après l'EPA.

Nous avons utilisé un modèle à marche aléatoire pour u_{it}. Rao et Yu (1994) ont utilisé un modèle autorégressif stationnaire pour u_{it}. You et coll. (2003) ont montré que le modèle à marche aléatoire sur u_{it} fournissait un modèle mieux ajusté à l'estimation du taux de chômage que le modèle autorégressif AR(1). Datta et coll. (1999) ont également utilisé un modèle à marche aléatoire pour estimer les taux de chômage au niveau de l'État aux États-Unis.

2.3 Modèle de lien loglinéaire

Cependant, l'une des limites du modèle (3) est due au fait que le modèle de lien pour le paramètre d'intérêt, c'est-à-dire le taux réel de chômage θ_{it}, est un modèle linéaire avec effets aléatoires normaux. Puisque θ_{it} est un nombre positif compris entre 0 et 1, et qu'il est proche de 0, le modèle de lien linéaire avec effets aléatoires normaux pourrait introduire des estimations négatives de θ_{it} pour certains petits domaines. Afin d'éviter ce problème, You, Chen et Gambino (2002) ont proposé un modèle de lien loglinéaire pour θ_{it} de la forme suivante :

$$\log(\theta_{it}) = x'_{it}\beta + v_i + u_{it}, \quad i = 1, \ldots, m, \; t = 1, \ldots, T. \qquad (6)$$

You et Rao (2002) ont également étudié le modèle de lien loglinéaire pour le modèle de Fay-Herriot en tant que modèles non appariés d'échantillonnage et de lien avec application à l'estimation du sous-dénombrement au recensement canadien. Les résultats de You et Rao (2002) et de You et coll. (2002) montrent que le modèle de lien loglinéaire donne de fort bons résultats lorsqu'il est appliqué à des problèmes d'estimation sur petits domaines. Ici, nous utiliserons par conséquent le modèle de lien loglinéaire (6) pour estimer le taux réel de chômage θ_{it}.

3. Variance d'échantillonnage

En général, nous pouvons obtenir des estimations directes de la variance d'échantillonnage d'après les données d'enquête. Toutefois, ces estimations directes sont instables si les tailles d'échantillon sont faibles. Dans les modèles au niveau du domaine pour l'estimation sur petits domaines, on suppose habituellement que les variances d'échantillonnage sont connues (par exemple, Fay et Herriot 1979; Datta et coll. 1999; You et Rao 2002). Sous cette hypothèse, des estimations fiables (lissées) des variances d'échantillonnage sont construites en utilisant d'autres données et modèles auxiliaires, habituellement en se servant de fonctions de variance généralisées (par exemple, Dick 1995; Datta et coll. 1999). Alternativement, dans le présent article, nous modélisons la matrice de variance-covariance d'échantillonnage en nous servant des estimations directes de manière particulière qui rend non nécessaire l'hypothèse que les variances et les covariances d'échantillonnage sont connues dans le modèle. Donc, nous simplifions le problème du lissage de la variance d'échantillonnage inconnue et intégrons la modélisation de la variance d'échantillonnage dans le modèle complet.

3.1 Lissage de la matrice de covariance d'échantillonnage

You et coll. (2000, 2003) ont suivi deux étapes pour lisser la matrice de covariance d'échantillonnage. La première consiste à obtenir un coefficient de variation lissé ou constant en calculant pour chaque RMR/CU le coefficient de variation moyen sur une certaine période, que nous dénotons $\overline{\mathrm{CV}}_i$, où $i = 1, 2, ..., m$. La deuxième étape consiste à obtenir le coefficient de corrélation avec décalage moyen sur le temps et toutes les RMR/CU, dénoté $\overline{\rho}_{|t-s|}$ pour le décalage temporel $|t-s|$. Cette étape requiert d'importantes ressources de calcul. Nous avons utilisé des données de l'EPA couvrant une période de trois années (1999 à 2001) pour calculer les coefficients de corrélation lissés. Dans le modèle, nous traitons les valeurs lissées temporellement et spatialement comme étant les valeurs réelles. Nous obtenons pour le coefficient de corrélation avec décalage d'un mois (décalage -1) la valeur $\overline{\rho}_1 = 0,48$, pour le coefficient de corrélation avec décalage -2, la valeur $\overline{\rho}_2 = 0,31$, pour le décalage -3, la valeur $\overline{\rho}_3 = 0,21$, pour le décalage -4, la valeur $\overline{\rho}_4 = 0,16$, pour le décalage -5, la valeur $\overline{\rho}_5 = 0,11$ et pour le décalage -6, $\overline{\rho}_6 = 0,1$. Après décalage -6, la valeur est inférieure à 0,1. Les coefficients de corrélation avec décalage diminuent à mesure qu'augmente le décalage, ce qui est en harmonie avec le processus de renouvellement de l'échantillon intégré dans le plan de sondage de l'EPA. La figure 1 montre les coefficients de corrélation avec décalage lissés pour les estimations du taux de chômage d'après l'EPA.

En utilisant ces coefficients de variation lissés et les coefficients de corrélation avec décalage, nous pouvons obtenir une matrice de covariance lissée $\hat{\Sigma}_i$ ayant pour éléments diagonaux $\hat{\sigma}_{it}^2 = (\overline{\mathrm{CV}}_i)^2 y_{it}^2$ et pour éléments non diagonaux $\hat{\sigma}_{its} = \overline{\rho}_{|t-s|}\hat{\sigma}_{it}\hat{\sigma}_{is}$. Nous traitons ensuite la matrice lissée $\hat{\Sigma}_i$ comme étant connue dans le modèle. L'étude de You et coll. (2000, 2003) donne à penser que l'utilisation de la matrice lissée $\hat{\Sigma}_i$ dans le modèle peut améliorer considérablement les estimations en ce qui concerne la réduction du CV comparativement aux estimations HB obtenues en utilisant dans le modèle les estimations directes par sondage de Σ_i. Pour plus de renseignements sur le résultat, voir You et coll. (2003).

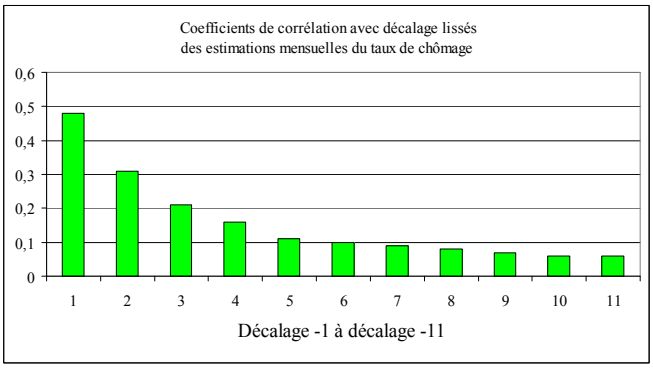

Figure 1 Coefficients de corrélation avec décalage lissés des taux de chômage

3.2 Approche de la modélisation avec CV égaux

Le principal problème de la méthode de You et coll. (2000, 2003) est que les matrices lissées de covariance d'échantillonnage dépendent des estimations directes par sondage y_{it}, alors que ces dernières ne sont pas fiables pour certaines petites régions. Il convient de souligner que la variance d'échantillonnage réelle peut s'écrire sous la forme $\sigma_{it}^2 = \theta_{it}^2(\mathrm{CV}_{it})^2$. Partant de l'hypothèse d'un CV constant au cours du temps pour un domaine donné, You et coll. (2003) ont proposé dans leurs conclusions d'utiliser des estimations de la forme $\tilde{\sigma}_{it}^2 = \theta_{it}^2(\overline{\mathrm{CV}}_i)^2$ et $\tilde{\sigma}_{its} = \overline{\rho}_{|t-s|}(\tilde{\sigma}_{it}\tilde{\sigma}_{is})$ pour les variances et les covariances lissées,

respectivement. Alors, la nouvelle matrice de covariance d'échantillonnage lissée $\tilde{\Sigma}_i$ a pour éléments diagonaux $\tilde{\sigma}_{it}^2$ et pour éléments non diagonaux $\tilde{\sigma}_{its}$. Cependant, sous cette méthode, la matrice de covariance d'échantillonnage $\tilde{\Sigma}_i$ devient inconnue dans le modèle, puisque $\tilde{\sigma}_{it}^2$ et $\tilde{\sigma}_{its}$ dépendent des paramètres inconnus θ_{it}, et que ceux-ci sont reliés à un modèle de lien. L'avantage de cette méthode tient au fait que la structure de la matrice de covariance d'échantillonnage est clairement spécifiée dans le modèle. Elle est meilleure que la méthode de lissage en ce sens que la covariance d'échantillonnage est spécifiée explicitement au lieu d'être considérée comme connue.

3.3 Approche de modélisation avec effets de plan égaux

Une autre approche de modélisation est fondée sur l'hypothèse d'effets de plan constants au cours du temps, comme l'ont proposé Singh, You et Mantel (2005), ainsi que Singh, Folsom et Vaish (2005) pour lisser la variance d'échantillonnage σ_{it}^2. L'effet de plan (deff) pour le i^{e} domaine au temps t peut s'écrire approximativement sous la forme

$$\text{deff}_{it} = \frac{\sigma_{it}^2}{\theta_{it}(1-\theta_{it})/n_{it}},$$

où n_{it} est la taille de l'échantillon correspondant. Alors, la variance d'échantillonnage σ_{it}^2 peut s'écrire sous la forme $\sigma_{it}^2 = \theta_{it}(1-\theta_{it}) \cdot \text{deff}_{it}/n_{it}$. Soit $\tau_{it} = \text{deff}_{it}/n_{it} = \sigma_{it}^2/(\theta_{it}(1-\theta_{it}))$. Alors, nous pouvons estimer τ_{it} en utilisant les estimations directes de θ_{it} et de σ_{it}^2 comme $\hat{\tau}_{it} = \hat{\sigma}_{it}^2/(y_{it}(1-y_{it}))$. Pour chaque domaine, sous l'hypothèse d'un effet de plan constant et d'une taille d'échantillon constante au cours du temps, nous pouvons obtenir un facteur moyen lissé $\bar{\tau}_i$ donné par $\bar{\tau}_i = \sum_{t=1}^{T} \hat{\tau}_{it}/T$. Alors, nous pouvons calculer une variance d'échantillonnage lissée de la forme $\tilde{\sigma}_{it}^2 = \theta_{it}(1-\theta_{it}) \cdot \bar{\tau}_i$, qui de nouveau dépend également de θ_{it}. La covariance d'échantillonnage reste de la forme $\tilde{\sigma}_{its} = \bar{\rho}_{|t-s|}(\tilde{\sigma}_{it}\tilde{\sigma}_{is})$, comme dans You et coll. (2003). Soulignons que $\bar{\tau}_i$ est une moyenne mobile de $\hat{\tau}_{it}$ sur la période T dans le modèle. Toutefois, en pratique, une autre option consiste à utiliser une série chronologique plus longue pour obtenir une estimation plus stable de $\bar{\tau}_i$ pour chaque domaine, au besoin. Ici, nous utiliserons le modèle à effet de plan constant pour l'estimation du taux de chômage basée sur le facteur à moyenne mobile lissé $\bar{\tau}_i$ car nous empruntons de l'information provenant de la période passée T.

4. Inférence bayésienne hiérarchique

À la présente section, nous proposons un modèle loglinéaire transversal et chronologique intégré pour estimer le taux de chômage. Nous appliquons l'approche hiérarchique bayésienne au modèle. Nous obtenons les estimations des moyennes et des variances a posteriori en utilisant la méthode d'échantillonnage de Gibbs.

4.1 Modèle hiérarchique bayésien intégré

Nous proposons maintenant le modèle loglinéaire transversal et chronologique intégré sous un cadre hiérarchique bayésien qui suit :

- Conditionnellement à $\theta_i = (\theta_{i1},...,\theta_{iT})'$, $[y_i|\theta_i] \sim \text{ind } N(\theta_i, \Sigma_i(\theta_i))$;
- Conditionnellement à β, u_{it} et σ_v^2, $[\log(\theta_{it})|\beta, u_{it}, \sigma_v^2] \sim \text{ ind } N(x_{it}'\beta + u_{it}, \sigma_v^2)$;
- Conditionnellement à $u_{i,t-1}$ et σ_ε^2, $[u_{it}|u_{i,t-1}, \sigma_\varepsilon^2] \sim \text{ ind } N(u_{i,t-1}, \sigma_\varepsilon^2)$;
- $\Sigma_i(\theta_i)$ dépend de θ_i avec les éléments diagonaux $\tilde{\sigma}_{it}^2 = \theta_{it}(1-\theta_{it}) \cdot \bar{\tau}_i$ et les éléments non diagonaux $\tilde{\sigma}_{its} = \bar{\rho}_{|t-s|}(\tilde{\sigma}_{it}\tilde{\sigma}_{is})$.
- Marginalement, β, σ_v^2 et σ_ε^2 sont mutuellement indépendants, leurs priors étant donnés par $\beta \propto 1$, $\sigma_v^2 \sim \text{IG}(a_1, b_1)$, et $\sigma_\varepsilon^2 \sim \text{IG}(a_2, b_2)$, où IG dénote une loi gamma inverse et a_1, b_1, a_2, b_2 sont des constantes positives connues auxquelles nous donnons habituellement une valeur très faible pour refléter nos connaissances vagues au sujet de σ_v^2 et σ_ε^2.

Remarques :

1. Nous utilisons dans le modèle HB proposé un modèle de lien loglinéaire pour le paramètre de petit domaine d'intérêt θ_{it}, comme l'ont suggéré You et coll. (2002) et You et Rao (2002).
2. La matrice de covariance d'échantillonnage Σ_i est inconnue dans le modèle et est spécifiée sous la forme d'une fonction du paramètre de petit domaine inconnu θ_i, comme l'ont suggéré You et Rao (2002) et You et coll. (2003).
3. Nous avons utilisé l'hypothèse d'effets de plan constants pour les petits domaines comme l'ont suggéré Singh, You et Mantel (2005).
4. Le modèle HB proposé surmonte les limites du modèle de You et coll. (2000, 2003) en ce qui concerne la modélisation loglinéaire et la spécification de la modélisation de la matrice de covariance d'échantillonnage inconnue. En particulier, nous modélisons cette matrice à l'aide de paramètres de petit domaine θ_i en utilisant les estimations lissées des effets de plan pour chaque domaine.

Nous souhaitons estimer le taux de chômage réel θ_{it} et, en particulier, le taux de chômage courant θ_{iT}. Dans l'analyse HB, θ_{iT} est estimé par sa moyenne a posteriori

$E(\theta_{iT} \mid y)$ et l'incertitude associée à l'estimateur est mesurée par la variance a posteriori $V(\theta_{iT} \mid y)$. Nous utilisons la méthode d'échantillonnage de Gibbs (Gelfand et Smith 1990; Gelman et Rubin 1992) pour obtenir la moyenne et la variance a posteriori de θ_{iT}.

4.2 Inférence sous échantillonnage de Gibbs

La méthode d'échantillonnage de Gibbs est une méthode itérative d'échantillonnage Monte Carlo par chaînes de Markov utilisée pour simuler le tirage d'échantillons à partir d'une loi conjointe de variables aléatoires en procédant à un échantillonnage à partir de densités sous un espace de dimension réduite pour faire des inférences au sujet des lois conjointe et marginale (Gelfand et Smith 1990). Cette méthode est appliquée principalement à l'inférence dans un cadre bayésien où l'on s'intéresse à la loi a posteriori des paramètres. Supposons que la densité conditionnelle de $y_i \mid \theta$ est $f(y_i \mid \theta)$ pour $i = 1, ..., n$ et que l'information a priori au sujet de $\theta = (\theta_1, ..., \theta_k)'$ est résumée par une densité a priori $\pi(\theta)$. Soit $\pi(\theta \mid y)$ la densité a posteriori de θ sachant les données $y = (y_1, ..., y_n)'$. En pratique, il pourrait être difficile de tirer directement des échantillons à partir de $\pi(\theta \mid y)$, à cause de l'intégration en grande dimension par rapport à θ. Cependant, nous pouvons utiliser l'échantillonneur de Gibbs pour construire une chaîne de Markov $\{\theta^{(g)} = (\theta_1^{(g)}, ..., \theta_k^{(g)})'\}$ avec $\pi(\theta \mid y)$ comme limite. À titre d'illustration, posons que $\theta = (\theta_1, \theta_2)'$. En partant d'un ensemble initial de valeurs $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})'$, nous générons $\theta^{(g)} = (\theta_1^{(g)}, \theta_2^{(g)})'$ par échantillonnage de $\theta_1^{(g)}$ à partir de $\pi(\theta_1 \mid \theta_2^{(g-1)}, y)$ et $\theta_2^{(g)}$ à partir de $\pi(\theta_2 \mid \theta_1^{(g-1)}, y)$. Sous certaines conditions de régularité, $\theta^{(g)} = (\theta_1^{(g)}, \theta_2^{(g)})'$ converge en loi vers $\pi(\theta \mid y)$ quand $g \to \infty$. Si la valeur de g est grande, l'inférence marginale au sujet de $\pi(\theta_i \mid y)$ peut être fondée sur les échantillons marginaux $\{\theta_i^{(g+k)}; k = 1, 2, ...\}$.

Pour les modèles HB intégrés proposés, afin d'obtenir l'estimation a posteriori du taux de chômage, nous mettons en œuvre la méthode d'échantillonnage de Gibbs en produisant des échantillons à partir des lois conditionnelles complètes des paramètres β, σ_v^2 et σ_ε^2, u_{it} et θ_i. Ces lois conditionnelles complètes sont données en annexe. Les lois de β, σ_v^2 et σ_ε^2, u_{it} sont des lois normales standard ou des lois gamma inverses qu'il est facile d'échantillonner. Cependant, la loi conditionnelle de θ_i n'a pas de forme explicite. Nous utilisons l'algorithme de Metropolis-Hastings dans l'échantillonneur de Gibbs (Chib et Greenberg 1995) pour mettre à jour θ_i. À l'instar de You et coll. (2002) et de You et Rao (2002), nous pouvons écrire la loi conditionnelle complète de θ_i dans l'échantillonneur de Gibbs sous la forme

$$\theta_i \mid Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u \propto h(\theta_i) f(\theta_i),$$

où

$$h(\theta_i) = \left|\sum_i(\theta_i)\right|^{-1} \exp\{-\frac{1}{2}(y_i - \theta_i)' \sum_i^{-1}(y_i - \theta_i)\}$$

et

$$f(\theta_i) = \exp\left\{-\frac{1}{2\sigma_v^2}(\log(\theta_i) - \mathrm{x}_i'\beta - u_i)'(\log(\theta_i) - \mathrm{x}_i'\beta - u_i)\right\} \cdot \left(\prod_{t=1}^{T} \frac{1}{\theta_{it}}\right).$$

Pour mettre à jour θ_i, nous procédons comme il suit :

1. Pour $t = 1, ..., T$, tirer $\theta_{it}^{(k+1)} \sim \log N(x_{it}'\beta^{(k+1)} + u_{it}^{(k+1)}, \sigma_v^{2(k+1)})$, ce qui nous donne alors $\theta_i^{(k+1)} = (\theta_{i1}^{(k+1)}, ..., \theta_{iT}^{(k+1)})'$.
2. Calculer la probabilité de rejet

$$\alpha(\theta_i^{(k)}, \theta_i^{(k+1)}) = \min\{\frac{h(\theta_i^{(k+1)})}{h(\theta_i^{(k)})}, 1\}.$$

3. Générer $\lambda \sim \text{Uniforme}(0,1)$, si $\lambda < \alpha(\theta_i^{(k)}, \theta_i^{(k+1)})$, alors accepter $\theta_i^{(k+1)}$; sinon, rejeter $\theta_i^{(k+1)}$ et fixer $\theta_i^{(k+1)} = \theta_i^{(k)}$.

Pour appliquer l'échantillonnage de Gibbs, nous suivons les recommandations de Gelman et Rubin (1992) et nous exécutons indépendamment $L (L > 2)$ chaînes parallèles, chacune de longueur $2d$. Nous supprimons les d premières itérations de chaque chaîne. La surveillance de la convergence est fondée sur le facteur de réduction d'échelle possible proposé dans Gelman et Rubin (1992) et adopté par You et coll. (2003) pour estimer θ_{iT}. Une description détaillée figure dans You et coll. (2003). Les estimations de la moyenne *a posteriori* $E(\theta_{iT} \mid y)$ et de la variance a posteriori $V(\theta_{iT} \mid y)$ sont obtenues en se basant sur les échantillons créés à partir de l'échantillonneur de Gibbs.

5. Application aux données de l'EPA

5.1 Estimation

Nous utilisons les estimations du taux de chômage fondées sur les données de l'EPA recueillies de janvier à juin 2005, y_{it}, dans notre analyse de données. En plus des estimations directes y_{it} et des matrices de covariance d'échantillonnage utilisées dans les modèles pour petits domaines, nous avons besoin de variables auxiliaires administratives dans nos modèles. Pour l'estimation du taux de chômage, nous utilisons comme données auxiliaires les taux de bénéficiaires de l'assurance-emploi (a.-e.) au niveau local. Le taux de bénéficiaires est calculé sous forme du ratio du nombre de personnes faisant la demande de prestations d'assurance-emploi sur le nombre de personnes dans la population active. Le Canada compte 72 RMR/CU.

Un CU (Miramichi) ne possède pas de données sur l'assurance-emploi. Donc, nous considérons dans le modèle $m = 71$ RMR/CU. Dans chaque domaine, nous considérons six estimations mensuelles consécutives y_{it} allant de janvier 2005 à juin 2005, de sorte que $T = 6$. Pour les données recueillies de janvier à juin 2005, le taux de chômage moyen global (sur les 71 RMR/CU et les six mois) est de 0,076 et le taux moyen global de bénéficiaires de l'assurance-emploi est de 0,059. Pour le modèle pour petits domaines proposé, le paramètre d'intérêt θ_{iT} est le taux de chômage réel pour le domaine i en juin 2005, où $i = 1, \ldots, 71$. Pour appliquer l'échantillonneur de Gibbs, nous avons utilisés dix exécutions parallèles comptant chacune 2 000 itérations. Nous avons supprimé les 1 000 premières itérations à titre de période de « rodage ». Les hyperparamètres des composantes de variance incluses dans le modèle sont fixés à 0,0001 pour refléter nos connaissances vagues au sujet de σ_v^2 et de σ_e^2.

Nous présentons maintenant les estimations a posteriori des taux de chômage sous le modèle HB intégré proposé décrit à la section 4.1 en utilisant la méthode d'échantillonnage de Gibbs. La figure 2 donne les estimations directes d'après l'EPA et les estimations fondées sur le modèle HB des taux de chômage en juin 2005 pour les 71 RMR/CU au Canada. Ces 71 RMR/CU figurent par ordre de taille de population, en commençant à gauche par le plus petit CU (Dawson Creek, C.-B.) et en terminant à droite par la plus grande RMR (Toronto, Ont.). Dans le cas des estimations ponctuelles, les estimations HB correspondent à un lissage modéré des estimations directes d'après l'EPA. Pour les RMR ayant les populations les plus grandes et, par conséquent, les tailles d'échantillon les plus grandes, les estimations directes et les estimations HB sont très proches, comme il fallait s'y attendre, particulièrement pour Toronto, Montréal et Vancouver; pour les CU plus petits, l'écart entre les estimations directes et HB est important pour certaines régions.

La figure 3 donne les CV des estimations. Le CV de l'estimation HB correspond au ratio de la racine carrée de la variance a posteriori à la moyenne a posteriori. Il est manifeste, si l'on examine la figure 3, que les CV des estimations directes sont très grands, particulièrement pour les CU dont les CV sont très grands et instables. Comparativement aux estimations directes, les CV des estimations HB sont très petits et stables. Le gain d'efficacité obtenu grâce aux estimations HB est évident, particulièrement pour les CU dont la taille de population est faible. Plus précisément, nous avons calculé la réduction en pourcentage des CV pour les estimateurs HB en nous basant sur les données de juin 2005. La réduction en pourcentage du CV est calculée comme la différence entre le CV direct et le CV HB par rapport au CV direct. La réduction moyenne du CV est de 63 % pour les CU et de 35 % pour les RMR. Comme prévu, le modèle proposé a permis de réduire considérablement le CV comparativement aux estimations directes, particulièrement pour les petits CU dont la taille d'échantillon est faible.

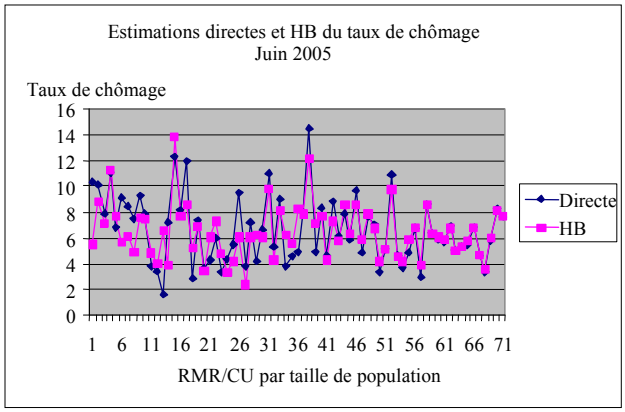

Figure 2 Comparaison des estimations directes et HB

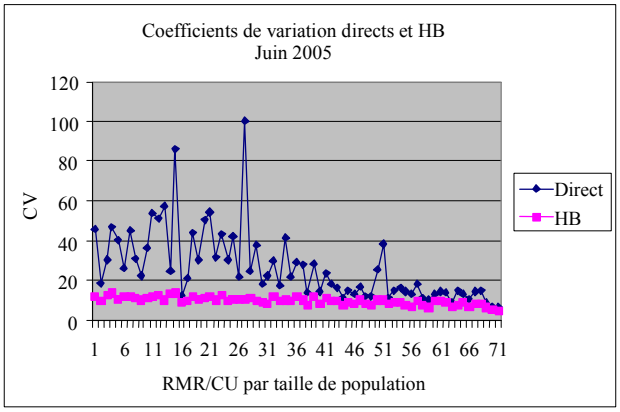

Figure 3 Comparaison des CV directs et HB

5.2 Ajustement du modèle au moyen de la loi prédictive a posteriori

Pour vérifier l'ajustement global du modèle proposé, nous utilisons la méthode de la loi prédictive a posteriori. Soit y_{rep} l'observation répétée sous le modèle. La loi prédictive a posteriori de y_{rep} sachant les données observées y_{obs} est définie comme étant

$$f(y_{\text{rep}} \mid y_{\text{obs}}) = \int f(y_{\text{rep}} \mid \theta) f(\theta \mid y_{\text{obs}}) d\theta.$$

Selon cette approche, nous pouvons définir une mesure de divergence $D(y, \theta)$ qui dépend des données y et du paramètre θ, et comparer la valeur observée $D(y_{\text{obs}}, \theta \mid y_{\text{obs}})$ à la loi prédictive a posteriori de $D(y_{\text{rep}}, \theta \mid y_{\text{obs}})$, tout écart significatif indiquant une défaillance du modèle. Meng (1994), ainsi que Gelman, Carlin, Stern et Rubin (1995) ont proposé la valeur p prédictive a posteriori de la forme

$$p = P(D(y_{\text{rep}}, \theta) \geq D(y_{\text{obs}}, \theta) \mid y_{\text{obs}}).$$

Il s'agit d'une extension naturelle de la valeur p habituelle dans un contexte bayésien. Si un modèle est ajusté aux données observées, les deux valeurs de la mesure de divergence sont semblables. Autrement dit, si le modèle est ajusté correctement aux données observées, $D(y_{\text{obs}}, \theta \mid y_{\text{obs}})$ devrait se situer près de la partie centrale de l'histogramme des valeurs $D(y_{\text{rep}}, \theta \mid y_{\text{obs}})$ si y_{rep} est généré de façon répétée à partir de la loi prédictive a posteriori. Conséquemment, la valeur p prédictive a posteriori devrait s'approcher de 0,5 si le modèle est bien ajusté aux données. Les valeurs p extrêmes (proches de 0 ou de 1) impliquent que l'ajustement n'est pas bon. Nous pouvons estimer la valeur p prédictive a posteriori de la façon suivante. Soit θ^* un tirage à partir de la loi a posteriori $f(\theta \mid y_{\text{obs}})$, et soit y^*_{rep} un tirage à partir de $f(y_{\text{rep}} \mid \theta^*)$. Alors, marginalement, y^*_{rep} est un échantillon provenant de la loi prédictive a posteriori $f(y_{\text{rep}} \mid y_{\text{obs}})$. Il est relativement facile de calculer la valeur p en utilisant les valeurs simulées de θ^* provenant de l'échantillonneur de Gibbs. Pour chaque valeur simulée θ^*, nous pouvons simuler y^*_{rep} à l'aide du modèle et calculer $D(y^*_{\text{rep}}, \theta^*)$ et $D(y_{\text{obs}}, \theta^*)$. Alors, la valeur p est estimée par la proportion de fois que $D(y^*_{\text{rep}}, \theta^*)$ excède $D(y_{\text{obs}}, \theta^*)$.

Pour le modèle HB proposé, la mesure de divergence utilisée pour l'ajustement global est donnée par $d(y, \theta) = \sum_{i=1}^{m} (y_i - \theta_i)' \sum_i^{-1} (y_i - \theta_i)$. Cette mesure a été utilisée par Datta et coll. (1999) et par You et coll. (2003). Nous avons calculé la valeur p en combinant les valeurs simulées de θ^* et y^* provenant des dix exécutions parallèles. Nous avons obtenu une valeur p moyenne estimée d'environ 0,38. Donc, rien n'indique que l'ajustement global du modèle est insuffisant.

Certains ont critiqué la vérification de l'ajustement du modèle à l'aide de la valeur p prédictive a posteriori, la jugeant modérée à cause du double usage des données observées. La double utilisation des données peut induire un comportement non naturel, comme l'ont démontré Bayarri et Berger (2000). Ces auteurs ont proposé deux mesures de la valeur p de rechange pour la vérification du modèle qu'ils ont nommées la valeur p prédictive a posteriori partielle et la valeur p prédictive conditionnelle. Cependant, leurs méthodes sont plus difficiles à appliquer et à interpréter (Rao 2002; Sinharay et Stern 2003). Comme le soulignent Sinharay et Stern (2003), la valeur p prédictive a posteriori est particulièrement utile si nous voyons dans le modèle courant un point final plausible auquel des modifications ne seront apportées que si un manque considérable d'ajustement est constaté.

Pour comparer le modèle proposé au modèle de You et coll. (2003), nous avons calculé la mesure de divergence de Laud et Ibrahim (1995) fondée sur la loi prédictive a posteriori. La mesure de divergence attendue de Laud et Ibrahim (1995) est donnée par $d(y^*, y_{\text{obs}}) = E(k^{-1} \| y^* - y_{\text{obs}} \|^2 \mid y_{\text{obs}})$, où k est la dimension de y_{obs} et y^* est un échantillon tiré de la loi prédictive a posteriori $f(y \mid y_{\text{obs}})$. De deux modèles, nous préférons celui qui donne la valeur la plus faible de cette mesure. Comme Datta, Day et Maiti (1998) et You et coll. (2003), nous avons utilisé, pour approximer la mesure de divergence $d(y^*, y_{\text{obs}})$, les échantillons simulés à partir de la loi prédictive a posteriori. En utilisant les sorties multiples de l'échantillonnage de Gibbs, nous avons obtenu une mesure de divergence de l'ordre de 8 à 9 pour le modèle proposé et d'environ 12 à 14 pour le modèle de You et coll. (2003). Donc, la mesure de divergence donne à penser que le modèle HB intégré proposé pour l'estimation du taux de chômage d'après l'EPA est mieux ajusté.

5.3 Diagnostic du biais par analyse de régression

Afin d'évaluer le biais que pourrait introduire le modèle, nous utilisons une méthode simple d'analyse de régression par les moindres carrés ordinaires pour les estimations directes d'après l'EPA ainsi que pour les estimations fondées sur le modèle HB. La méthode de régression est proposée par Brown, Chamber, Heady et Heasman (2001). Si les estimations fondées sur le modèle sont proches des taux réels de chômage, les estimateurs directs d'après l'EPA devraient se comporter comme des variables aléatoires dont les valeurs prévues correspondent aux valeurs des estimations fondées sur le modèle. Nous représentons graphiquement les estimations HB fondées sur le modèle en abscisse X et les estimations directes d'après l'EPA en ordonnée Y et nous voyons dans quelle mesure la droite de régression s'approche de $Y = X$. En termes de régression, fondamentalement, nous ajustons le modèle de régression $Y = \alpha X$ aux données et estimons le coefficient α. Des estimations fondées sur le modèle dont le biais est faible devraient donner une valeur de α proche de 1. Pour les données de juin 2005, soit Y les estimations directes du taux de chômage et X les estimations HB fondées sur le modèle. Nous obtenons une estimation de la valeur de α de 1,0207 avec une erreur-type de 0,0281. La figure 4 donne le diagramme de dispersion avec la droite de régression ajustée.

Les résultats de la régression ne révèlent aucun écart significatif par rapport à $Y = X$. Par conséquent, nous concluons que les estimations fondées sur le modèle calculées d'après le modèle proposé concordent avec les estimations directes d'après l'EPA sans inclusion d'un biais éventuel supplémentaire. Les résultats pourraient également indiquer qu'il n'existe aucune preuve d'un biais dû à une erreur de spécification possible du modèle.

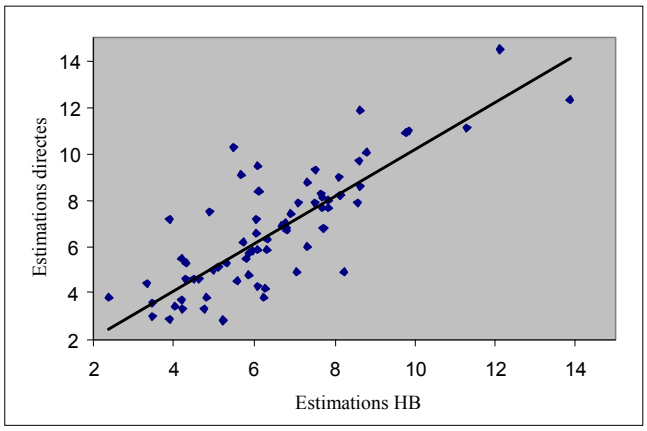

Figure 4 Diagramme de dispersion avec droite de régression

6. Conclusion et futurs travaux

Nous avons passé en revue certains modèles pour petits domaines, y compris le modèle de Fay-Herriot et le modèle transversal et chronologique de You et coll. (2003). Étant donné les travaux antérieurs, nous avons proposé un modèle non linéaire transversal et chronologique intégré en vue d'obtenir des estimations fondées sur un modèle des taux de chômage pour les RMR/CU au Canada en utilisant les données de l'EPA. Le modèle proposé surmonte les limites constatées lors des travaux antérieurs. En particulier, nous pouvons modéliser la variance d'échantillonnage sous forme d'une fonction de la moyenne de petits domaines en émettant l'hypothèse d'un coefficient de variation constant pour un domaine donné ou d'un effet de plan constant pour un domaine donné. Notre analyse des données révèle que le modèle proposé est relativement bien ajusté aux données. Les estimations hiérarchiques bayésiennes fondées sur le modèle améliorent significativement les estimations directes par sondage en ce qui concerne la réduction du coefficient de variation, surtout pour les CU dont la population est faible.

Nous prévoyons adopter l'approche de modélisation de rechange pour la variance d'échantillonnage. Récemment, You et Dick (2004), ainsi que You et Chapman (2006) ont utilisé l'approche HB pour modéliser directement la variance d'échantillonnage, sans spécifier la forme de cette dernière sous le cadre du modèle de Fay-Herriot. Le modèle tient compte automatiquement de la variabilité de l'estimation des variances d'échantillonnage. En particulier, You et Dick (2004) ont appliqué le modèle au problème de l'estimation du sous-dénombrement au recensement et ont obtenu des estimations HB efficaces de ce sous-dénombrement pour divers petits domaines au Canada. Il serait intéressant d'adapter la même idée au modèle transversal et chronologique et de comparer les résultats à ceux des présents travaux. L'objectif de la comparaison est d'établir un modèle fiable et facile à mettre en œuvre de l'estimation des taux de chômage fondés sur un modèle ajusté aux données de l'EPA pour les petits domaines.

Nous prévoyons produire les estimations fondées sur le modèle pour une période relativement longue, par exemple 24 mois allant de 2004 à 2005. Nous comparerons ces estimations aux estimations directes pour les 24 mois, particulièrement pour les grandes RMR, afin d'étudier les effets de lissage du modèle proposé. Les estimations fondées sur le modèle devraient suivre le profil des estimations directes d'après l'EPA pour les grandes RMR, ce qui indiquerait que le lissage des effets de la série temporelle sont raisonnables. Le but est de vérifier la robustesse des estimations fondées sur le modèle proposé au fil du temps.

Annexe

Nous présentons ci-après les lois conditionnelles complètes pour l'échantillonneur de Gibbs sous le modèle HB proposé. Soit $Y = (Y_1', ..., Y_m')'$, $X = (X_1', ..., X_m')'$, $\theta = (\theta_1', ..., \theta_m')'$ et $u = (u_1', ..., u_m')'$, avec $Y_i' = (y_{i1}, ..., y_{iT})$, $X_i' = (x_{i1}, ..., x_{iT})$, $\theta_i' = (\theta_{i1}, ..., \theta_{iT})$, et $u_i' = (u_{i1}, ..., u_{iT})$. Nous obtenons les lois conditionnelles complètes comme il suit :

- $\beta \mid Y, \sigma_v^2, \sigma_\varepsilon^2, u, \theta \sim$
  $N((X'X)^{-1}X'(\log(\theta) - u), \sigma_v^2 (X'X)^{-1})$;

- $\sigma_v^2 \mid Y, \beta, \sigma_\varepsilon^2, u, \theta \sim$
  $\mathrm{IG}\left(\left(a_1 + mT/2, b_1 + \sum_{i=1}^{m}\sum_{t=1}^{T}(\log(\theta_{it}) - x_{it}'\beta - u_{it})^2\right)/2\right)$;

- $\sigma_\varepsilon^2 \mid Y, \beta, \sigma_v^2, u, \theta \sim$
  $\mathrm{IG}\left(\left(a_2 + m(T-1)/2, b_2 + \sum_{i=1}^{m}\sum_{t=2}^{T}(u_{it} - u_{i,t-1})^2\right)/2\right)$;

- Pour $i = 1, ..., m$,
  $u_{i1} \mid Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i2}, \theta \sim$
  $N\left(\left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1}\left(\frac{\log(\theta_{i1}) - x_{i1}'\beta}{\sigma_v^2} + \frac{u_{i2}}{\sigma_\varepsilon^2}\right), \left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1}\right)$;

- Pour $i = 1, ..., m$, et $2 \le t \le T-1$,
  $u_{it} \mid Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i,t-1}, u_{i,t+1}, \theta \sim$
  $N\left(\left(\frac{1}{\sigma_v^2} + \frac{2}{\sigma_\varepsilon^2}\right)^{-1}\left(\frac{\log(\theta_{it}) - x_{it}'\beta}{\sigma_v^2} + \frac{u_{i,t-1} + u_{i,t+1}}{\sigma_\varepsilon^2}\right),\right.$
  $\left.\left(\frac{1}{\sigma_v^2} + \frac{2}{\sigma_\varepsilon^2}\right)^{-1}\right)$;

- Pour $i = 1, ..., m$,

$$u_{iT} \mid Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i,T-1}, \theta \sim N\left(\left(\frac{1}{\sigma_v^2}+\frac{1}{\sigma_\varepsilon^2}\right)^{-1}\left(\frac{\log(\theta_{iT})-x_{iT}'\beta}{\sigma_v^2}+\frac{u_{i,T-1}}{\sigma_\varepsilon^2}\right), \left(\frac{1}{\sigma_v^2}+\frac{1}{\sigma_\varepsilon^2}\right)^{-1}\right);$$

- Pour $i = 1, ..., m$,

$$\theta_i \mid Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u \propto \left|\sum_i\right|^{-1/2} \exp\left\{-\frac{1}{2}(y_i-\theta_i)'\sum_i^{-1}(y_i-\theta_i)\right\} \times \exp\left\{-\frac{1}{2\sigma_v^2}\sum_{t=1}^{T}(\log(\theta_{it})-x_{it}'\beta-u_{it})^2\right\}\left(\prod_{t=1}^{T}\frac{1}{\theta_{it}}\right).$$

Remerciements

L'auteur remercie le rédacteur en chef, le rédacteur adjoint et un examinateur de leurs commentaires et suggestions. Les travaux ont été financés en partie par le fonds global de financement de la recherche de la Direction de la méthodologie de Statistique Canada.

Bibliographie

Bayarri, M.J., et Berger, J.O. (2000). *P* values for composite null models. *Journal of the American Statistical Association*, 95, 1127-1142.

Brown, G., Chambers, R., Heady, P. et Heasman, D. (2001). Evaluation of small area estimation methods - An application to unemployment estimates from the UK LFS. *Proceedings of Statistics Canada Symposium 2001 Achieving Data Quality in a Statistical Agency*: *A Methodological Perspective*, CD-ROM.

Chib, S., et Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49, 327-335.

Datta, G.S., Day, B. et Maiti, T. (1998). Multivariate Bayesian small area estimation: An application to survey and satellite data. *Sankhyā*, 60, 344-362.

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.

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, 83-97.

Dick, P. (1995). Modélisation du sous-dénombrement net dans le recensement du Canada de 1991. *Techniques d'enquête*, 21, 51-61.

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.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. et Lindeyer, J. (1998). *Methodology of the Canadian Labour Force Survey*, Statistique Canada, Catalogue No. 71-526.

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., Carlin, J.B., Stern, H.S. et Rubin, D.B. (1995). *Bayesian Data Analysis.* Londres : Chapman and Hall.

Gelman, A., et Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.

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, 1423-1431.

Laud, P., et Ibrahim, J. (1995). Predictive model selection. *Journal of Royal Statistical Society*, Séries B, 57, 247-262.

Meng, X.L. (1994). Posterior predictive *p* value. *The Annals of Statistics*, 22, 1142-1160.

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.

Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.

Singh, A.C., Folsom, R.E., Jr. et Vaish, A.K. (2005). Small area modeling for survey data with smoothed error covariance structure via generalized design effects. *Federal Committee on Statistical methods Conference proceedings*, *Washington*, *D.C.*, *www.fcsm.gov*.

Singh, A., You, Y. et Mantel, H. (2005). Use of generalized design effects for variance function modeling in small area estimation from survey data. Présentation au 2005 Statistical Society of Canada Annual Meeting, Regina, SK.

Sinharay, S., et Stern, H.S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, 111, 209-221.

Wang, J., et Fuller, W.A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y., et Chapman, B. (2006). Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage. *Techniques d'enquête*, 32, 107-114.

You. Y., Chen, E. et Gambino, G. (2002). Nonlinear mixed effects cross-sectional and time series models for unemployment rate estimation. *2002 Proceedings of the American Statistical Association, Section on Government Statistics*, Alexandria, VA : American Statistical Association. 3883-3888.

You, Y., et Dick, P. (2004). Hierarchical Bayes small area inference to the 2001 census undercoverage estimation. *2004 Proceedings of the American Statistical Association*, *Section on Government Statistics*, Alexandria, VA : American Statistical Association, 1836-1840.

You, Y., et Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 3-15.

You, Y., Rao, J.N.K. et Gambino, J. (2000). Hierarchical Bayes estimation of unemployment rates for sub-provincial regions using cross-sectional and time series data. *American Statistical Association 2000 Proceedings of the Section on Government Statistics and Section on Social Statistics*, 160-165.

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, 27-36.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2008
Vol. 34, N° 1, pp. 33-40
Statistique Canada, N° 12-001-X au catalogue

Estimation pour petits domaines sous une contrainte

Junyuan Wang, Wayne A. Fuller et Yongming Qu [1]

Résumé

La prédiction sur petits domaines fondée sur des effets aléatoires, appelée (MPLSBE), est une méthode de construction d'estimations pour de petites régions géographiques ou de petites sous-populations en utilisant les données d'enquête existantes. Souvent, le total des prédicteurs sur petits domaines est forcé d'être égal à l'estimation par sondage directe et ces prédicteurs sont alors dits calés. Nous passons en revue plusieurs prédicteurs calés et présentons un critère qui unifie leur calcul. Nous dérivons celui qui est l'unique meilleur prédicteur linéaire sans biais sous ce critère et discutons de l'erreur quadratique moyenne des prédicteurs calés. L'imposition de la contrainte comporte implicitement la possibilité que le modèle de petit domaine soit spécifié incorrectement et que les prédicteurs présentent un biais. Nous étudions des modèles augmentés contenant une variable explicative supplémentaire pour lesquels les prédicteurs sur petits domaines ordinaires présentent la propriété d'autocalage. Nous démontrons à l'aide de simulations que les prédicteurs calés ont un biais un peu plus faible que le prédicteur MPLSBE habituel. Cependant, si le biais est une préoccupation, une meilleure approche consiste à utiliser un modèle augmenté contenant une variable auxiliaire supplémentaire qui est fonction de la taille du domaine. Dans les simulations, les prédicteurs fondés sur le modèle augmenté ont une EQM plus petite que MPLSBE quand le modèle incorrect est utilisé pour la prédiction. De surcroît, l'EQM augmente très légèrement comparativement à celle de MPLSBE si la variable auxiliaire est ajoutée au modèle correct.

Mots clés : Modèle des composantes de la variance; meilleure prédiction linéaire sans biais; calage; convergence sous le plan; modèles linéaires mixtes.

1. Introduction

Dans certaines situations, il est souhaitable d'établir d'après les données d'enquête existantes des estimateurs fiables pour de petites régions géographiques ou de petites sous-populations. Cependant, les tailles des échantillons de domaine peuvent être telles que les estimateurs habituels produisent des erreurs-types inacceptablement grandes et il est alors raisonnable d'utiliser un estimateur fondé sur un modèle. Voir Rao (2003) pour une discussion complète de l'estimation pour petits domaines.

Un modèle pour l'estimation pour petits domaines peut s'écrire

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + b_i, \tag{1}$$

$$Y_i = y_i + e_i, \qquad i = 1, \ldots, n, \tag{2}$$

où les y_i sont les moyennes inobservables de petits domaines, les Y_i sont les estimateurs d'après les données d'enquête observables, les $\boldsymbol{x}_i'$ sont des vecteurs connus, $\boldsymbol{\beta}$ est le vecteur des paramètres de régression, les b_i sont des variables aléatoires indépendantes et de même loi telles que $E(b_i) = 0$ et $V(b_i) = \sigma_b^2$, et les e_i sont les erreurs d'échantillonnage telles que $E(e_i \mid y_i) = 0$ et $V(e_i \mid y_i) = \sigma_{ei}^2$. En combinant (1) et (2), nous obtenons

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + b_i + e_i, \qquad i = 1, \ldots, n, \tag{3}$$

qui est un cas particulier du modèle linéaire mixte.

En supposant que les composantes de la variante σ_b^2 et σ_{ei}^2 sont connues, le meilleur estimateur linéaire sans biais de $\boldsymbol{\beta}$ est

$$\hat{\boldsymbol{\beta}} = \left[\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right]^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$$

$$= \left[\sum_{i=1}^{n}(\sigma_b^2+\sigma_{ei}^2)^{-1}\boldsymbol{x}_i\,\boldsymbol{x}_i'\right]^{-1}\left[\sum_{i=1}^{n}(\sigma_b^2+\sigma_{ei}^2)^{-1}\boldsymbol{x}_i\,Y_i\right], \tag{4}$$

où $\boldsymbol{X}' = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{Y}' = (Y_1, \ldots, Y_n)$ et $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{Y}) = \mathrm{diag}(\sigma_b^2 + \sigma_{e1}^2, \ldots, \sigma_b^2 + \sigma_{en}^2)$. En outre, le meilleur prédicteur linéaire sans biais (MPLSB) de y_i est

$$\tilde{y}_i^H = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + \gamma_i(Y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}), \tag{5}$$

où

$$\gamma_i = (\sigma_b^2 + \sigma_{ei}^2)^{-1}\sigma_b^2. \tag{6}$$

Voir Henderson (1963) et Rao (2003). Quand les composantes de la variance sont inconnues, nous les remplaçons dans (4) et (6) par des estimateurs pour obtenir $\hat{y}_i^H$, c'est-à-dire le MPLSB empirique ou MPLSBE.

Souvent, l'estimateur direct du total de tous les domaines étudiés est considéré comme étant d'une précision adéquate. Si cela est le cas, le praticien pourrait choisir d'utiliser l'estimateur du total convergent sous le plan et exiger que la somme pondérée des prédicteurs sur petits domaines soit égale à cet estimateur. Donc, il est désirable que les prédicteurs sur petits domaines $\hat{y}_i$ satisfassent

1. Junyuan Wang, Directeur de Biostatistics, Global Biostatistics and Programming, Wyeth Research, 35 Cambridge Park Drive, Cambridge, MA 02140; Wayne A. Fuller, Professeur, Department of Statistics, Iowa State University, Ames, IA 50011; Yongming Qu, chercheur scientifique principal, Eli Lilly and Company, Lilly Corporation Center, Indianapolis, IN 26285.

$$\sum_{i=1}^{n} \omega_i \hat{y}_i = \sum_{i=1}^{n} \omega_i Y_i, \tag{7}$$

où les ω_i sont les poids d'échantillonnage, tels que $\sum_{i=1}^{n} \omega_i Y_i$ est un estimateur du total (ou de la moyenne) convergent sous le plan. Un certain nombre de méthodes, souvent appelées « étalonnage » ou « calage », ont été proposées pour construire des prédicteurs qui satisfont (7). Voir, par exemple, Mantel, Singh et Barcau (1993) et You et Rao (2003).

Afin de passer ce genre de méthodes en revue, représentons par $\tilde{\boldsymbol{y}}^H = (\tilde{y}_1^H, ..., \tilde{y}_n^H)'$ le prédicteur MPLSB de $\boldsymbol{y} = (y_1, ..., y_n)'$ défini en (5), où

$$\tilde{\boldsymbol{y}}^H = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{b}}, \tag{8}$$

$\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{b}}$ représentant toute solution de

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1} \\ \boldsymbol{\Sigma}_e^{-1}\boldsymbol{X} & \boldsymbol{\Sigma}_e^{-1} + \boldsymbol{\Sigma}_b^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\mathbf{Y} \\ \boldsymbol{\Sigma}_e^{-1}\mathbf{Y} \end{bmatrix}, \tag{9}$$

$\boldsymbol{\Sigma}_b = \sigma_b^2 \boldsymbol{I}_n$, et $\boldsymbol{\Sigma}_e = \text{diag}(\sigma_{e1}^2, ..., \sigma_{en}^2)$. L'équation (9) est appelée l'équation du modèle mixte. Trouver une solution de cette équation équivaut à trouver une solution au problème de minimisation

$$\min_{\boldsymbol{\beta},\boldsymbol{b}} \{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{b})'\boldsymbol{\Sigma}_e^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{b}) + \boldsymbol{b}'\boldsymbol{\Sigma}_b^{-1}\boldsymbol{b}\}. \tag{10}$$

Pfeffermann et Barnard (1991) ont proposé le prédicteur modifié

$$\hat{\boldsymbol{y}}^{\text{PB}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{PB}} + \hat{\boldsymbol{b}}^{\text{PB}}, \tag{11}$$

où $\hat{\boldsymbol{\beta}}^{\text{PB}}$ et $\hat{\boldsymbol{b}}^{\text{PB}}$ représentent toute solution du problème de minimisation (10) avec $\boldsymbol{\beta}$ et $\boldsymbol{b}$ soumis à la contrainte

$$\sum_{i=1}^{n} \omega_i(\boldsymbol{x}_i'\boldsymbol{\beta} + b_i) = \sum_{i=1}^{n} \omega_i Y_i. \tag{12}$$

Cela mène au prédicteur

$$\hat{y}_i^{\text{PB}} = \tilde{y}_i^H + [\text{Var}(\tilde{y}_\cdot)]^{-1}\text{cov}(\tilde{y}_i^H, \tilde{y}_\cdot)\left[\sum_{j=1}^{n} \omega_j Y_j - \tilde{y}_\cdot\right], \tag{13}$$

où $\tilde{y}_\cdot = \sum_{i=1}^{n} \omega_i \tilde{y}_i^H$, $\text{cov}(\tilde{y}_i^H, \tilde{y}_\cdot) = \omega_i \gamma_i \sigma_{ei}^2 + \sum_{j=1}^{n} \omega_j (1 - \gamma_i)(1 - \gamma_j) \boldsymbol{x}_i' V(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_j$, et $\text{Var}(\tilde{y}_\cdot) = \sum_{i=1}^{n} \omega_i \text{cov}(\tilde{y}_i^H, \tilde{y}_\cdot)$.

Isaki, Tsay et Fuller (2000) ont imposé la contrainte par une méthode qui, approximativement, consiste à construire les meilleurs prédicteurs de $n - 1$ quantités qui ne sont pas corrélées avec $\sum_{i=1}^{n} \omega_i Y_i$. Après certaines opérations matricielles, le prédicteur d'Isaki-Tsay-Fuller (ITF) peut être réécrit sous la forme

$$\hat{y}_i^{\text{ITF}} = \hat{y}_i^H + \left[\sum_{j=1}^{n} \omega_j^2 \widehat{\text{Var}}(Y_j)\right]^{-1} \omega_i \widehat{\text{Var}}(Y_i)\left(\sum_{j=1}^{n} \omega_j Y_j - \sum_{j=1}^{n} \omega_j \hat{y}_j^H\right), \tag{14}$$

où $\widehat{\text{Var}}(Y_i)$ est un estimateur de $\sigma_b^2 + \sigma_{ei}^2$.

Soulignons que le prédicteur de Pfeffermann-Barnard (PB) (13) et le prédicteur ITF (14) sont de la forme

$$\hat{y}_i^{\boldsymbol{a}} = \hat{y}_i + a_i\left(\sum_{j=1}^{n} \omega_j Y_j - \sum_{j=1}^{n} \omega_j \hat{y}_j\right), \tag{15}$$

où $\sum_{i=1}^{n} \omega_i a_i = 1$. Autrement dit, nous pouvons considérer qu'imposer la contrainte (7) revient à un problème d'ajustement. Pour qu'un prédicteur ajusté $\hat{y}_i^{\boldsymbol{a}}$ satisfasse (7), nous allouons la différence $\sum_{j=1}^{n} \omega_j Y_j - \sum_{j=1}^{n} \omega_j \hat{y}_j$ aux prédicteurs sur petits domaines $\hat{y}_i$ en utilisant a_i.

À l'aide du modèle au niveau des unités, You et Rao (2002) ont proposé un estimateur de $\boldsymbol{\beta}$ tel que les prédicteurs résultants satisfont (7). Ils ont dit que ces prédicteurs étaient autocalés. En appliquant leur méthode au modèle au niveau du domaine (3), nous obtenons

$$\hat{y}_i^{\text{YR}} = \hat{\gamma}_i Y_i + (1 - \hat{\gamma}_i)\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\text{YR}}, \tag{16}$$

où

$$\hat{\boldsymbol{\beta}}_{\text{YR}} = \left[\sum_{i=1}^{n} \omega_i(1 - \hat{\gamma}_i)\boldsymbol{x}_i\boldsymbol{x}_i'\right]^{-1} \sum_{i=1}^{n} \omega_i(1 - \hat{\gamma}_i)\boldsymbol{x}_i Y_i. \tag{17}$$

Tout prédicteur qui a la propriété d'autocalage, tel que le prédicteur (16) de You et Rao (YR), est un prédicteur de la forme (15), puisque la différence $\sum_{j=1}^{n} \omega_j Y_j - \sum_{j=1}^{n} \omega_j \hat{y}_j$ est égale à zéro.

À la section 2, nous dérivons le « meilleur » prédicteur de la forme (15). Les résultats conduisent à une vision unifiée de plusieurs prédicteurs fondés sur le MPLSB. À la section 3, nous proposons une autre approche qui possède la propriété d'autocalage. À la section 4, nous décrivons brièvement l'erreur quadratique moyenne (EQM) et à la section 5, nous recourons à des études par simulation pour comparer les prédicteurs. À la section 6, nous présentons nos conclusions et une discussion.

2. « Meilleur » prédicteur linéaire sans biais sous une contrainte

Afin de trouver le « meilleur » prédicteur linéaire sans biais de $\boldsymbol{y}$ qui satisfait la contrainte (7), nous commençons par supposer que les paramètres pour les composantes de la variance sont connus. Selon le lemme 1 de Pfeffermann et Barnard (1991), il est impossible de comparer, composante par composante, des prédicteurs qui satisfont la contrainte (7) pour trouver le meilleur d'entre eux. Par conséquent, un certain critère global est requis. L'un des critères naturels est

$$Q(\hat{\boldsymbol{y}}^{\boldsymbol{a}}) = \sum_{i=1}^{n} \varphi_i E(\hat{y}_i^{\boldsymbol{a}} - y_i)^2, \tag{18}$$

où les φ_i, $i = 1, ..., n$ sont un ensemble choisi de poids positifs.

Théorème 1. *Soit le modèle à effets aléatoires*

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + b_i + e_i, \qquad i = 1, ..., n,$$

où les b_i *sont indépendants et de même loi de moyenne nulle et de variance* σ_b^2, *les* e_i *sont indépendants et de loi de moyenne nulle et de variance* σ_{ei}^2, *et* $\boldsymbol{b} = (b_1, ..., b_n)'$ *est indépendant de* $\boldsymbol{e} = (e_1, ..., e_n)'$. *Supposons que* σ_b^2 *et* σ_{ei}^2 *sont connues, et que* $\boldsymbol{\beta}$ *est inconnu. Soit* $\tilde{y}_i^H$ *le MPLSB de* y_i *défini en* (5). *Soit*

$$\hat{y}_i^{\breve{\boldsymbol{a}}} = \tilde{y}_i^H + \breve{a}_i\left(\sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \tilde{y}_j^H\right), \qquad (19)$$

où $\breve{a}_i = (\sum_{i=1}^n \varphi_i^{-1}\omega_i^2)^{-1}\varphi_i^{-1}\omega_i$ *et* ω_i *sont les poids fixes de* (7). *Alors* $\hat{\boldsymbol{y}}^{\breve{\boldsymbol{a}}} = (\hat{y}_1^{\breve{\boldsymbol{a}}}, ..., \hat{y}_n^{\breve{\boldsymbol{a}}})'$ *est, parmi tous les prédicteurs linéaires sans biais, le seul qui satisfait* (7) *et minimise le critère* (18).

Preuve : Voir l'annexe A.

Remarque 1. Quand les composantes de la variance sont inconnues, nous les remplaçons dans (6) par des estimateurs appropriés pour obtenir le MPLSB empirique, ou MPLSBE, dénoté par $\hat{y}_i^H$. Donc, nous obtenons le prédicteur modifié

$$\hat{y}_i^{\breve{\boldsymbol{a}}} = \hat{y}_i^H + \breve{a}_i\left(\sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \hat{y}_j^H\right). \qquad (20)$$

Remarque 2. Le critère (18) définit une « fonction de perte », où le choix des poids φ_i dépend du problème étudié. Par exemple, un statisticien peut décider d'attribuer des poids plus élevés aux domaines « plus importants » et des poids plus faibles aux domaines « moins importants ». Souvent, φ_i est une fonction des composantes de la variance. Dans certains cas, on peut choisir φ_i de sorte que les prédicteurs dérivés possèdent certaines propriétés souhaitables. Par exemple, $\varphi_i = [\widehat{\mathrm{Var}}(Y_i)]^{-1}$ donne les prédicteurs ITF, qui sont les prédicteurs MPLSB (au sens classique) dans le sous-espace qui est orthogonal à $(\omega_1, ..., \omega_n)'$ dans l'espace couvert par $\boldsymbol{Y}$.

Remarque 3. Si $\varphi_i = \omega_i[\mathrm{cov}(\tilde{y}_i^H, \tilde{y})]^{-1}$, où $\tilde{y} = \sum_{j=1}^n \omega_j \tilde{y}_j^H$, nous obtenons le prédicteur (13) dérivé par Pfeffermann et Barnard (1991). Quand $\varphi_i = [\widehat{\mathrm{Var}}(\hat{y}_j^H)]^{-1}$, nous obtenons celui utilisé par Battese, Harter et Fuller (1988).

3. Un autre moyen d'imposer la contrainte

Nous avons discuté d'une famille de prédicteurs pour lesquels le total des prédicteurs sur petits domaines est égal au total des estimations d'enquête directes. L'imposition de la contrainte (7) comporte implicitement la possibilité que le prédicteur sur petits domaines du total présente un biais parce que le modèle (3) est mal spécifié. Dans les applications pratiques, l'erreur de spécification du modèle est une préoccupation valide, car le mécanisme réel de production de $\boldsymbol{Y}$ est inconnu.

Une erreur de spécification courante survient quand les variables explicatives utilisées dans le modèle ne sont pas les mêmes que celles qui ont produit $\boldsymbol{Y}$. Donc, la direction du biais global pourrait ne pas être la même que celle du biais pour un petit domaine particulier. Le cas échéant, les prédicteurs de la forme (15) pourraient accroître le biais pour certains petits domaines comparativement au biais avant ajustement. Mantel et coll. (1993) ont conclu qu'en général, l'effet de l'étalonnage est une légère amélioration du biais global au prix d'une certaine détérioration des autres mesures d'évaluation.

Puisque le biais n'est pas nul s'il existe une corrélation non nulle entre ω_i et $(Y_i - \hat{y}_i)$, on peut le réduire en incluant ω_i dans le modèle. Autrement dit, pour un modèle donné, une approche consiste à utiliser le modèle augmenté

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{e}, \qquad (21)$$

où $\boldsymbol{X}_1 = (\boldsymbol{X}, \boldsymbol{\omega})$ et $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)'$ pour obtenir le MPLSB ou MPLSBE. Quand $\boldsymbol{\omega}$ est dans le modèle, l'ajustement nécessaire pour satisfaire la contrainte (7) sera souvent beaucoup plus faible que celui requis pour le modèle sans $\boldsymbol{\omega}$.

Grâce à l'approche du modèle augmenté, nous pouvons aller une étape plus loin et construire des prédicteurs qui satisfont la contrainte (7). En premier lieu, supposons que les variances σ_{ei}^2 sont connues. Notons que

$$\sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \hat{y}_j^H = \sum_{i=1}^n \omega_i(1-\gamma_i)(Y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) \qquad (22)$$

et $\omega_i(1-\gamma_i)\mathrm{Var}(Y_i) = \omega_i\sigma_{ei}^2$. En appliquant la théorie du modèle linéaire, nous pouvons montrer que le prédicteur construit avec le modèle augmenté

$$\boldsymbol{Y} = \boldsymbol{X}_2\boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{e}, \qquad (23)$$

où $\boldsymbol{X}_2 = (\boldsymbol{X}, \boldsymbol{\omega}_e)$ et $\boldsymbol{\omega}_e = (\omega_1\sigma_{e1}^2, ..., \omega_n\sigma_{en}^2)'$ possède la propriété d'autocalage quand nous utilisons l'estimateur par les moindres carrés généralisés (MCG) de $\boldsymbol{\beta}$. Notons que cette approche donne un prédicteur différent de celui de You-Rao (16).

Si les σ_{ei}^2 sont inconnues, nous remplaçons σ_{ei}^2 dans $\boldsymbol{\omega}_e$ par son estimateur $\hat{\sigma}_{ei}^2$. À condition que les $\hat{\sigma}_{ei}^2$ dans $\boldsymbol{\omega}_e$ soient les mêmes que les $\hat{\sigma}_{ei}^2$ utilisés pour construire γ_i, les prédicteurs ont la propriété d'autocalage. Si les σ_{ei}^2 sont de la forme $\sigma_e^2 f(\boldsymbol{u}_i)$, où σ_e^2 est inconnue, mais que $\boldsymbol{u}_i$ et $f(.)$

sont connues, nous pouvons construire les variables pour le modèle (23) en utilisant $\boldsymbol{\omega}_e = (\omega_1 f(\boldsymbol{u}_1), \ldots, \omega_n f(\boldsymbol{u}_n))'$ sans estimer σ_e^2. Par exemple, si $\sigma_{ei}^2 = m_i^{-1}\sigma_e^2$, le $\boldsymbol{\omega}_e$ pour le modèle (23) est $(m_i^{-1}\omega_1, \ldots, m_n^{-1}\omega_n)'$.

4. L'EQM des prédicteurs modifiés

Nous pouvons montrer que tout prédicteur de la forme (15) peut s'écrire

$$\tilde{\boldsymbol{y}}^a = \boldsymbol{Y} - \boldsymbol{C}_a^{-1}\boldsymbol{B}\boldsymbol{C}_a(\boldsymbol{I}_n - \boldsymbol{\Gamma})(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \qquad (24)$$

en posant que $\boldsymbol{C}_a = \boldsymbol{A}_a\boldsymbol{T}$, où

$$\boldsymbol{A}_a = \begin{pmatrix} 1 & \boldsymbol{0}'_{n-1} \\ -\boldsymbol{a}_{n-1} & \boldsymbol{I}_{n-1} \end{pmatrix},$$

$$\boldsymbol{a}_{n-1} = (a_2, \ldots, a_n)'.$$

Par conséquent, nous pouvons utiliser l'estimateur de la variance de $\tilde{\boldsymbol{y}}^a$ défini en (14) et proposé dans Isaki et coll. (2000) pour estimer l'EQM de tout prédicteur de la forme (15). Souvent, l'EQM d'un prédicteur ajusté est proche de celle du prédicteur avant l'ajustement.

Comme le modèle augmenté (23) possède la propriété d'autocalage, nous pouvons employer, pour estimer l'EQM, la formule utilisée pour les prédicteurs MPLSBE habituels.

5. Étude par simulation

5.1 Conditions de simulation

Afin d'étudier les propriétés empiriques des prédicteurs sur petits domaines décrits aux sections 2 et 3, nous utilisons des données conçues pour simuler une grande enquête nationale dans laquelle on souhaite produire des estimations par État. Le tableau 1 contient les populations approximatives des 50 États des États-Unis en l'an 2000. Les tailles d'échantillon m_i données dans le tableau sont approximativement proportionnelles aux racines carrées des populations des États.

Nous avons produit en tout 10 000 échantillons pour l'étude par simulation. Chacun était composé d'observations produites à l'aide du modèle

$$Y_{ij} = \boldsymbol{x}_i'\boldsymbol{\beta} + b_i + \varepsilon_{ij}, \qquad (25)$$

où $\boldsymbol{x}_i' = (1, z_i)$, $\boldsymbol{\beta}' = (6{,}0,\ 3{,}0)$, $z_i = \mathrm{Pop}_i^{0,2} - \overline{\mathrm{Pop}^{0,2}}$, Pop_i est la population de l'État i en millions, $\overline{\mathrm{Pop}^{0,2}}$ est la moyenne de $\mathrm{Pop}_i^{0,2}$, $b_i \sim NI(0, 1)$, et $\varepsilon_{ij} \sim NI(0, 16)$. Les b_i et les e_{ij} sont indépendants. Pour les observations au niveau de l'État, le modèle devient

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + b_i + e_i, \qquad (26)$$

où $Y_i = m_i^{-1}\sum_{j=1}^{m_i} Y_{ij}$ et $e_i = m_i^{-1}\sum_{j=1}^{m_i}\varepsilon_{ij}$. Avec les tailles d'échantillon données au tableau 1, le γ_i défini en (6) est 0,784 pour l'État le plus grand (Californie) et 0,304 pour l'État le plus petit (Wyoming).

Tableau 1 Populations et tailles d'échantillon pour la simulation

État	Population (en milliers)	Taille de l'échantillon (m_i)	État	Population (en milliers)	Taille de l'échantillon (m_i)
1	33 640	58	26	4 000	20
2	21 160	46	27	3 610	19
3	19 360	44	28	3 240	18
4	16 000	40	29	3 240	18
5	12 250	35	30	2 890	17
6	12 250	35	31	2 890	17
7	11 560	34	32	2 560	16
8	10 240	32	33	2 560	16
9	8 410	29	34	2 250	15
10	8 410	29	35	1 960	14
11	7 840	28	36	1 690	13
12	7 290	27	37	1 690	13
13	6 250	25	38	1 690	13
14	6 250	25	39	1 210	11
15	5 760	24	40	1 210	11
16	5 760	24	41	1 210	11
17	5 760	24	42	1 210	11
18	5 290	23	43	1 000	10
19	5 290	23	44	810	9
20	5 290	23	45	810	9
21	4 840	22	46	810	9
22	4 410	21	47	640	8
23	4 410	21	48	640	8
24	4 410	21	49	640	8
25	4 000	20	50	490	7

Pour étudier les propriétés des cinq prédicteurs, c'est-à-dire MPLSBE, Pfeffermann-Bernard (PB), Isaki-Tsay-Fuller (ITF), You-Rao (YR) et le modèle augmenté (23) (AUG2), nous nous sommes servis de deux modèles d'estimation. Le premier est un modèle spécifié incorrectement avec $\boldsymbol{x}_i' = 1$ dans la notation de (26). Nous l'appelons modèle (A). De même, nous dénommons modèle (B) le modèle générateur de données (26) avec $\boldsymbol{x}_i' = (1, z_i)$.

Suivant la méthode décrite dans Wang et Fuller (2003), l'estimateur de σ_b^2 est

$$\hat{\sigma}_b^2 = \max\{0{,}5[\hat{V}(\tilde{\sigma}_b^2)]^{0,5},\ \tilde{\sigma}_b^2\}, \qquad (27)$$

où

$$\tilde{\sigma}_b^2 = \sum_{i=1}^{50} c_i\left[\frac{50}{50-k}(Y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\mathrm{MCO}})^2 - \hat{\sigma}_{ei}^2\right], \qquad (28)$$

$$\hat{V}(\tilde{\sigma}_b^2) = \sum_{i=1}^{50} c_i^2\left\{\left[\frac{50}{50-k}(Y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\mathrm{MCO}})^2 - \hat{\sigma}_{ei}^2\right] - \tilde{\sigma}_b^2\right\}^2, \qquad (29)$$

$c_i = (\sum_{i=1}^{50} m_i^{0,5})^{-1} m_i^{0,5}$, $\hat{\boldsymbol{\beta}}_{\mathrm{MCO}}$ est l'estimateur par les moindres carrés ordinaires du coefficient de régression de

Y_i sur $\boldsymbol{x}_i$, k est la dimension du vecteur $\boldsymbol{x}_i$, $\hat{\sigma}_{ei}^2 = m_i^{-1} s_i^2$ et $s_i^2 = (m_i - 1)^{-1} \sum_{j=1}^{m_i} (Y_{ij} - Y_i)^2$ est la variance d'échantillon pour le domaine i.

Le prédicteur MPLSBE est

$$\hat{y}_i = \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_{\mathrm{MCG}} + \hat{\gamma}_i (Y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_{\mathrm{MCG}}), \qquad (30)$$

où

$$\hat{\gamma}_i = (\hat{\sigma}_b^2 + \hat{\sigma}_{ei}^2)^{-1} \hat{\sigma}_b^2, \qquad (31)$$

et

$$\hat{\boldsymbol{\beta}}_{\mathrm{MCG}} = \left[\sum_{i=1}^{n} (\sigma_b^2 + \hat{\sigma}_{ei}^2)^{-1} \boldsymbol{x}_i \boldsymbol{x}_i' \right]^{-1} \left[\sum_{i=1}^{n} (\hat{\sigma}_b^2 + \hat{\sigma}_{ei}^2)^{-1} \boldsymbol{x}_i Y_i \right] \qquad (32)$$

est l'estimateur par les moindres carrés généralisés de $\boldsymbol{\beta}$. La contrainte considérée est $\sum_{i=1}^{50} \omega_i \hat{y}_i^{\boldsymbol{a}} = \sum_{i=1}^{50} \omega_i Y_i$, où

$$\omega_i = \left(\sum_{i=1}^{50} \mathrm{Pop}_i \right)^{-1} \mathrm{Pop}_i. \qquad (33)$$

Avec le prédicteur MPLSBE, nous dérivons les prédicteurs PB et ITF en utilisant (13) et (14). Nous dérivons le prédicteur YR en utilisant (16) avec $\hat{\gamma}_i$ défini en (31) et $\hat{\boldsymbol{\beta}}_{\mathrm{YR}}$ défini en (17). Le prédicteur AUG2 est de la forme (30) avec $\boldsymbol{x}_i' = (1, \omega_i \hat{\sigma}_{ei}^2)$ sous le modèle augmenté (A) et $\boldsymbol{x}_i' = (1, z_i, \omega_i \hat{\sigma}_{ei}^2)$ sous le modèle augmenté (B).

Pour chacun des dix prédicteurs, nous calculons le critère

$$Q(\hat{\boldsymbol{y}}) = 0{,}02 \sum_{i=1}^{50} \varphi_i (\hat{y}_i - y_i)^2 \qquad (34)$$

où $\varphi_i = (\gamma_i m_i^{-1} \sigma_e^2)^{-1}$. Notons que φ_i^{-1} est la variance du prédicteur de y_i construite à l'aide des paramètres connus.

5.2 Résultats de la simulation

L'estimateur de σ_b^2 construit sous le modèle (B) a une moyenne Monte Carlo de 1,001 et un écart-type de 0,386. Quand nous utilisons le modèle (A) pour l'estimation, la moyenne Monte Carlo de l'estimateur de σ_b^2 est 1,720 et l'écart-type est de 0,521. Le carré de la moyenne de $3z_i$ est 0,636. Par conséquent, la méthode d'estimation avec l'utilisation du modèle (A) intègre une part importante de l'effet fixe de domaine du modèle (B) dans l'effet aléatoire. Quand $\hat{\sigma}_b^2$ est plus grand, $\hat{\gamma}_i$ est plus grand et une plus forte proportion de Y_i est utilisée pour construire les prédicteurs, ce qui annule en partie l'effet de l'erreur de spécification du modèle. Sous le modèle augmenté (A), l'estimateur de σ_b^2 est égal à 1,197, c'est-à-dire qu'une part nettement plus faible de l'effet fixe de domaine du modèle (B) est intégrée dans l'effet aléatoire.

Le tableau 2 contient certaines statistiques sommaires pour les prédicteurs fondés sur 10 000 échantillons simulés. Le biais empirique dans $\sum \omega_i (Y_i - \hat{y}_i)$ est nul pour tous les prédicteurs quand le modèle (B) ou sa version augmentée est utilisé, parce que le modèle de prédiction concorde avec le modèle de génération des données. L'écart-type simulé de la différence $\sum \omega_i (Y_i - \hat{y}_i)$ est égal à 0,022 pour le prédicteur MPLSBE habituel. L'écart-type de $\sum \omega_i (Y_i - \hat{y}_i)$ est nul pour les quatre autres prédicteurs, parce que ceux-ci satisfont la contrainte $\sum_{i=1}^{50} \omega_i \hat{y}_i = \sum_{i=1}^{50} \omega_i Y_i$.

Tableau 2
Propriétés Monte Carlo des prédicteurs pour petits domaines (moyenne de 10 000 échantillons générés par le modèle B)

	Quantité	MPLSBE	PB	ITF	YR	AUG2
		Prédicteur construit sous le modèle (A)				
$\sum \omega_i (Y_i - \hat{y}_i)$	Moyenne	-0,100	0,000	0,000	0,000	0,000
	(É.-T.)	(0,027)	(0,000)	(0,000)	(0,000)	(0,000)
$Q(\hat{\boldsymbol{y}})$	Moyenne	1,438	1,446	1,419	1,558	1,298
		Prédicteur construit sous le modèle (B)				
$\sum \omega_i (Y_i - \hat{y}_i)$	Moyenne	-0,000	0,000	0,000	0,000	0,000
	(É.-T.)	(0,022)	(0,000)	(0,000)	(0,000)	(0,000)
$Q(\hat{\boldsymbol{y}})$	Moyenne	1,203	1,202	1,202	1,208	1,219

La prédiction fondée sur le modèle (A) ou sa version augmentée est biaisée, parce que le modèle de génération des données (B) contient une fonction de la taille de population. La moyenne simulée de la différence pondérée $\sum \omega_i (Y_i - \hat{y}_{iA})$ est -0,100, où $\hat{y}_{iA}$ est le prédicteur MPLSBE. La valeur de la statistique t pour le biais pondéré est -3,70. La variance simulée de la moyenne pondérée de la prédiction est égale à

$$V \left\{ \sum_{i=1}^{50} \omega_i \hat{y}_{iA} \right\} = 0{,}060.$$

L'erreur quadratique moyenne estimée de la prédiction sur le modèle (A) de $\sum \omega_i y_i$ pour les données générées par le modèle (B) est

$$\mathrm{EQM} \left\{ \sum_{i=1}^{50} \omega_i (\hat{y}_{iA} - y_i) \right\} = V \left\{ \sum_{i=1}^{50} \omega_i \hat{y}_{iA} \right\} + \mathrm{Biais}^2$$
$$= 0{,}060 + (-0{,}100)^2 = 0{,}070. \qquad (35)$$

La variance de $\sum \omega_i Y_i$ est

$$V \left\{ \sum_{i=1}^{50} \omega_i Y_i \right\} = \sum_{i=1}^{50} \omega_i^2 (\sigma_b^2 + m_i^{-1} \sigma_e^2) = 0{,}0622. \qquad (36)$$

Donc, l'utilisation de $\sum \omega_i Y_i$ comme estimateur de $\sum \omega_i y_i$ est environ 12,5 % plus efficace que celle du prédicteur $\sum \omega_i \hat{y}_{iA}$ fondé sur le modèle (A). Étant donné le calage, les quatre prédicteurs (PB, ITF, YR et AUG2) de $\sum \omega_i y_i$ ont la même EQM que la moyenne estimée directement $\sum \omega_i Y_i$. Le carré du biais représenterait une proportion beaucoup plus importante de l'erreur quadratique moyenne si les petits domaines étaient plus nombreux.

La valeur du critère Q pour MPLSBE est 1,20 sous le modèle (B). Donc, l'estimation des paramètres accroît la variance moyenne des prédicteurs d'environ 20 % comparativement à l'utilisation des paramètres connus. Si les prédictions sont faites en utilisant des valeurs connues de σ_{ei}^2, la valeur du critère Q pour MPLSBE est 1,06. Par conséquent, l'estimation de $\hat{\sigma}_{ei}^2$ est celle qui contribue le plus à l'accroissement de la variabilité. Comme le biais est nul quand le modèle (B) est utilisé pour l'estimation, les ajustements que produisent les prédicteurs contraints sont faibles, si bien que les prédicteurs ajustés donnent des valeurs de critères fort semblables à celles des prédicteurs non ajustés. Les prédicteurs YR ont des valeurs de critère un peu plus grandes que les prédicteurs PB et ITF correspondants, parce que le prédicteur YR utilise un estimateur inefficace de $\boldsymbol{\beta}$. Les prédicteurs reposant sur le modèle augmenté ont une valeur de critère Q légèrement plus grande, parce que le modèle contient une variable redondante. L'accroissement de moins de 2 % de Q est de l'ordre de n^{-1}, c'est-à-dire la perte attendue en raison de l'ajout d'un paramètre inutile dans une prédiction par les moindres carrés.

La valeur du critère Q pour MPLSBE sous le modèle (A) est de 1,438 comparativement à 1,203 pour MPLSBE sous le modèle (B). Il s'agit de la pénalité due à l'erreur de spécification du modèle. Parmi les méthodes d'ajustement fondées sur le modèle (A), la méthode ITF est celle dont la valeur de Q la plus faible. Les différences entre les prédicteurs PB, ITF et MPLSBE sont faibles. La valeur de Q est environ 8 % plus grande pour la méthode You-Rao que pour MPLSBE. Le prédicteur fondé sur le modèle augmenté a une valeur de Q environ 11 % plus petite que le prédicteur MPLSBE. Dans un certain sens, le modèle augmenté est spécifié moins incorrectement.

Non seulement l'approche d'augmentation donne lieu au calage des prédicteurs sur petits domaines, mais elle réduit aussi le biais au niveau du domaine quand le modèle est spécifié incorrectement. Les tableaux 3 et 4 donnent les propriétés Monte Carlo des prédicteurs pour certains domaines. Dans les tableaux, les biais estimés sont normalisés par $(\gamma_i m_i^{-1} \sigma_e^2)^{0,5}$, la racine carrée de l'EQM du prédicteur MPLSB avec paramètres connus, et les EQM estimées sont normalisées par $\gamma_i m_i^{-1} \sigma_e^2$. Quand le modèle correct (B) est utilisé, le biais Monte Carlo au niveau du domaine approche de zéro. Voir le tableau 3. En outre, l'écart est faible entre l'EQM calculée pour les diverses méthodes, le modèle augmenté produisant une EQM un peu plus grande (moins de 2 %). De nouveau, le tableau 3 montre que le processus de calage a peu d'effet sur l'EQM des prédicteurs au niveau du domaine comparativement au prédicteur MPLSBE.

Tableau 3
Propriétés Monte Carlo des prédicteurs individuels de domaine en utilisant des versions du modèle (B) (10 000 échantillons générés par le modèle B)

État	Quantité	EMPLSB	PB	ITF	YR	AUG2
1	Biais	0,011	0,012	0,012	0,013	0,011
	EQM	1,100	1,101	1,105	1,104	1,120
2	Biais	0,000	0,001	0,001	0,002	0,001
	EQM	1,072	1,072	1,072	1,076	1,092
14	Biais	-0,001	-0,001	-0,001	-0,001	-0,001
	EQM	1,058	1,058	1,058	1,058	1,074
26	Biais	0,015	0,015	0,015	0,015	0,018
	EQM	1,078	1,077	1,078	1,079	1,092
38	Biais	-0,005	-0,005	-0,005	-0,006	-0,003
	EQM	1,123	1,122	1,122	1,132	1,135
50	Biais	0,012	0,012	0,012	0,009	0,014
	EQM	1,222	1,222	1,222	1,247	1,246

Si nous utilisons un modèle spécifié incorrectement, tel que le modèle (A), le biais de la somme des prédicteurs MPLSBE en tant qu'estimateur du total est négatif dans notre exemple, parce que les États pour lesquels le biais est négatif ont une pondération ω_i élevée. Voir le tableau 4. Des méthodes d'ajustement, comme PB ou ITF, répartissent le biais entre tous les petits domaines. Donc, l'ajustement réduit le biais négatif des prédicteurs de domaine ayant un grand biais négatif et augmente le biais positif des prédicteurs ayant un grand biais positif. Par conséquent, l'EQM est plus petite pour les grands états et elle est un peu plus grande pour les petits états. Le prédicteur YR produit un biais plus important que le prédicteur ITF. Par ailleurs, les prédicteurs construits avec $\boldsymbol{x}_i' = (1, \omega_i \hat{\sigma}_{ei}^2)$, c'est-à-dire le modèle augmenté (A), sont nettement supérieurs à ceux construits sous le modèle (A). Le biais au niveau des domaines, que ceux-ci soient grands ou petits, est réduit.

Tableau 4
Propriétés Monte Carlo des prédicteurs individuels de domaine en utilisant des versions du modèle (A) (10 000 échantillons générés par le modèle B).

État	Quantité	MPLSBE	PB	ITF	YR	AUG2
1	Biais	-0,597	-0,225	-0,052	-0,473	-0,030
	EQM	1,471	1,165	1,130	1,331	1,139
2	Biais	-0,496	-0,227	-0,170	-0,358	-0,070
	EQM	1,330	1,134	1,115	1,207	1,124
14	Biais	-0,121	0,004	-0,031	0,055	-0,025
	EQM	1,100	1,085	1,086	1,089	1,105
26	Biais	0,057	0,157	0,115	0,249	0,053
	EQM	1,126	1,148	1,136	1,188	1,132
38	Biais	0,380	0,453	0,406	0,601	0,202
	EQM	1,340	1,405	1,361	1,571	1,233
50	Biais	0,922	0,980	0,931	1,178	0,537
	EQM	2,196	2,316	2,215	2,767	1,577

6. Conclusion

Dans le présent article, nous passons en revue plusieurs prédicteurs calés. Nous jetons un nouveau regard sur la contrainte d'étalonnage (7). Nous considérons l'imposition de la contrainte comme un problème d'ajustement et nous présentons un critère qui unifie la dérivation des prédicteurs calés. L'approche du critère pour résoudre le problème permet de considérer d'autres prédicteurs.

L'imposition de la contrainte comprend implicitement la possibilité que le modèle pour petits domaines soit spécifié incorrectement et que les prédicteurs présentent un biais. Quand le modèle est mal spécifié, l'ajustement par calage corrige le biais global, mais non le biais au niveau du petit domaine. L'approche du modèle augmenté aboutit à un prédicteur autocalé et réduit le biais au niveau du petit domaine. En outre, l'estimation de la variance des prédicteurs calés extérieurement est relativement complexe, tandis que l'estimation de la variance des prédicteurs autocalés est simple. Brièvement, si l'on est préoccupé par le biais, l'utilisation du modèle augmenté autocalé est préférable au calage externe.

7. Annexe

Preuve du théorème 1 : Soit $\tilde{y}_i^H$ un prédicteur MPLSB de y_i et soit $\hat{y}_i$ un prédicteur linéaire sans biais de y_i. Selon les résultats standard pour le MPLSB (voir, par exemple, Robinson (1991) et Harville (1976)), nous avons

$$\text{cov}(\tilde{y}_i^H - y_i,\ \hat{y}_j - \tilde{y}_j^H) = 0, \quad \text{si} \quad i \neq j, \qquad \text{(A.1)}$$

pour $i = 1, ..., n$ et $j = 1, ..., n$. Soit $R(\hat{\boldsymbol{y}}^{\boldsymbol{a}})$ la série de prédicteurs linéaires sans biais qui satisfont (7). Pour tout $\hat{\boldsymbol{y}} \in R(\hat{\boldsymbol{y}}^{\boldsymbol{a}})$, selon (A.1), nous avons

$$E\{(\hat{y}_i - y_i)^2\} = E\{(\tilde{y}_i^H - y_i)^2\} + E\{(\hat{y}_i - \tilde{y}_i^H)^2\}. \qquad \text{(A.2)}$$

Par conséquent,

$$Q(\hat{\boldsymbol{y}}) = \sum_{i=1}^n \varphi_i E\{(\hat{y}_i - y_i)^2\}$$

$$= \sum_{i=1}^n \varphi_i E\{(\tilde{y}_i^H - y_i)^2\} + \sum_{i=1}^n \varphi_i E\{(\hat{y}_i - \tilde{y}_i^H)^2\}. \qquad \text{(A.3)}$$

Puisque $\hat{\boldsymbol{y}}$ satisfait (7), nous avons $\sum_{i=1}^n \omega_i \hat{y}_i = \sum_{i=1}^n \omega_i Y_i$ et, pour le $\hat{y}_i^{\bar{a}}$ défini en (19),

$$\hat{y}_i^{\bar{a}} = \tilde{y}_i^H + \breve{a}_i \left[\sum_{j=1}^n \omega_j (Y_j - \tilde{y}_j^H) \right]$$

$$= \tilde{y}_i^H + \breve{a}_i \left[\sum_{j=1}^n \omega_j (\hat{y}_j - \tilde{y}_j^H) \right].$$

Selon (A.1),

$$Q(\hat{y}_i^{\bar{a}}) = \sum_{i=1}^n \varphi_i E(\hat{y}_i^{\bar{a}} - y_i)^2$$

$$= \sum_{i=1}^n \varphi_i E\left\{ \left[(\tilde{y}_i^H - y_i) + \breve{a}_i \left(\sum_{j=1}^n \omega_j \hat{y}_j - \sum_{j=1}^n \omega_j \tilde{y}_j^H \right) \right]^2 \right\}$$

$$= \sum_{i=1}^n \varphi_i E\{(\tilde{y}_i^H - y_i)^2\} + E\left\{ \left[\sum_{j=1}^n \omega_j (Y_j - \tilde{y}_j^H) \right]^2 \right\} \sum_{i=1}^n \varphi_i \breve{a}_i^2. \qquad \text{(A.4)}$$

Puisque $\breve{a}_i = (\sum_{i=1}^n \varphi_i^{-1} \omega_i^2)^{-1} \varphi_i^{-1} \omega_i$ dans (A.4), nous avons

$$Q(\hat{y}_i^{\bar{a}}) = \sum_{i=1}^n \varphi_i E\{(\tilde{y}_i^H - y_i)^2\} + E\left\{ \left[\sum_{j=1}^n \omega_j (\hat{y}_j - \tilde{y}_j^H) \right]^2 \right\} \left(\sum_{i=1}^n \varphi_i^{-1} \omega_i^2 \right)^{-1}. \qquad \text{(A.5)}$$

Notons que

$$E\left\{ \left[\sum_{j=1}^n \omega_j (\hat{y}_j - \tilde{y}_j^H) \right]^2 \right\} \leq \sum_{j=1}^n \sum_{k=1}^n \omega_j \omega_k g_j g_k$$

$$= \left(\sum_{i=1}^n \omega_i g_i \right)^2, \qquad \text{(A.6)}$$

où $g_j = \{E[(\hat{y}_j - \tilde{y}_j^H)^2]\}^{0.5}$. En vertu de l'inégalité de Cauchy,

$$\left(\sum_{j=1}^n \omega_j g_j \right)^2 \leq \left(\sum_{i=1}^n \varphi_i^{-1} \omega_i^2 \right) \left(\sum_{i=1}^n \varphi_i g_i^2 \right)$$

$$= \left(\sum_{i=1}^n \varphi_i^{-1} \omega_i^2 \right) \left[\sum_{i=1}^n \varphi_i E\{(\hat{y}_j - \tilde{y}_j^H)^2\} \right]. \qquad \text{(A.7)}$$

En combinant (A.3), (A.5), (A.6) et (A.7), nous avons $Q(\hat{y}_i^{\bar{a}}) \leq Q(\hat{\boldsymbol{y}})$.

Pour démontrer le caractère unique de $\hat{y}_i^{\bar{a}}$, nous devons vérifier quand les inégalités (A.6) et (A.7) deviennent des égalités. L'inégalité (A.6) devient une égalité si, et uniquement si,

$$\hat{y}_j - \tilde{y}_j^H = c_j^0 + c_j^1 (\hat{y}_1 - \tilde{y}_1^H) \qquad \text{(A.8)}$$

pour certaines constantes c_j^0 et c_j^1, $j = 2, ..., n$. L'inégalité (A.7) devient une égalité si, et uniquement si,

$$\sqrt{\varphi_i^{-1} \omega_i^2} \sqrt{v_j g_j^2} - \sqrt{v_j^{-1} \omega_j^2} \sqrt{\varphi_i g_i^2} = 0, \qquad \text{(A.9)}$$

ou, de manière équivalente,

$$\varphi_i^2 \omega_i^{-2} E\{(\hat{y}_i - \tilde{y}_i^H)^2\} = v_j^2 \omega_j^{-2} E\{(\hat{y}_j - \tilde{y}_j^H)^2\}. \qquad \text{(A.10)}$$

En outre,

$$\sum_{i=1}^{n} \omega_i \hat{y}_i = \sum_{i=1}^{n} \omega_i Y_i. \qquad \text{(A.11)}$$

En combinant (A.8), (A.10) et (A.11), nous obtenons que l'égalité tient si, et uniquement si, $\hat{y}_j = \hat{y}_i^{\bar{a}}$. Donc, nous avons montré que $\hat{y}_i^{\bar{a}}$ est le seul prédicteur linéaire sans biais qui satisfait (7) et minimise le critère (18).

Remerciements

Cette étude a été financée en partie par l'entente de coopération 68-3A75-14 conclue entre le Natural Resources Conservation Service de l'USDA et la Iowa State University.

Bibliographie

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of American Statistical Association*, 28-36.

Harville, D.A. (1976). Extension of the Gauss-Markov Theorem to include the estimation of random effects. *Annals of Statistics*, 384-395.

Henderson, C.R. (1963). Selection index and expected genetic advance. Dans *Statistical Genetics and Plant Breeding*, (Éds., W.D. Hanson et H.F. Robinson), National Academy of Sciences–National Research Council, Washington, Publication 982, 141-163.

Isaki, C.T., Tsay, J.H. et Fuller, W.A. (2000). Estimation des facteurs de correction au recensement. *Techniques d'enquête*, 31-42.

Mantel, H.J., Singh, A.C. et Barcau, M. (1993). Benchmarking of small area estimators. Dans *Proceedings of International Conference on Establishment Surveys*, American Statistical Association, Washington, DC, 920-925.

Pfeffermann, D., et Barnard, C.H. (1991). Some new estimators for small-area means with application to the assessment of Farmland values. *Journal of Business & Economic Statistics*, 73-84.

Rao, J.N.K. (2003). *Small area estimation*. New York : John Wiley & Sons, Inc.

Robinson, G.K. (1991). That MPLSB is a good thing: the estimation of random effects. *Statistical Science*, 15-51.

Wang, J., et Fuller, W.A. (2003). The mean square error of small area estimators constructed with estimated area variances. *Journal of American Statistical Association*, 716-723.

You, Y., et Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 431-439.

You, Y., et Rao, J.N.K. (2003). Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference*, 197-208.

Techniques d'enquête, juin 2008
Vol. 34, N° 1, pp. 41-54
Statistique Canada, N° 12-001-X au catalogue

Une répartition bayésienne des électeurs indécis

Balgobin Nandram et Jai Won Choi [1]

Résumé

Aux États-Unis, les données provenant des sondages électoraux sont habituellement présentées dans des tableaux de contingence à double entrée et de nombreux sondages sont réalisés avant qu'ait lieu l'élection réelle en novembre. Par exemple, pour l'élection du gouverneur de l'État de l'Ohio en 1998, trois sondages (Buckeye State Poll) ont eu lieu, un en janvier, un en avril et un en octobre; la première catégorie des tableaux représente les candidats (par exemple Fisher, Taft et autre) et la deuxième représente l'intention courante de vote (votera vraisemblablement ou ne votera vraisemblablement pas pour le gouverneur de l'Ohio). Le nombre d'électeurs indécis est important dans l'une ou dans les deux catégories pour les trois sondages et nous utilisons une méthode bayésienne pour les répartir entre les trois candidats. Nous pouvons ainsi modéliser divers scénarios de données manquantes sous les hypothèses d'ignorabilité et de non-ignorabilité, et nous utilisons un modèle Dirichlet-Multinomial pour estimer les probabilités de cellule qui nous aideront à prédire le gagnant. Nous proposons un modèle de non-réponse non ignorable variable en fonction du temps pour les trois tableaux. Ici, un modèle de non-réponse non ignorable est centré sur un modèle de non-réponse ignorable afin d'induire une certaine souplesse et une certaine incertitude au sujet de l'ignorabilité ou de la non-ignorabilité. Nous considérons également deux autres modèles concurrents, à savoir un modèle de non-réponse ignorable et un modèle de non-réponse non ignorable. Ces deux derniers modèles reposent sur l'hypothèse d'un processus stochastique commun pour obtenir un renforcement par emprunt de données au cours du temps. Nous utilisons des méthodes de Monte Carlo par chaîne de Markov pour ajuster les modèles. Nous construisons aussi un paramètre qui peut éventuellement être utilisé pour prédire le gagnant parmi les candidats à l'élection de novembre.

Mots clés : Monte Carlo par chaîne de Markov; échantillonneur de Metropolis; modèle Dirichlet-Multinomial; modèle variable en fonction du temps; tableau de contingence à double entrée.

1. Introduction

Il est courant, en pratique, d'utiliser des tableaux de contingence à double entrée pour présenter les données de sondage. Notre application a pour but de prédire le gagnant d'une élection en utilisant des tableaux construits en se basant sur une courte série de sondages électoraux réalisés avant l'élection réelle. Dans le cas de nombreux sondages, des données manquent, si bien que la classification des individus échantillonnés n'est que partielle. Little et Rubin (2002, section 1.3) définissent les trois mécanismes de création de données manquantes [manquent complètement au hasard (MCAR pour « missing completely at random »), manquent au hasard (MAR pour « missing at random »), ne manquent pas au hasard (MNAR pour « missing not at random »)]; des modèles à données manquantes ignorables sont utilisés pour analyser les données issues des mécanismes MAR et MCAR, et des modèles à données manquantes non ignorables sont utilisés pour les données provenant du mécanisme MNAR. Donc, pour le tableau à double entrée, il existe à la fois une non-réponse partielle (l'une des deux catégories manque) et une non-réponse totale (les deux catégories manquent). Comme le mécanisme de création des données manquantes pourrait ne pas être connu, il serait peut-être préférable d'utiliser un modèle dans lequel est incluse une différence entre les données observées et les données manquantes (c'est-à-dire des données manquantes non ignorables). Pour un tableau de contingence général $r \times c$, nous examinons la question de l'estimation des probabilités de cellule du tableau à double entrée. Elle est importante, car, si le nombre d'électeurs indécis est considérable, la prédiction du résultat d'une élection fondée uniquement sur les données partiellement observées pourrait être incorrecte.

Comme dans Nandram, Cox et Choi (2005), nous avons essentiellement quatre tableaux à double entrée, à savoir le tableau contenant les cas complets et trois tableaux supplémentaires. Le premier de ceux-ci ne contient que les classifications par ligne (non-réponse partielle), le deuxième ne contient que les classifications par colonne (non-réponse partielle) et le troisième ne contient aucune classification (non-réponse totale). Nous avons étendu les modèles de non-réponse ignorable et non ignorable établis pour les tableaux de contingence à double entrée par Nandram et coll. (2005) afin d'inclure une troisième catégorie (c'est-à-dire le temps dans une brève série de sondages électoraux). Nous avons étendu ces modèles encore davantage afin d'intégrer une structure de non-réponse non ignorable variable en fonction du temps, qui peut fournir une prédiction plus efficace. Une méthode bayésienne nous

1. Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester MA 01609. Courriel : balnan@wpi.edu.; Jai Won Choi, Department of Biostatistics, Medical College of Georgia, 1420 15th Street, Augusta, GA 30912. Courriel : jchoi@MCG.edu.

permet de modéliser divers scénarios de données manquantes sous les hypothèses d'ignorabilité et de non-ingorabilité, et nous obtenons un modèle de non-réponse non ignorable variable en fonction du temps.

Notre application, qui est celle de l'élection du gouverneur de l'Ohio, pose plusieurs problèmes connexes. Nous classons les personnes échantillonnées en fonction de deux types d'attributs et nous analysons les cellules des tableaux de contingence ainsi obtenues. Cependant, nous ne disposons que d'une classification partielle des individus, parce que certains sont classés en fonction d'un attribut au plus et d'autres ne sont pas classés du tout. Plus précisément, nous utilisons des données tabulaires provenant des sondages électoraux réalisés en Ohio pour étudier la relation entre une mesure de l'intention de voter (votera vraisemblablement ou ne votera vraisemblablement pas) et le candidat préféré (Fisher, Taft et autre) pour illustrer notre méthode. Il est intéressant de constater que l'intention de voter est reliée à la préférence concernant les candidats. En outre, il est souhaitable de faire un ajustement pour tenir compte des électeurs indécis, parce que leur proportion est habituellement élevée et qu'ils décident souvent du résultat final d'une élection.

Nous ne savons pas s'il convient d'utiliser un modèle de non-réponse ignorable ou un modèle de non-réponse non ignorable, mais il peut exister une incertitude quant à l'ignorabilité des électeurs indécis dans les sondages électoraux. Faisant référence au sondage électoral dans l'État de l'Ohio, Chen et Stasny (2003) ont déclaré que l'hypothèse de la non-ignorabilité de la non-réponse est peut-être raisonnable dans cette étude, parce que les gens pourraient hésiter à indiquer leur préférence si elle va à un candidat impopulaire ou si, au moment du sondage, elle n'est pas suffisamment ferme ou exacte en regard des normes de l'interview. Ils ont également fait remarquer que Chang et Krosnick (2001) utilisent des modèles de non-réponse ignorable pour leurs analyses, mais qu'ils laissent entendre que la non-réponse pourrait être reliée aux données inobservées proprement dites. Chen et Stasny (2003) ajustent trois modèles de non-réponse ignorable (A, B et C) et un modèle de non-réponse non ignorable (D). Nous comparons nos résultats aux leurs.

Nandram et Choi (2002a, b) utilisent un modèle d'extension pour étudier les données binaires avec non-réponse non ignorable. Le modèle d'extension, qui est un modèle de non-réponse non ignorable, dégénère en un modèle de non-réponse ignorable (dans l'esprit de Draper, 1995). Cette dégénérescence a lieu quand un paramètre du modèle de non-réponse non ignorable est fixé à une certaine valeur; une bonne description du concept de centrage est donnée dans Nandram et coll. (2005, section 1.2). Cette procédure étant difficile à exécuter telle qu'elle est décrite, nous utilisons une méthode de rechange qui, comme dans Nandram et coll. (2005), nous permet de donner une expression de l'incertitude au sujet de l'ignorabilité. Elle repose sur le concept du centrage d'un modèle de non-réponse non ignorable sur un modèle de non-réponse ignorable que nous avons utilisé dans plusieurs de nos articles pour exprimer l'incertitude au sujet de l'ignorabilité ou de la non-ignorabilité. Ici, pour la non-réponse non ignorable, nous essayons d'appliquer une méthode apparentée, mais les questions qui se posent pour un tableau de contingence à double entrée sont plus complexes, surtout quand une troisième catégorie (par exemple le temps) est incluse dans ces tableaux.

En suivant l'approche de Chen et Fienberg (1974), Chen et Stasny (2003) décrivent les deux questions dont nous discutons dans le présent article. Dans le cas des tableaux de contingence à double entrée, ils ne peuvent traiter que la non-réponse partielle; la non-réponse totale est exclue de leur analyse. Toutefois, ils supposent que les données manquent au hasard et montrent comment obtenir des estimateurs du maximum de vraisemblance sous leur modèle. Ils utilisent également un modèle de non-réponse non ignorable (D), qu'ils affirment être leur meilleur modèle. Little et Rubin (2002, chapitre 15) soulignent qu'un des problèmes que pose le modèle de non-réponse non ignorable est que les paramètres sont trop nombreux et que beaucoup d'entre-eux ne sont pas identifiés, de sorte qu'ils essayent d'apporter une correction en utilisant des modèles log-linéaires hiérarchiques. [Voir Nandram et coll. (2005) pour le cas où il y a trois tableaux supplémentaires.]

Notre méthode diffère de celle de Chen et Stasny (2003). La principale différence est que nous suivons une approche bayésienne, ce qui nous permet d'utiliser une méthode ne s'appuyant pas sur la théorie asymptotique, d'intégrer la non-réponse non ignorable dans la modélisation et d'obtenir un modèle de non-réponse non ignorable variable en fonction du temps en vue d'estimer les proportions d'électeurs revenant aux trois candidats. Pour essayer de prédire le gagnant de manière plus convaincante, nous avons également construit un nouveau paramètre relativement facile à analyser dans le contexte du paradigme bayésien. La méthode bayésienne permet de modéliser divers scénarios de donnée manquantes sous deux hypothèses distinctes (c'est-à-dire données manquantes ignorables et non ignorables). Notre idée consiste à partir d'un modèle de non-réponse ignorable, qui est alors étendu à un modèle de non-réponse non ignorable, ainsi qu'au modèle de non-réponse non ignorable variable en fonction du temps. Il convient de souligner que la non-réponse totale est également incluse dans notre modélisation, alors que les autres chercheurs la considèrent comme un problème distinct qu'ils corrigent par repondération (par exemple, voir

la discussion dans Kalton et Kasprzyk 1986). Pourtant, la non-réponse totale pourrait également être non ignorable et il serait alors nécessaire d'inclure simultanément les non-réponses partielle et totale.

Dans le présent article, notre principale contribution consiste à introduire une méthode bayésienne pour analyser des données provenant d'un tableau de contingence $r \times c$, en présence à la fois de non-réponse partielle et totale, et dans des conditions où le mécanisme de création des données manquantes peut être non ignorable avec une structure variable en fonction du temps. À la section 2, nous décrivons les donnés catégoriques sur l'intention de voter et le candidat préféré avec une structure variable en fonction du temps. À la section 3, nous exposons la méthode appliquée pour estimer les probabilités de cellule en intégrant les deux types de données manquantes, et nous montrons comment étendre un modèle de non-réponse ignorable à un modèle de non-réponse non ignorable et à un modèle de non-réponse variant en fonction du temps. Nous montrons aussi comment utiliser les méthodes de Monte Carlo par chaîne de Markov pour ajuster le modèle de non-réponse non ignorable. À la section 4, nous analysons les données des sondages électoraux réalisés en Ohio pour démontrer la souplesse de nos méthodes. Enfin, à la section 5, nous présentons nos conclusions.

2. Données des sondages électoraux de 1998 en Ohio

Le Center for Survey Research (CSR) de la Ohio State University a réalisé le Buckeye State Poll (BSP), ou sondage électoral dans l'État de l'Ohio, durant les élections de 1998 en vue de désigner le sénateur, le gouverneur, le procureur général, le secrétaire d'État, le trésorier et le maire de Columbus. Au cours de certains mois avant les élections, le CSR a réalisé des sondages préélectoraux dans le cadre du BPS et a inclus des questions supplémentaires pour recueillir des renseignements sur la probabilité de voter et sur le candidat préféré. Dans le BPS, les ménages sont échantillonnés par la méthode de composition aléatoire (CA) et un adulte par ménage est sélectionné pour l'interview par la méthode du dernier anniversaire (Lavrakas 1993).

Suit une brève description de la méthode CA. Les entreprises de sondage y recourent à grande échelle, avant tout en vue d'obtenir un échantillon représentatif de la population globale d'électeurs. L'échantillonnage par la méthode CA repose sur l'hypothèse qu'il n'est pas possible d'obtenir un échantillon représentatif en utilisant les numéros de téléphone répertoriés dans l'annuaire. Chaque numéro de téléphone contient dix chiffres, dont les trois premiers forment l'indicateur régional, les trois suivants forment le préfixe (aussi appelé indicatif de central), tandis que les quatre derniers (suffixe) identifient un abonné particulier ou un ménage (un ménage peut avoir plus d'un numéro de téléphone). Les codes régionaux sont de nature géographique et identifient habituellement les localités dans un État, et les indicatifs de central peuvent aussi avoir une orientation géographique. Il existe dix millions de numéros à composer, mais approximativement moins de 25 % de ceux-ci sont des numéros de téléphone réels. Par conséquent, du temps et de l'argent sont gaspillés en composant des numéros non attribués. Nous discutons de ce point plus en détail à la section 3.

Chen et Stasny (2003), ainsi que Chang et Krosnick (2001) ont analysé les données provenant de trois sondages prévisionnels préélectoraux du BPS. Des détails sur chacun de ces trois sondages figurent au tableau 1. Dans le cadre de ces sondages, on a déterminé trois fois le candidat préféré des répondants (en janvier, en avril et en octobre) pour la course au poste de gouverneur de l'Ohio de novembre 1998. En outre, on a posé deux questions aux répondants pour déterminer la probabilité qu'ils se rendent aux urnes le jour de l'élection. Chang et Krosnick (2001) ont également utilisé des variables filtres (telles que l'inscription sur la liste des électeurs, la probabilité autodéclarée de voter et le fait d'avoir voté lors de la dernière élection importante, *etc*.) pour déterminer quelles personnes étaient le plus susceptibles de voter. Donc, la prédiction est fondée uniquement sur les répondants susceptibles de voter. Les personnes inscrites sur les listes d'électeurs sont classées en trois catégories, à savoir votera probablement, ne votera probablement pas et indécis. Chang et Krosnick (2001) ont montré que la répartition déterministe des répondants indécis améliore la prévision des préférences des électeurs pour les candidats, comparativement à l'exclusion de tous les répondants indécis. Chen et Stasny (2003) ont utilisé des modèles probabilistes pour répartir les électeurs indécis et ont comparé leurs prévisions à celles de Chang et Krosnick (2001).

L'ensemble de données de Chen et Stasny (2003) diffère légèrement du nôtre, parce que nous utilisons les nombres d'indécis (non-réponse totale) sur les deux variables. Un électeur peut être indécis pour au moins une des deux variables catégoriques à chacun des trois sondages. Chen et Stasny (2003) n'étudient les données comprenant les indécis que pour exactement une variable, et non les deux. Dans le tableau 1, pour les électeurs indécis sur les deux variables, les chiffres pour les sondages de janvier, d'avril et d'octobre sont de 5, 3 et 4, respectivement; ils sont représentés en caractères gras. En fait, l'inclusion de ces dénombrements dans notre modèle est une extension des modèles de Chen et Stasny qui généralise considérablement notre méthode.

Nous décrivons brièvement le tableau de contingence 2×3 des données électorales en Ohio selon l'intention de voter (IV) et le candidat préféré (CAN). Ici, IV est une variable binaire comportant deux niveaux, à savoir votera vraisemblablement et ne votera vraisemblablement pas; CAN comporte trois niveaux, à savoir Fisher, Taft et autre. Les variables IV et CAN comprennent aussi des électeurs indécis. La majeure partie de ceux-ci proviennent des groupes d'électeurs qui « voteront vraisemblablement » et « ne voteront vraisemblablement pas », et les chiffres sont de 173, 142 et 138 pour janvier, avril et octobre, respectivement; les nombres d'électeurs indécis pour Fisher, Taft et autre sont beaucoup plus faibles.

Dans le sondage de janvier 1998, environ 73 % des électeurs sont complètement classés, 27 % n'ont pas décidé de leur candidat préféré, 1 % seulement ne savent pas s'ils voteront ou non et cinq personnes seulement sont entièrement non classées parmi les 648 participants. L'ensemble de données utilisé dans notre étude est présenté au tableau 1 sous forme de tableau de contingence 2×3 de l'intention de voter et du candidat préféré. Notre problème est de prédire le candidat gagnant en estimant la proportion de votes finaux pour chaque candidat.

Tableau 1
Classification des données du sondage de l'État d'Ohio d'octobre 1998 selon l'intention de voter et le candidat

	Candidat				
Intention	**Fisher**	**Taft**	**Autre**	**Indécis**	**Total**
a. Janvier 1998					
Votera vraisemblablement	127	183	8	109	427
Ne votera vraisemblablement pas	57	94	4	59	214
Indécis	0	2	0	**5**	7
Total	184	279	12	173	648
b. Avril 1998					
Votera vraisemblablement	114	135	1	61	311
Ne votera vraisemblablement pas	104	149	3	78	334
Indécis	2	6	0	**3**	11
Total	220	290	4	142	656
c. Octobre 1998					
Votera vraisemblablement	112	140	23	61	336
Ne votera vraisemblablement pas	96	108	21	73	298
Indécis	7	11	1	**4**	23
Total	215	259	45	138	657

NOTA : Ces données sont tirées de Chang et Krosnick (2001); Chen et Stasny (2003) ont utilisé un ensemble de données fort semblable; ils n'ont pas utilisé 5, 3 et 4, c'est-à-dire le nombre d'électeurs indécis dans les deux variables.

Les échantillons obtenus en janvier, en avril et en octobre sont indépendants. Il n'y a pas de suréchantillonnage pour une sous-population particulière ni de pondération de l'échantillon original. À l'instar de nombreux sondages téléphoniques, la base de sondage par composition aléatoire (CA) souffre souvent du problème de sous-dénombrement. Comme la couverture téléphonique n'est pas uniforme selon l'âge, la race, le sexe, le revenu et la géographie, il est nécessaire de poststratifier l'échantillon original afin de réduire le biais de couverture par pondération correcte des données originales.

Nous exécutons un test préliminaire de l'hétérogénéité des proportions de cellule entre les trois sondages. En supposant que les données manquent au hasard, nous ajoutons les électeurs indécis. Nous supposons que, pour chaque ligne (colonne), les électeurs indécis sont ajoutés proportionnellement aux effectifs de cellule. Soit n_{tjk} les effectifs corrigés de cellule avec $n_t = \sum_{j=1}^{r}\sum_{k=1}^{c} n_{tjk}$, et soit p_{tjk} les proportions de cellule. Pour un modèle à proportions hétérogènes, nous supposons que

$$\boldsymbol{n}_t \,|\, \boldsymbol{p}_t \overset{\text{ind}}{\sim} \text{Multinomiale}(n_t, \boldsymbol{p}_t) \quad \text{et} \quad \boldsymbol{p}_t \overset{\text{iid}}{\sim} \text{Dirichlet}(\mathbf{1}), t = 1, \ldots, T,$$

où $\mathbf{1}$ est un vecteur rc de valeurs unitaires.

Pour un modèle à proportions homogènes, nous supposons que

$$\boldsymbol{n}_t \,|\, \boldsymbol{p} \overset{\text{ind}}{\sim} \text{Multinomiale}(n_t, \boldsymbol{p}), t = 1, \ldots, T, \text{ et } \boldsymbol{p} \sim \text{Dirichlet}(\mathbf{1}).$$

Alors, le facteur de Bayes d'hétérogénéité contre homogénéité est

$$\text{FB} = \frac{1}{\{(rc-1)!\}^2}\left[\prod_{j=1}^{r}\prod_{k=1}^{c}\left\{\frac{\prod_{t=1}^{T} n_{tjk}!}{(\sum_{t=1}^{T} n_{tjk})!}\right\}\right]\frac{\{\sum_{t=1}^{T} n_t + rc - 1\}!}{\prod_{t=1}^{T}(n_t + rc - 1)!}.$$

Donc, en utilisant les effectifs corrigés de cellule, le logarithme du facteur de Bayes (LFB) est approximativement égal à 12,4, ce qui est un indice très prononcé d'hétérogénéité et appuie notre modèle de non-réponse variable en fonction du temps.

De manière comparable, nous avons calculé les facteurs de Bayes de $\boldsymbol{p}_1 = \boldsymbol{p}_2 \neq \boldsymbol{p}_3$ ou $\boldsymbol{p}_1 \neq \boldsymbol{p}_2 = \boldsymbol{p}_3$ comparativement à l'homogénéité; les LFB sont de 7,6 et 4,4, respectivement. Donc, la variation en fonction du temps a lieu pour les deux périodes, c'est-à-dire de janvier à avril et d'avril à octobre.

3. Méthodologie

Nous avons construit un modèle de non-réponse non ignorable variable en fonction du temps pour les données du sondage de l'État de l'Ohio de 1998. Aux fins de comparaison, nous avons également envisagé deux autres

modèles, c'est-à-dire un modèle de non-réponse ignorable et un modèle de non-réponse non ignorable. Ces deux derniers modèles ne varient pas en fonction du temps, parce que nous supposons que les trois points dans le temps proviennent du même processus stochastique (autrement dit, qu'il n'y a pas de corrélation entre les périodes). Notre principale contribution est le modèle de non-réponse variable en fonction du temps. Nous avons utilisé les modèles de non-réponse ignorable et non ignorable pour un seul point dans le temps dans Nandram et coll. (2005). Bien que ces deux modèles ne conviennent pas dans le présent contexte, il s'agit de choix naturels pour justifier notre modèle de non-réponse non ignorable variable en fonction du temps. Essentiellement, nous partons du modèle de non-réponse ignorable que nous étendons à un modèle de non réponse non ignorable, lequel est à son tour étendu à un modèle variable en fonction du temps.

Dans la sélection par CA, nous recourons à la stratification et à la mise en grappes pour réduire l'excédent de numéros artificiels. Nous utilisons une stratification selon le code régional et certains indicatifs de central; un classement par ordre géographique (État ou région) avec sélection systématique fournit une stratification implicite des indicatifs de central. Si l'un de ceux-ci est utilisé pour former une strate, il existe encore des dizaines de milliers de numéros à composer, c'est-à-dire un gaspillage encore important avec de nombreux numéros artificiels. La procédure de Mitofsky-Waksberg (voir Waksberg 1978) est un plan d'échantillonnage en grappes à deux degrés stratifié utilisé pour réduire la quantité de numéros artificiels. Les zones locales sont divisées en parties égales et un échantillon aléatoire des indicatifs de central est tiré avec remise parmi ceux admissibles (d'après la mesure de taille de chaque zone locale). Dans certaines zones locales, un nombre fixe de numéros de téléphone sont produits au hasard, sans remise, et composés. Donc, nous avons des probabilités différentielles de sélection (c'est-à-dire des tailles de grappe inégales) dont il faut tenir compte dans une analyse complète. Il existe d'autres variantes de cette procédure. La méthode CA était adéquate au moment des élections de l'Ohio de 1998, mais étant donné les innovations technologiques récentes (par exemple, téléphone cellulaire, courrier électronique, Internet), son utilité pourrait avoir diminué. Dans le présent article, notre méthode et nos modèles ne comprennent pas de stratification, ni de mise en grappes ni de probabilités différentielles de sélection.

Nous utilisons nos modèles pour estimer les proportions d'électeurs qui voteront pour Fisher, Taft et d'autres au moment du sondage d'octobre. Puis, en supposant qu'aucun changement catastrophique n'aura lieu au moment de l'élection de novembre, nous prédisons les proportions d'électeurs qui voteront pour Fisher, Taft et d'autres. De cette manière, nous pouvons prédire le gagnant de l'élection de novembre. Nous trouvons fort séduisante la suggestion faite par un examinateur d'utiliser un modèle de mélange pour couvrir la possibilité d'une catastrophe.

Aux sections 3.1 et 3.2, nous décrivons la notation et les trois modèles. À la section 3.3, nous montrons comment ajuster le modèle de non-réponse non ignorable variable en fonction du temps. Les modèles de non-réponse ignorable et non ignorable peuvent être ajustés de la même manière (voir Nandram et coll., 2005, pour des précisions). À la section 3.4, nous montrons comment spécifier les deux paramètres (μ_0 et c_0^2), et à la section 3.5, comment calculer l'estimation pour le sondage d'octobre et la prédiction pour l'élection de novembre.

3.1 Notation

Soit $I_{tjk\ell} = 1$ si le ℓ^{e} électeur appartient à la j^{e} ligne et à la k^{e} colonne dans le tableau à double entrée au temps t et $I_{tjk\ell} = 0$ sinon, $t = 1, ..., T$, $j = 1, ..., r$, $k = 1, ..., c$, $\ell = 1, ..., L$. Autrement dit, $I_{tjk\ell} = 1$ dénote la cellule du tableau $r \times c$ à laquelle appartient un électeur. Dans notre application, $T = 3$, $r = 2$ et $c = 3$. Soit $J_{ts\ell} = 1$ si le ℓ^{e} électeur se trouve dans le tableau s ($s = 1, 2, 3, 4$) et $J_{ts\ell} = 0$ sinon, $s = 1, ..., 4$, $\sum_{s=1}^{4} J_{ts\ell} = 1$; $J_{ts\ell}$ indique à quel tableau appartient un individu et $\boldsymbol{J}_{t\ell} = (J_{t1\ell}, J_{t2\ell}, J_{t3\ell}, J_{t4\ell})$.

Soit $y_{tsjk} = \sum_{\ell=1}^{n} I_{tjk\ell} J_{ts\ell}$, $s = 1, 2, 3, 4$ les effectifs de cellule pour les quatre tableaux à chaque sondage. Ici, les y_{t1jk} sont observées et les $y_{tsjk}, s = 2, 3, 4, t = 1, ..., T$ manquent (c'est-à-dire sont des variables latentes). Pour y_{t1jk}, nous savons que $\sum_{j=1}^{r}\sum_{k=1}^{c} y_{t1jk} = n_{t0}$, le nombre d'individus pour lesquels les données sont complètes. Pour y_{t2jk}, nous savons que $\sum_{k=1}^{c} y_{t2jk} = u_{tj}$, où les totaux de marge des lignes $u_{tj}, j = 1, ..., r$ sont observés. Pour y_{t3jk}, nous savons que $\sum_{j=1}^{r} y_{t3jk} = v_{tk}$, où les totaux de marge des colonnes $v_{tk}, k = 1, ..., c$ sont observés. Pour y_{t4jk}, nous savons que $\sum_{j=1}^{r}\sum_{k=1}^{c} y_{t4jk} = w_t$ (non-réponse totale). Dans la présente analyse, n_{t0}, $\boldsymbol{u}_t$, $\boldsymbol{v}_t$ et w_t sont maintenus fixes (c'est-à-dire analyse avec totaux de marge fixes) et sont connus.

Partout où cela est commode, nous utiliserons des notations telles que

$$\sum_{s,j,k} y_{tsjk} \equiv \sum_{s=1}^{4}\sum_{j=1}^{r}\sum_{k=1}^{c} y_{tsjk}, \quad \prod_{s,j,k} \pi_{tsjk} \equiv \prod_{s=1}^{4}\prod_{j=1}^{r}\prod_{k=1}^{c} \pi_{tsjk}$$

et $\boldsymbol{y}_{t(1)} = (\boldsymbol{y}_{t2}, \boldsymbol{y}_{t3}, \boldsymbol{y}_{t4})$, $\boldsymbol{y}_{t(2)} = (\boldsymbol{y}_{t1}, \boldsymbol{y}_{t3}, \boldsymbol{y}_{t4})$, *etc.*, où $\boldsymbol{y}_{ts} = (y_{tsjk}, j = 1, ..., r, k = 1, ..., c, t = 1, ..., T, s = 1, 2, 3, 4)$. En outre, nous posons que $\boldsymbol{y}_1 = (\boldsymbol{y}_{11}, ..., \boldsymbol{y}_{T1})$ et $\boldsymbol{y}_{(1)} = (\boldsymbol{y}_{1(1)}, ..., \boldsymbol{y}_{T(1)})$ avec $\boldsymbol{y}_{(1)} = (\boldsymbol{y}_{t(1)}, ..., \boldsymbol{y}_{t(4)})$. De plus, $\sum_{s,j,k}^{4,r,c} y_{tsjk} = n_t$. Nous utiliserons aussi $y_{ts\cdot\cdot} = \sum_{j,k} y_{tsjk}$, $y_{t\cdot jk} = \sum_{s} y_{tsjk}$, *etc.*, $\boldsymbol{y}_t = (\boldsymbol{y}_{t1}, \boldsymbol{y}_{t2}, \boldsymbol{y}_{t3}, \boldsymbol{y}_{t4})$ et $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_T)$.

3.2 Modèles de non-réponse

En posant que $\boldsymbol{I}_{t\ell} = (I_{tjk\ell}, t = 1,...,T, j = 1, ..., r, k = 1, ..., c, \ell = 1, ..., L)$, pour tous les modèles, nous prenons

$$\boldsymbol{I}_{t\ell} \mid \boldsymbol{p}_t \overset{\text{iid}}{\sim} \text{Multinomiale}\{1, \boldsymbol{p}_t\}, \tag{1}$$

où

$$\sum_{j=1}^{r}\sum_{j=1}^{c} p_{tjk} = 1,\ p_{tjk} \geq 0, t = 1, ..., T, j = 1, ..., r, k = 1, ..., c.$$

Pour le modèle de non-réponse ignorable, nous prenons

$$\boldsymbol{J}_{t\ell} \mid \boldsymbol{\pi}_t \overset{\text{iid}}{\sim} \text{Multinomiale}\{1, \boldsymbol{\pi}_t\}. \tag{2}$$

Autrement dit, il n'y a aucune dépendance à l'égard de la situation de cellule d'un individu. Pour les modèles de non-réponse non ignorable, nous prenons

$$\boldsymbol{J}_{t\ell} \mid \{I_{tjk\ell} = 1, I_{tj'k'\ell} = 0, j \neq j', k \neq k', \boldsymbol{\pi}_{tjk}\} \overset{\text{iid}}{\sim} \text{Multinomiale}\{1, \boldsymbol{\pi}_{tjk}\}. \tag{3}$$

L'hypothèse (3) spécifie que les probabilités qu'un individu appartienne à l'un des quatre tableaux dépend des deux caractéristiques (c'est-à-dire classifications de ligne et de colonne) de cet individu. De cette manière, nous intégrons l'hypothèse que les données manquantes sont non ignorables. Notons, sachant les paramètres spécifiés dans (1) à (3), que le comportement d'un électeur est corrélé à celui d'un autre à un même temps *t*, mais que ces comportements sont indépendants au cours du temps. Il mérite d'être souligné ici que les paramètres de (2) sont identifiables, mais que ceux de (3) ne le sont pas. C'est là que survient la difficulté dans le modèle de non-réponse non ignorable et il convient d'y accorder une attention particulière.

Il découle de (1) et de (2) que, pour le modèle ignorable,

$$g(\boldsymbol{p}, \boldsymbol{\pi} \mid \boldsymbol{y}) \propto \prod_{t=1}^{T}\left[\left[\prod_{s=1}^{4}\pi_{ts}^{y_{ts\cdot\cdot}}\right]\left[\prod_{s=1}^{4}\prod_{j=1}^{r}\prod_{k=1}^{c}\frac{p_{tjk}^{y_{tsjk}}}{y_{tsjk}!}\right]\right] \tag{4}$$

sachant $\sum_{k=1}^{c} y_{t2jk} = u_{tj},\ j = 1, ..., r, \sum_{j=1}^{r} y_{t3jk} = v_{tk}, k = 1, ..., c$ et $\sum_{j=1}^{r}\sum_{k=1}^{c} y_{t4jk} = w_t$. Notons que, sous des conditions d'ignorabilité, la fonction de vraisemblance dans (4) se divise en deux parties, l'une contenant les π_{ts} uniquement et l'autre, les p_{tjk}, et que les inférences au sujet de ces paramètres ne sont pas liées; voir la section 3.2 de Nandram et coll. (2005) pour la discussion originale de ce modèle. En outre, il découle de (1) et de (3) que, pour les modèles de non-réponse non ignorable, la fonction de vraisemblance augmentée pour $\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1$ est

$$g(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1) \propto \prod_{t=1}^{T}\left[\prod_{s,j,k}^{4,r,c}\frac{\pi_{sjk}^{y_{tsjk}}}{y_{tsjk}!}\right]\left[\prod_{j,k}^{r,c} p_{tjk}^{y_{t\cdot jk}}\right] \tag{5}$$

sachant $\sum_{k=1}^{c} y_{t2jk} = u_{tj},\ j = 1, ..., r,\ \sum_{j=1}^{r} y_{t3jk} = v_{tk}, k = 1, ..., c,$ et $\sum_{j=1}^{r}\sum_{k=1}^{c} y_{t4jk} = w_t$; voir Nandram et coll. (2005) pour une description de l'identifiabilité dans une situation semblable.

Pour les modèles de non-réponse ignorable et non ignorable, nous prenons

$$\boldsymbol{p}_t \mid \boldsymbol{\mu}_2, \tau_2 \overset{\text{iid}}{\sim} \text{Dirichlet}\,(\boldsymbol{\mu}_2\,\tau_2), t = 1,..., T+1, \tag{6}$$

où nous considérons la prédiction au temps $T+1$, c'est-à-dire une étape plus loin (novembre). La structure probabiliste de (6) permet un « emprunt d'information » au cours du temps. Notons que le vecteur $\boldsymbol{x}$ de dimension *k* suit une loi de Dirichlet si $p(\boldsymbol{x}|\boldsymbol{\alpha}) = \prod_{j=1}^{k} x_j^{\alpha_j - 1}/D(\boldsymbol{\alpha})$, $x_j \geq 0,\ j = 1,...,k, \sum_{j=1}^{k} x_j = 1$, où $D(\boldsymbol{\alpha})$ est la fonction de Dirichlet et $\alpha_j > 0,\ j = 1,...,k$. Pour une référence rapide, consulter Ghosh et Meeden (1997, pages 42, 50 et 127) en ce qui concerne la distribution dans l'urne de Pólya et, plus exactement, son utilisation en tant que loi *a priori* conjuguée dans l'échantillonnage multinomial; depuis la publication de notre premier article (c'est-à-dire Nandram 1998), nous avons utilisé extensivement la combinaison Dirichlet-Multinomiale dans nos travaux de recherche.

Décrivons maintenant les modèles stochastiques pour les $\boldsymbol{\pi}_{tjk}$. Pour le modèle de non-réponse ignorable, nous prenons

$$\boldsymbol{\pi}_t \overset{\text{iid}}{\sim} \text{Dirichlet}\,(\boldsymbol{1}),\ t = 1, ..., T, \tag{7}$$

où **1** est un vecteur de dimension quatre de valeurs unitaires. Nous avons besoin de (7), parce que T est petit (c'est-à-dire $T = 3$ dans notre application). Donc, nous utilisons la loi *a priori* uniforme dans R^4 (essentiellement non informatif); sinon, nous devrons spécifier les paramètres inconnus de la loi de Dirichlet en ne disposant pratiquement d'aucune données. Pour les modèles de non-réponse non ignorable, nous prenons

$$\boldsymbol{\pi}_{tjk} \mid \boldsymbol{\mu}_1, \tau_1 \overset{\text{iid}}{\sim} \text{Dirichlet}\,(\boldsymbol{\mu}_1\tau_1), t = 1,...,T,\ j = 1,...,r, k = 1,...,c. \tag{8}$$

D'abord, nous notons que (8) fournit un « emprunt d'information » au cours du temps. Fait plus important encore, parce que les $\boldsymbol{\pi}_{tjk}$ ne sont pas identifiables, il en est de même de $\boldsymbol{\mu}_1$ et de τ_1. Un moyen éventuel de résoudre ce dilemme est de « centrer » le modèle de non-réponse non ignorable sur le modèle de non-réponse ignorable.

Pour le modèle de non-réponse variable en fonction du temps, nous prenons

$$\boldsymbol{p}_t \mid \boldsymbol{p}_{t-1}, \tau_2 \overset{\text{iid}}{\sim} \text{Dirichlet}\,(\boldsymbol{p}_{t-1}\,\tau_2), t = 1,...,T+1, \tag{9}$$

où $\boldsymbol{p}_0$ est également connue. Notons que

$$E\{\boldsymbol{p}_t \mid \boldsymbol{p}_{t-1}, \tau_2\} = \boldsymbol{p}_{t-1}, t = 1, ..., T+1;$$

de sorte que $\{p_t\}$, *a priori*, est une martingale vectorielle. Ici, T est petit (c'est-à-dire $T = 3$). Donc, cette structure variable en fonction du temps semble plus appropriée et pourrait accroître la précision. Notons également que nous avons pris $\boldsymbol{p}_0 \sim$ Dirichlet (**1**).

Enfin, nous spécifions les densités *a priori* pour les hyperparamètres. Premièrement, nous prenons

$$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \overset{\text{iid}}{\sim} \text{Dirichlet } (\mathbf{1}), \tag{10}$$

c'est-à-dire des densités *a priori* essentiellement non informatives.

Enfin, τ_1 et τ_2 sont des variables aléatoires indépendantes et identiquement distribuées provenant de

$$f(x) = 1/(1+x)^2, \; x \geq 0. \tag{11}$$

De nouveau, il s'agit d'une densité *a priori* essentiellement non informative. Notons que $\boldsymbol{\mu}_1$ et τ_1 n'existent pas dans le modèle de non-réponse ignorable. Gelman (2006) a recommandé d'utiliser des lois *a priori* tels que (11) au lieu des priors gamma diffus appropriés dont le comportement est mauvais.

Pour les modèles de non-réponse non ignorable, nous devons spécifier plus prudemment la densité *a priori* de τ_1, parce que les $\boldsymbol{\pi}_{tjk}$ ne sont pas identifiables. Ici, nous essayons de « centrer » les modèles de non-réponse non ignorable sur le modèle de non-réponse ignorable. Dans (8), le paramètre τ_1 nous renseigne sur la mesure dans laquelle le modèle de non-réponse non ignorable est proche du modèle de non-réponse ignorable. Par exemple, si τ_1 est petit, les $\boldsymbol{\pi}_{tjk}$ seront très différentes et si τ_1 est grand, les $\boldsymbol{\pi}_{tjk}$ seront fort semblables. Donc, l'inférence *a priori* sera sensible au choix de τ_1, qui doit donc être fait avec prudence. Nous aimerions choisir pour τ_1 une densité *a priori* telle que le modèle de non-réponse non ignorable reste proche du modèle de non-réponse ignorable. Donc, nous prenons

$$\tau_1 \sim \text{Gamma } (1/c_0^2, \; 1/\mu_0 c_0^2), \tag{12}$$

où $E(\tau_1) = \mu_0$ et $\text{CV}(\tau_1) = c_0$, CV étant le coefficient de variation; nous devons spécifier μ_0 ainsi que c_0. Nous utilisons la loi *a priori* (12), parce que, par un choix approprié de μ_0 et c_0, il est possible de centrer le modèle de non-réponse non ignorable sur le modèle de non-réponse ignorable. Naturellement, d'autres lois *a priori* appropriées commodes comprenant des paramètres tels que μ_0 et c_0 peuvent être utilisés pour faciliter le centrage. À la section 3.4, nous utiliserons des échantillons provenant de la densité *a posteriori* de τ_1 sous le modèle de non-réponse ignorable pour spécifier μ_0 et c_0.

Pour chacun des trois modèles, il est facile d'écrire la densité *a priori* conjointe des paramètres. Par exemple, pour le modèle variable en fonction du temps, la densité *a priori* conjointe est

$$p(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\mu}_1, \tau_1, \tau_2) \propto \tau_1^{1/c_0^2 - 1} e^{-\tau_1/\mu_0 c_0^2} \frac{1}{(1+\tau_2)^2}$$
$$\times \prod_{t=1}^{T} \left\{ \frac{\prod_{j=1}^{r} \prod_{k=1}^{c} p_{tjk}^{p_{t-1jk}\tau_2 - 1}}{D(\boldsymbol{p}_{t-1}\tau_2)} \prod_{j=1}^{r} \prod_{k=1}^{c} \frac{\prod_{s=1}^{4} \pi_{tsjk}^{\mu_{1s}\tau_1 - 1}}{D(\boldsymbol{\mu}_1 \tau_1)} \right\}, \tag{13}$$

où $D(\cdot)$ est la fonction de Dirichlet.

3.3 Ajustement du modèle de non-réponse non ignorable variable en fonction du temps

Si nous combinons la fonction de vraisemblance (5) et la densité *a priori* conjointe (13) par la voie du théorème de Bayes, nous obtenons la densité *a posteriori* conjointe des paramètres $\boldsymbol{\pi}$, $\boldsymbol{p}$ $\boldsymbol{\mu}_1$, τ_1 τ_2 et des variables latentes $\boldsymbol{y}_{(1)}$ qui suit

$$\pi(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\mu}_1, \tau_1, \tau_2, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1)$$
$$\propto \tau_1^{1/c_0^2 - 1} e^{-\tau_1/\mu_0 c_0^2} \frac{1}{(1+\tau_2)^2} \prod_{t=1}^{T} \left[\prod_{s,j,k}^{4,r,c} \frac{\pi_{tsjk}^{y_{tsjk}}}{y_{tsjk}!} \right] \left[\prod_{j,k}^{r,c} p_{tjk}^{y_{t\cdot jk}} \right]$$
$$\times \prod_{t=1}^{T} \left\{ \frac{\prod_{j=1}^{r} \prod_{k=1}^{c} p_{tjk}^{p_{t-1jk}\tau_2 - 1}}{D(\boldsymbol{p}_{t-1}\tau_2)} \prod_{j=1}^{r} \prod_{k=1}^{c} \frac{\prod_{s=1}^{4} \pi_{tsjk}^{\mu_{1s}\tau_1 - 1}}{D(\boldsymbol{\mu}_1 \tau_1)} \right\} \tag{14}$$

sachant $\sum_{k=1}^{c} y_{t2jk} = u_{tj}$, $j = 1, ..., r$, $\sum_{j=1}^{r} y_{t3jk} = v_{tk}$, $k = 1, ..., c$, et $\sum_{j=1}^{r} \sum_{k=1}^{c} y_{t4jk} = w_t$, $t = 1, ..., T$.

La densité *a posteriori* (14) est complexe, de sorte que nous utiliserons des méthodes de Monte Carlo par chaîne de Markov pour l'ajuster. Cependant, il est facile d'ajuster le modèle variable en fonction du temps en utilisant l'échantillonneur matriciel de Metropolis-Hastings comme nous le décrirons. En outre, de manière semblable, en utilisant l'échantillonneur de Gibbs (Ritter et Tanner 1992), il est facile d'ajuster les modèles de non-réponse ignorable et non ignorable. Nous tirons un échantillon à partir de la densité *a posteriori* conjointe afin de conduire une inférence au sujet des paramètres. Plus précisément, nous devons faire une inférence au sujet de $\boldsymbol{p}_t$. Pour exécuter l'échantillonneur de Metropolis-Hastings, nous avons besoin de la densité *a posteriori* conditionnelle de chacun des paramètres, sachant les autres.

En premier lieu, nous considérons les fonctions de masse de probabilité *a posteriori* conditionnelles de $\boldsymbol{y}_{ts}$, $s = 2, 3, 4$, $t = 1, ..., T$, sachant $\boldsymbol{y}_{t(s)}$, $\boldsymbol{p}_t$, $\boldsymbol{\pi}_{tjk}$, $j = 1, ..., r$, $k = 1, ..., c$. De (14), il découle clairement que, sous la densité *a posteriori* conditionnelle, les $\boldsymbol{y}_{ts}$, $t = 1, ..., T$, $s = 2, 3, 4$, sont des vecteurs aléatoires multinomiaux indépendants. En

particulier, en posant que $\boldsymbol{p}=(p_{tjk}, t=1,...,T,\ j=1,...,r, k=1,...,c)$ et $\boldsymbol{\pi}=(\boldsymbol{\pi}_{tjk}, t=1,...,T,\ j=1...,r, k=1,...,c)$,

$$\boldsymbol{y}_{t2j} \mid \{\boldsymbol{y}_{t1}, \boldsymbol{p}, \boldsymbol{\pi}\} \overset{\text{ind}}{\sim} \text{Multinomiale}(u_{tj}, \boldsymbol{q}_{tj}^{(2)}) \quad j=1,...,r),$$

$$\boldsymbol{y}_{t3k} \mid \{\boldsymbol{y}_{t1}, \boldsymbol{p}, \boldsymbol{\pi}\} \overset{\text{ind}}{\sim} \text{Multinomiale}(v_{tk}, \boldsymbol{q}_{tk}^{(3)}),\ k=1,...,c),$$

$$\boldsymbol{y}_{t4} \mid \{\boldsymbol{y}_{t1}, \boldsymbol{p}, \boldsymbol{\pi}\} \sim \text{Multinomiale}(w_t, \boldsymbol{q}_t^{(4)}), \qquad (15)$$

où $q_{tjk}^{(2)} = \pi_{t2jk}\, p_{tjk} / \sum_{k'=1}^{c} \pi_{t2jk'}\, p_{tjk'}$, $k=1,...,c$, $q_{tjk}^{(3)} = \pi_{t3jk}\, p_{tjk} / \sum_{j'=1}^{r} \pi_{t3j'k}\, p_{tj'k}$, $j=1,...,r$ et $q_{tjk}^{(4)} = \pi_{t4jk} p_{tjk} / \sum_{j'=1}^{r}\sum_{k'=1}^{c} \pi_{t4j'k'} p_{tj'k'}$, $j=1,...,r$, $k=1,...,c$, $t=1,...,T$.

La densité *a posteriori* conditionnelle de $\boldsymbol{\pi}_{tjk}$ est donnée par

$$\boldsymbol{\pi}_{tjk} \mid \{\boldsymbol{\mu}, \tau, \boldsymbol{y}_t\} \overset{\text{ind}}{\sim} \text{Dirichlet}(y_{t1jk}+\mu_{t1}\tau_1, y_{t2jk}+\mu_{t2}\tau_1, y_{t3jk}+\mu_{t3}\tau_1, y_{t4jk}+\mu_{t4}\tau_1) \qquad (16)$$

avec indépendance sur $t=1,...,T,\ j=1,...,r, k=1,...,c$.

La densité *a posteriori* conditionnelle pour $\boldsymbol{p}_t$, $t=1,...,T$ est plus compliquée. Nous notons que

$$\pi(\boldsymbol{p}_0 \mid \text{sinon}, \boldsymbol{y}_1) \propto \frac{\prod_{j=1}^{r}\prod_{k=1}^{c} p_{1jk}^{p_{0jk}\tau_2 - 1}}{D(\boldsymbol{p}_0\tau_2)} \qquad (17)$$

et

$$\pi(\boldsymbol{p}_t \mid \text{sinon}, \boldsymbol{y}_1) \propto \left\{\prod_{j=1}^{r}\prod_{k=1}^{c} p_{tjk}^{y_{t\cdot jk}+p_{t-1jk}\tau_2-1}\right\} \frac{\prod_{j=1}^{r}\prod_{k=1}^{c} p_{t+1jk}^{p_{tjk}\tau_2-1}}{D(\boldsymbol{p}_t\tau_2)},\ t=1,...,T, \qquad (18)$$

où « sinon » fait référence à tous les paramètres compris dans $(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\mu}_1, \tau_1, \tau_2, \boldsymbol{y}_{(1)})$ sauf $\boldsymbol{p}_0$ dans (17) ou $\boldsymbol{p}_t$ dans (18). Nous montrons comment tirer les échantillons de (17) et de (18) à l'annexe A.

Ensuite, nous considérons les hyperparamètres. En posant que $\delta_s = \prod_{t=1}^{T}\prod_{j=1}^{r}\prod_{k=1}^{c} \pi_{sjk}$, et $\boldsymbol{\pi}=(\pi_{tjk}, t=1,...,T, j=1,...,r,\ k=1,...,c)$, la densité *a posteriori* conditionnelle conjointe de $\boldsymbol{\mu}_1, \tau_1$ est

$$p(\boldsymbol{\mu}_1, \tau_1 \mid \boldsymbol{\pi}) \propto \frac{\prod_{s=1}^{4} \delta_s^{\mu_{1s}\tau_1}}{\{D(\boldsymbol{\mu}_1\tau_1)\}^{rcT}} \tau_1^{1/c_0^2-1} e^{-\tau_1/\mu_0 c_0^2},$$

où

$$\sum_{s=1}^{4} \mu_{1s} = 1,\ \mu_{1s} \geq 0, s=1,2,3,4, \tau_1 > 0.$$

Il n'est pas nécessaire d'obtenir un échantillon directement de $p(\boldsymbol{\mu}_1 \mid \tau_1, \boldsymbol{\pi})$. Toutefois, en représentant par $\boldsymbol{\mu}_{1(s)}$ le vecteur de toutes les composantes de $\boldsymbol{\mu}_1$ sauf μ_{1s}, nous avons

$$p(\mu_{1s} \mid \boldsymbol{\mu}_{1(s)}, \tau_1, \boldsymbol{\pi}) \propto \frac{\delta_s^{\mu_{1s}\tau_1}}{\{\Gamma(\mu_{1s}\tau_1)\}^{rcT}} \frac{\delta_4^{(1-\mu_{11}-\mu_{12}-\mu_{13})\tau_1}}{\{\Gamma((1-\mu_{11}-\mu_{12}-\mu_{13})\tau_1)\}^{rcT}},$$

$$0 \leq \mu_{1s} \leq 1 - \sum_{s'=1,\, s'\neq s}^{3} \mu_{1s'},\ s=1,2,3. \qquad (19)$$

Nous utilisons une méthode de grille pour tirer un échantillon à partir de $p(\mu_{1s} \mid \boldsymbol{\mu}_{1(s)}, \tau_1, \boldsymbol{\pi})$. Nous commençons par utiliser 50 mailles (autrement dit, nous divisons l'étendue de μ_{1s}, $(0, 1-\sum_{s'=1, s'\neq s}^{3}\mu_{1s'})$, en 50 intervalles égaux) pour former une fonction de masse de probabilité approximative de μ_{1s}, $s=1,2,3$. Nous tirons d'abord une variable aléatoire de la fonction de masse de probabilité pour indiquer lequel des 50 intervalles est sélectionné. Puis, pour μ_{1s}, nous tirons une variable aléatoire uniforme dans cet intervalle. Cette procédure est efficace, parce que μ_{1s} est bornée, que les intervalles sont très étroits et qu'il est très « peu coûteux » de construire la fonction de masse de probabilité discrète pour chaque μ_{1s}, $s=1,2,3$. Enfin, nous obtenons μ_{14} à partir de sa densité *a posteriori* conditionnelle en prenant $\mu_{14} = 1 - \sum_{s=1}^{3} \mu_{1s}$.

La densité *a posteriori* conditionnelle de τ_1 est

$$p(\tau_1 \mid \boldsymbol{\mu}_1, \boldsymbol{\pi}) \propto \left[\prod_{s=1}^{4} \frac{\delta_s^{\mu_{1s}\tau_1}}{\{\Gamma(\mu_{1s}\tau_1)\}^{rcT}}\right] \tau_1^{1/c_0^2-1} e^{-\tau_1/\mu_0 c_0^2},\ \tau_1 > 0. \qquad (20)$$

Pour tirer une valeur aléatoire de (20), nous procédons de la même manière que pour (19), sauf que nous transformons τ_1 de la moitié positive de la droite réelle en $(0,1)$. (Il est plus commode d'obtenir une approximation par grille d'une densité dans un intervalle borné.) Donc, en posant que $\tau_1 = \phi/(1-\phi)$ dans (20), nous avons

$$p(\phi \mid \boldsymbol{\mu}_1, \boldsymbol{\pi}) \propto \frac{1}{(1-\phi)^2} \left\{\left[\prod_{s=1}^{4} \frac{\delta_s^{\mu_{1s}\tau_1}}{\{\Gamma(\mu_{1s}\tau_1)\}^{rcT}}\right] \tau_1^{1/c_0^2-1} e^{-\tau_1/\mu_0 c_0^2}\right\}_{\tau_1=\frac{\phi}{1-\phi}},\ 0<\phi<1.$$

De nouveau, nous commençons par utiliser 50 intervalles de même largeur pour tirer ϕ, et la valeur aléatoire pour τ_1 est $\phi/(1-\phi)$.

En posant que $p=(p_{tjk}, t=1,...,T, j=1,...,r, k=1,...,c)$, la densité *a posteriori* conditionnelle de τ_2 est

$$\pi(\tau_2 \mid \boldsymbol{p}) \propto \frac{1}{(1+\tau_2)^2} \prod_{t=1}^{T} \left\{\frac{\prod_{j=1}^{r}\prod_{k=1}^{c} p_{tjk}^{p_{t-1jk}\tau_2-1}}{D(\boldsymbol{p}_{t-1}\tau_2)}\right\},\ \tau_2>0. \qquad (21)$$

Nous obtenons un échantillon de manière comparable à τ_1 dans (20).

Nous possédons une vaste expérience de l'utilisation de l'approximation par grille. Cependant, il faut être prudent lorsque l'on applique cette méthode pour des paramètres dont les valeurs sont proches de 0 ou de 1 dans l'intervalle

[0, 1]. Il est alors nécessaire d'utiliser une approximation par grille dans un intervalle proche de la borne, ce que l'on peut obtenir par tâtonnement en examinant la sortie de l'échantillonneur à mesure que son exécution progresse. S'il est probable qu'un paramètre compris dans [0, 1] s'écarte de 0 ou de 1, alors la méthode par grille convient bien; cela est le cas pour les μ_{1s}. Cependant, lorsqu'un paramètre tel que τ_1 (qui peut être très grand) est transformé en ϕ dans l'intervalle [0, 1], ϕ peut être très grand (proche de 1). Si la valeur transformée est de l'ordre de 0,999, il est nécessaire d'ajuster la recherche par grille afin qu'elle se fasse dans un intervalle contenant 0,999. Cet ajustement doit se faire par tâtonnement; il faut examiner la sortie de ϕ à mesure que progresse l'exécution de l'échantillonneur et ajuster l'intervalle en conséquence. Par exemple, si l'intervalle [0, 1] contient 100 points de grille équidistants tels que 0,01, 0,02, 0,03, ..., 0,99, et que la valeur du paramètre se situe vraisemblablement autour de 0,999, même si nous procédons à un tirage uniforme dans l'intervalle de grille choisi, ces points de grille ne seront pas très efficaces.

Nous avons exécuté l'échantillonneur de Metropolis-Hastings en tirant une valeur aléatoire de chacune des densités (15), (16), (17), (18), (19), (20) et (21) en itérant la procédure complète jusqu'à la convergence. Il s'agit d'un exemple de l'échantillonneur Metropolis-Hastings (Ritter et Tanner 1992). Nous avons tiré un échantillon des densités *a posteriori* correspondant aux modèles de non-réponse ignorable et non ignorable de la même manière. Pour tous les modèles, nous avons utilisé un échantillon de $M = 1\,000$ tiré des densités *a posteriori* pour l'estimation et la prédiction. Nous avons surveillé la convergence de l'algorithme en examinant les tracés graphiques de chaque paramètre en fonction de l'ordre des itérations et étudié le coefficient d'autocorrélation. Nous avons utilisé un échantillonneur de Gibbs pour ajuster les modèles de non-réponse ignorable et non ignorable. Nous avons procédé à 1 000 itérations « d'apprentissage », puis nous avons sélectionné une itération sur dix. Cette procédure donne de bons résultats.

Cependant, pour le modèle variable en fonction du temps, nous avons utilisé une approximation par grille de la loi *a posteriori* conditionnelle de $\boldsymbol{p}_0$, mais les pas de Metropolis pour $\boldsymbol{p}_t, t = 1,...,T$. Les pas de Metropolis n'ont pas donné de bons résultats, parce que les probabilités de saut étaient 0,67, 0,65 et 0,73 pour les trois densités *a posteriori* conditionnelles de $\boldsymbol{p}_1$, $\boldsymbol{p}_2$ et $\boldsymbol{p}_3$, alors qu'il est recommandé qu'elles soient comprises entre 0,25 et 0,50 (Gelman, Roberts et Gilks 1996); les réglages fins n'ont produit aucune amélioration. Par conséquent, nous avons également utilisé des approximations par grille de ces trois densités *a posteriori* conditionnelles. Ces approximations sont très précises. Dans chaque cas, nous avons commencé avec 50 mailles et avons augmenté le nombre de mailles jusqu'à ce que nos estimations de tous les paramètres $\boldsymbol{p}_T$, $\boldsymbol{p}_{T+1}$, $\boldsymbol{\mu}_2$, τ_2 ne varient plus. Nous avons constaté que 200 mailles était un nombre adéquat dans tous les cas (c'est-à-dire pour $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \tau_1, \tau_2$). Nous avons aussi découvert que, si l'échantillonneur de Metropolis-Hastings ne fonctionne pas aussi bien que nous l'avions espéré, les estimations des proportions de cellule sont presque les mêmes pour les deux échantillonneurs. Nous avons exécuté 25 000 itérations de l'échantillonneur de Metropolis-Hastings, dont 5 000 itérations d'« apprentissage » suivies d'un élagage par sélection d'une itération sur vingt.

Enfin, nous avons mémorisé l'échantillon tiré de la densité *a posteriori* conjointe en prévision de futures analyses. Plus précisément, pour les modèles de non-réponse ignorable et non ignorable, nous avons besoin de l'échantillon de taille M tiré de $\{(\boldsymbol{\mu}_2^{(h)}, \tau_2^{(h)}, \boldsymbol{p}_{T-1}^{(h)}, \boldsymbol{p}^{(h)}), h = 1,...,M\}$, et, pour le modèle variable en fonction du temps, nous avons besoin de l'échantillon de taille M tiré de $(\tau_2^{(h)}, \boldsymbol{p}_{T-1}^{(h)}, \boldsymbol{p}^{(h)}), h = 1,...,M$.

3.4 Spécification de $\boldsymbol{\mu_0}$ et c_0^2

Enfin, nous décrivons comment spécifier μ_0 et c_0 dans (12). Cette étape est importante, parce qu'elle nous permet de « centrer » le modèle de non-réponse non ignorable sur le modèle de non-réponse ignorable (c'est-à-dire, un modèle d'expansion). Cette procédure s'inscrit dans la foulée de Nandram et coll. (2005).

Nous avons tiré un échantillon de $\boldsymbol{\pi}_t^{(h)}, t = 1,...,T, h = 1,...,M$, $M = 1\,000$ itérations à partir du modèle de non-réponse ignorable et calculé $\boldsymbol{\pi}^{(h)} = \sum_{t=1}^{T} \boldsymbol{\pi}_t^{(h)}/T, h = 1,...,M$. Puis, en utilisant l'échantillonneur de Gibbs, nous avons ajusté le modèle

$$\boldsymbol{\pi}^{(h)} \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_1\tau_1),$$

$$\boldsymbol{\mu}_1 \sim \text{Dirichlet}(\mathbf{1}),\ p(\tau_1) = 1/(1+\tau_1)^2, \tau_1 > 0,$$

où $\boldsymbol{\mu}_1$ et τ_1 *a priori* sont indépendants, pour obtenir un échantillon $\tau_1^{(h)}, h = 1,...,M$. Nous avons tiré 1 500 itérations avec un « apprentissage » de 500 pour obtenir $M = 1\,000$ itérations.

Enfin, en prenant $a = M^{-1}\sum_{h=1}^{M} \tau_1^{(h)}$ et $b = (M-1)^{-1}\sum_{h=1}^{M} (\tau_1^{(h)} - a)^2$, nous avons fixé que

$$c_0 = \sqrt{b}/a \text{ et } \mu_0 = a.$$

Pour les données électorales, notre méthode donne $c_0 = 0{,}031$ et $\mu_0 = 2{,}431$. Cette spécification maintiendra le modèle de non-réponse non ignorable proche du modèle de non-réponse ignorable, fournissant par conséquent un mécanisme possible de centrage.

Pour étudier la sensibilité à l'erreur de spécification de la densité *a priori* de τ_1, nous utilisons deux constantes, κ_1 et κ_2, telles que, *a priori*

$$\tau_1 \sim \text{Gamma}\,(1/\kappa_1^2\, c_0^2,\ 1/\kappa_1^2\,\kappa_2\,\mu_0\, c_0^2)$$

pour des valeurs variables de κ_1 et κ_2. Il convient de souligner que $E(\tau_1) = \kappa_2\mu_0$ et $\text{CV}(\tau_1) = \kappa_1 c_0$; donc, accroître la valeur de κ_2 signifie accroître celle de τ_1, ce qui, à son tour, signifie accroître la précision *a priori*, mais pas nécessairement *a posteriori*. Nous étudierons la sensibilité à la spécification de κ_1 et de κ_2 quand nous décrirons l'analyse des données.

3.5 Estimation et prédiction

Nous montrons comment améliorer l'estimation (c'est-à-dire, rao-blackwellisation) pour le sondage d'octobre et comment calculer la prédiction pour l'élection de novembre.

Pour les modèles de non-réponse ignorable et non ignorable,

$$g(\boldsymbol{p}_T \mid \boldsymbol{y}_1) = \int g(\boldsymbol{p}_T \mid \boldsymbol{\mu}_2, \tau_2)\,\pi(\boldsymbol{\mu}_2, \tau_2 \mid \boldsymbol{y}_1) d\,\boldsymbol{\mu}_2\, d\,\tau_2$$

$$\approx \frac{1}{M}\sum_{h=1}^{M} g\,(\boldsymbol{p}_T \mid \boldsymbol{\mu}_2^{(h)}, \tau_2^{(h)}), \qquad (22)$$

où $\boldsymbol{p}_T \mid \boldsymbol{\mu}_2, \tau_2 \sim \text{Dirichlet}\,(\boldsymbol{\mu}_2\tau_2)$, et, pour le modèle variable en fonction du temps,

$$g\,(\boldsymbol{p}_T \mid \boldsymbol{y}_1) = \int g(\boldsymbol{p}_T \mid \boldsymbol{p}_{T-1}, \tau_2)\,\pi\,(\boldsymbol{p}_{T-1}, \tau_2 \mid \boldsymbol{y}_1)\, d\,\boldsymbol{p}_T\, d\,\tau_2$$

$$\approx \frac{1}{M}\sum_{h=1}^{M} g(\boldsymbol{p}_T \mid \boldsymbol{p}_{T-1}^{(h)}, \tau_2^{(h)}), \qquad (23)$$

où $\boldsymbol{p}_T \mid \boldsymbol{p}_{T-1}, \tau_2 \sim \text{Dirichlet}\,(\boldsymbol{p}_{T-1}\tau_2)$.

Nous obtenons (prédisons) les proportions de cellule pour novembre comme il suit. Pour le modèle de non-réponse ignorable ou non ignorable, la densité *a posteriori* de $\boldsymbol{p}_{T+1}$ est

$$g\,(\boldsymbol{p}_{T+1} \mid \boldsymbol{y}_1) = \int g(\boldsymbol{p}_{T+1} \mid \boldsymbol{\mu}_2, \tau_2)\,\pi(\boldsymbol{\mu}_2, \tau_2 \mid \boldsymbol{y}_1)\, d\,\boldsymbol{\mu}_2\, d\,\tau_2$$

$$\approx \frac{1}{M}\sum_{h=1}^{M} g\,(\boldsymbol{p}_{T+1} \mid \boldsymbol{\mu}_2^{(h)}, \tau_2^{(h)}), \qquad (24)$$

où $\boldsymbol{p}_{T+1} \mid \boldsymbol{\mu}_2, \tau_2 \sim \text{Dirichlet}\,(\boldsymbol{\mu}_2\,\tau_2)$. Pour le modèle variable en fonction du temps,

$$g(\boldsymbol{p}_{T+1} \mid \boldsymbol{y}_1) = \int g\,(\boldsymbol{p}_{T+1} \mid \boldsymbol{p}_T, \tau_2)\,\pi\,(\boldsymbol{p}_T, \tau_2 \mid \boldsymbol{y}_1)\, d\,\boldsymbol{p}_T\, d\,\tau_2$$

$$\approx \frac{1}{M}\sum_{h=1}^{M} g\,(\boldsymbol{p}_{T+1} \mid \boldsymbol{p}_T^{(h)}, \tau_2^{(h)}), \qquad (25)$$

où $\boldsymbol{p}_{T+1} \mid \boldsymbol{p}_T, \tau_2 \sim \text{Dirichlet}\,(\boldsymbol{p}_T\,\tau_2)$.

Donc, d'après (22), (23), (24) et (25), l'estimation et la prédiction sont simples. Par exemple, considérons le modèle variable en fonction du temps. Dans le cas de l'estimation, pour chaque h, nous tirons de (24) une valeur aléatoire $\boldsymbol{p}_T \mid \boldsymbol{p}_{T-1}^{(h)}, \tau_2^{(h)} \sim \text{Dirichlet}(\boldsymbol{p}_{T-1}^{(h)}\tau_2^{(h)})$, dénotée par $\boldsymbol{p}_T^{(h)}, h = 1, \ldots, M$. Pour la prédiction, pour chaque h, nous tirons de (25) une valeur aléatoire $\boldsymbol{p}_{T+1} \mid \boldsymbol{p}_T^{(h)}$, $\tau_2^{(h)} \sim \text{Dirichlet}\,(\boldsymbol{p}_T^{(h)}\,\tau_2^{(h)})$, dénotée par $\boldsymbol{p}_T^{(h)}, h = 1, \ldots, M$. Donc, l'inférence au sujet de $\boldsymbol{p}_T$ et $\boldsymbol{p}_{T+1}$ est conduite de la manière habituelle. La méthode est semblable pour les modèles de non-réponse ignorable et non ignorable.

4. Analyse des données

Dans cette section, nous comparons nos modèles à ceux de Chen et Stasny (2003) et aux résultats réels (élection de novembre). Nous avons introduit un nouveau paramètre pour faciliter la prédiction du résultat de l'élection. Nous étudions aussi en détail la sensibilité de l'inférence au choix de κ_1 et de κ_2. En nous fondant sur notre méthode, nous avons spécifié le coefficient de variation, $c_0 = 0{,}031$, et la moyenne, $\mu_0 = 2{,}431$, de la distribution *a priori* de τ_1.

Au tableau 2, nous comparons l'inférence au sujet des proportions d'électeurs d'octobre affectées aux trois candidats d'après nos modèles et d'après ceux de Chen et Stasny (2003). Dans ce tableau, les résultats sont fondés sur le prior $\tau_1 \sim \text{Gamma}\,(1/c_0^2, 1/\mu_0\, c_0^2)$ (c'est-à-dire $\kappa_1 = \kappa_2 = 1$). Nous présentons aussi les proportions réelles tirées de Chang et Krosnick (2001). Les proportions réelles sont (0,45, 0,50, 0,05) pour Fisher, Taft et les autres. En utilisant notre modèle de non-réponse variable en fonction du temps, nous estimons que ces proportions sont (0,41, 0,50, 0,09). Ces valeurs se comparent bien aux résultats réels. Les estimations correspondantes sont (0,41, 0,51, 0,08) pour le modèle de non-réponse ignorable et (0,40, 0,50, 0,09) pour le modèle de non-réponse non ignorable. Le meilleur résultat de Chen et Stasny (2003) est celui qu'ils obtiennent à l'aide de leur modèle D qui produit les estimations (0,42, 0,51, 0,07). Nous fournissons des intervalles de crédibilité à 95 % pour nos estimations, mais dans le contexte de l'approche de Chen et Stasny (2003), il est un peu plus difficile de fournir des intervalles comparables. En outre, au tableau 2, nous présentons les estimations des proportions prédites pour l'élection de novembre. Les prédicteurs ponctuels sont comparables aux estimations ponctuelles, sauf pour la proportion prédite pour Taft sous le modèle de non-réponse ignorable. Toutefois, comme prévu, les intervalles de crédibilité à 95 % pour les proportions prédites sont beaucoup plus larges. Par exemple, sous le modèle variable en fonction du temps, l'intervalle de crédibilité à 95 % pour la proportion d'électeurs votant pour Taft au moment du sondage d'octobre est (0,41, 0,60) et pour la prédiction, il est (0,21, 0,78). Par conséquent, bien que les estimations et les prédictions ponctuelles indiquent

le gagnant, la variabilité ne révèle aucune différence entre Taft et Fisher. Nous reviendrons sur ce point plus loin.

Tableau 2
Comparaison de la proportion d'électeurs probables pour le sondage d'octobre 1998 et de la prédiction pour l'élection de novembre 1998 sous divers modèles avec le résultat réel

Intention	Fisher	Taft	Autre
Estimation d'échantillon	0,41	0,51	0,08
IC à 95 % approximatif	(0,35; 0,47)	(0,45; 0,57)	(0,05; 0,11)
Résultat réel	0,45	0,50	0,05
a. Estimation			
Modèles A, B, C de Chen/Stasny	0,41	0,51	0,08
Modèle D de Chen/Stasny	0,42	0,51	0,07
Modèle E de Chen/Stasny	0,41	0,51	0,08
Modèle de non-réponse ignorable	0,41	0,51	0,08
IC à 95 %	(0,35; 0,46)	(0,46; 0,57)	(0,05; 0,12)
Modèle de non-réponse non ignorable	0,41	0,50	0,09
IC à 95 %	(0,32; 0,51)	(0,40; 0,60)	(0,05; 0,17)
Modèle variable en fonction du temps	0,41	0,50	0,09
IC à 95 %	(0,32; 0,52)	(0,41; 0,60)	(0,05; 0,16)
b. Prédiction			
Modèle de non-réponse ignorable	0,41	0,54	0,05
IC à 95 %	(0,15; 0,70)	(0,25; 0,81)	(0,00; 0,22)
Modèle de non-réponse non ignorable	0,42	0,52	0,06
IC à 95 %	(0,15; 0,70)	(0,22; 0,79)	(0,00; 0,28)
Modèle variable en fonction du temps	0,41	0,50	0,09
IC à 95 %	(0,15; 0,71)	(0,21; 0,78)	(0,00; 0,31)

NOTA : $\tau_1 \sim \text{Gamma}(1/c_0^2, 1/\mu_0 c_0^2)$, où $c_0 = 0{,}031$ et $\mu_0 = 2{,}431$.

Bien que nos estimations à l'aide du modèle de non-réponse variable en fonction du temps soient proches des estimations réelles, les intervalles de crédibilité à 95 % pour p_{311} et p_{312} se chevauchent, si bien qu'il est difficile de prédire que Taft est le gagnant. Les intervalles de crédibilité à 95 % sont plus courts pour nos autres modèles, mais les estimations ponctuelles ne sont pas aussi bonnes et ces intervalles se chevauchent encore. L'une des faiblesses de notre analyse au tableau 2 tient au fait que nous n'avons pas tenu compte de la corrélation entre les deux estimations (autrement dit, nous devrions vraiment étudier la différence $p_{312} - p_{311}$, qui est la marge pour gagner).

Au tableau 3, nous présentons les estimations de $\Lambda_e = p_{312} - p_{311}$ et $\Lambda_p = p_{412} - p_{411}$ à $\kappa_1 = \kappa_2 = 1$ pour les trois modèles. Nous avons également inclus l'erreur-type numérique (ETN), qui est une mesure du degré de reproductibilité des résultats numériques; nous l'avons calculée par la méthode des moyennes par lot. De faibles ETN signifient que, si nous répétons le calcul complet de la même façon (c'est-à-dire en utilisant une autre série de 1 000 itérations), nous ne devrions observer que peu de différence entre les deux ensembles de réponse. Au tableau 3, les ETN sont petites. Les estimateurs et les prédicteurs ponctuels sont tous positifs, indiquant que Taft est le gagnant au sondage d'octobre ainsi qu'à l'élection de novembre. Cependant, la variabilité éclipse dans une certaine mesure ce résultat, parce que les écarts-types *a posteriori* (ETP) sont grands, et, comme prévu, encore plus dans le cas des prédictions. Par conséquent, pour les deux paramètres, les intervalles de crédibilité à 95 % contiennent la valeur 0. Donc, de nouveau, si l'on considère la variabilité, aucune différence ne se dégage entre Taft et Fisher.

Tableau 3
Comparaison des trois modèles pour l'estimation et la prédiction en utilisant les moyennes a posteriori (MP), les écarts-types a posteriori (ETP), les erreurs-types numériques (ETN) et les intervalles de crédibilité à 95 % pour Λ_e (Λ_p) et Δ_e (Δ_p)

	Modèle	MP	ETP	ETN	Intervalle
Λ_e	Ignorable	0,105	0,055	0,002	(-0,002; 0,209)
	Non ignorable	0,097	0,099	0,006	(-0,100; 0,280)
	Variable en fonction du temps	0,093	0,101	0,007	(-0,098; 0,276)
Λ_p	Ignorable	0,071	0,154	0,004	(-0,240; 0,362)
	Non ignorable	0,058	0,150	0,005	(-0,252; 0,369)
	Variable en fonction du temps	0,050	0,134	0,005	(-0,244; 0,314)
Δ_e	Ignorable	0,688	0,175	0,008	(0,295; 0,958)
	Non ignorable	0,663	0,200	0,012	(0,222; 0,959)
	Variable en fonction du temps	0,632	0,148	0,014	(0,336; 0,901)
Δ_p	Ignorable	0,688	0,175	0,008	(0,295; 0,960)
	Non ignorable	0,663	0,193	0,009	(0,253; 0,972)
	Variable en fonction du temps	0,648	0,155	0,011	(0,341; 0,923)

NOTA : Voir la note du tableau 2; $\Lambda_e = p_{312} - p_{311}$ (estimation, différence entre Taft et Fisher pour le sondage d'octobre); $\Lambda_p = p_{412} - p_{411}$ (prédiction, différence entre Taft et Fisher pour l'élection de novembre); $\Delta_e = \Pr(p_{312} > p_{311} | p_{311} + p_{312} + p_{313}, \boldsymbol{\alpha})$; et $\Delta_p = \Pr(p_{412} > p_{411} | p_{411} + p_{412} + p_{413}, \boldsymbol{\alpha})$; voir (26).

Nous cherchons un paramètre de rechange qui nous aidera à prédire le gagnant de manière plus convaincante. Nous posons la question suivante : « Quelle est la probabilité que la proportion de personnes votant pour Taft au moment du sondage d'octobre et au moment de l'élection de novembre soit plus grande que celle de personnes votant pour Fisher? »

Donc, nous considérons le paramètre $\Delta_e = \Pr(p_{312} > p_{311} | p_{311} + p_{312} + p_{313}, \boldsymbol{\alpha})$, où $\alpha_{jk} = \mu_{jk}\tau_2$, $j = 1, ..., r$, $k = 1, ..., c$, pour les modèles de non-réponse ignorable et non ignorable, et $\alpha_{jk} = p_{2jk}\tau_2$, $j = 1, ..., r$, $k = 1, ..., c$, pour le modèle de non-réponse variable en fonction du temps. Dans chaque cas, en posant $q_1 = p_{311}/p_{31\cdot}$, $q_2 = p_{312}/p_{31\cdot}$, et $q_3 = p_{313}/p_{31\cdot}$ avec $p_{31\cdot} = \sum_{k=1}^{3} p_{31k}$ et $\sum_{k=1}^{3} q_k = 1$, il est facile de montrer que $(q_1, q_2, q_3) \sim \text{Dirichlet}(\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3)$, où $\tilde{\alpha}_1 = \alpha_{11}$, $\tilde{\alpha}_2 = \alpha_{12}$ et $\tilde{\alpha}_3 = \alpha_{13} + \sum_{k=1}^{c} \alpha_{2k}$. Par conséquent, nous avons

$$\Delta_e = \Pr(q_2 > q_1 \mid \boldsymbol{\alpha})$$
$$= \int_0^{1/2} \left\{ \int_{q_1}^{1-q_1} \frac{q_1^{\tilde{\alpha}_1-1} q_2^{\tilde{\alpha}_2-1} (1-q_1-q_2)^{\tilde{\alpha}_3-1}}{D(\tilde{\alpha}_1, \tilde{\alpha}_2\ \tilde{\alpha}_3)} dq_2 \right\} dq_1.$$

Il est alors facile de montrer que

$$\Delta_e = 1 - F_{\tilde{\alpha}_1,\tilde{\alpha}_2+\tilde{\alpha}_3}(1/2) \int_0^{1/2} F_{\tilde{\alpha}_2,\tilde{\alpha}_3}\{q_1/(1-q_1)\}$$
$$\left\{ \frac{q_1^{\tilde{\alpha}_1-1}(1-q_1)^{\tilde{\alpha}_2+\tilde{\alpha}_3-1}}{F_{\tilde{\alpha}_1,\tilde{\alpha}_2+\tilde{\alpha}_3}(1/2) B(\tilde{\alpha}_1, \tilde{\alpha}_2+\tilde{\alpha}_3)} \right\} dq_1, \quad (26)$$

où

$$F_{\tilde{\alpha}_1,\tilde{\alpha}_2+\tilde{\alpha}_3}(a) = \int_0^a \frac{x^{\tilde{\alpha}_1-1}(1-x)^{\tilde{\alpha}_2+\tilde{\alpha}_3-1}}{B(\tilde{\alpha}_1,\tilde{\alpha}_2+\tilde{\alpha}_3)} dx$$

et

$$F_{\tilde{\alpha}_2,\tilde{\alpha}_3}(a) = \int_0^a \frac{x^{\tilde{\alpha}_2-1}(1-x)^{\tilde{\alpha}_3-1}}{B(\tilde{\alpha}_2,\tilde{\alpha}_3)} dx.$$

Nous notons que Δ_e est la probabilité que Taft obtienne une plus forte proportion de votes au sondage d'octobre et que Δ_p est la probabilité que Taft reçoive une plus forte proportion de votes à l'élection de novembre. Ces paramètres peuvent être fort utiles pour l'estimation (e) et la prédiction (p). Des paramètres tels que Δ_e et Δ_p sont difficiles à analyser selon une approche non bayésienne telle que celle de Chen et Stasny (2003); il s'agit effectivement d'une grande force du paradigme bayésien.

Il est facile de calculer (26) par une intégration de Monte Carlo. Pour chaque $\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3$, Δ_e $q_1 \sim$ Bêta$(\tilde{\alpha}_1, \tilde{\alpha}_2 + \tilde{\alpha}_3)$ tronquée à l'intervalle $(0, 1/2)$ est utilisée comme une fonction d'importance. Par conséquent, pour chaque $\tilde{\boldsymbol{\alpha}}^{(h)}$, $h = 1,...,M$, $M = 1\,000$ tiré de l'échantillonneur de Metropolis-Hastings (ou de l'échantillonneur de Gibbs), nous pouvons calculer $\Delta_e^{(h)}$. Nous obtenons l'inférence *a posteriori* au sujet de Δ_e de la manière empirique classique. Pour la prédiction, nous avons considéré $\Delta_p = \Pr(p_{412} > p_{411} | p_{411} + p_{412} + p_{413}, \boldsymbol{\alpha})$, où $\alpha_{jk} = \mu_{jk}\tau_2$, $j = 1,..., r$, $k = 1, ..., c$, pour les modèles de non-réponse ignorable et non ignorable, et $\alpha_{jk} = p_{3jk}\tau_2$, $j = 1, ..., r$, $k = 1, ..., c$, pour le modèle de non-réponse variable en fonction du temps. Notons que Δ_e et Δ_p sont les mêmes pour les modèles de non-réponse ignorable et non ignorable.

Au tableau 3, nous présentons aussi les estimations de Δ_e et Λ_p pour les trois modèles. Premièrement, soulignons de nouveau que toutes les ETN sont faibles. Les estimations de ces paramètres sont semblables pour les trois modèles et supérieures à 0,60, mais les intervalles de crédibilité à 95 % contiennent la valeur 0,5. Par conséquent, de nouveau, les moyennes *a posteriori* indiquent que Taft est le gagnant, mais la variation nullifie l'effet de Taft étant le gagnant. Nous remarquons une fois de plus que le modèle de non-réponse variable en fonction du temps fournit une inférence plus précise, quoique pas suffisamment. Les paramètres Δ_e et Δ_p sont plus sensibles, parce qu'ils limitent l'inférence à une région plus petite en conditionnant sur $p_{311} + p_{312} + p_{313}$ et $p_{411} + p_{412} + p_{413}$, et, d'un point de vue probabiliste, ils sont plus appropriés.

Enfin, nous étudions la sensibilité à l'inférence au sujet de Λ_p et Δ_p pour le modèle de non-réponse non ignorable et le modèle de non-réponse variable en fonction du temps. Nous ne présentons pas les résultats pour Λ_e et Δ_e, parce qu'ils sont semblables à ceux pour Λ_p et Δ_p. En outre, nous avons également abandonné le modèle de non-réponse ignorable, et nous ne présentons pas les intervalles de crédibilité à 95 %, parce que les densités *a posteriori* sont à peu près symétriques. Nous présentons nos résultats au tableau 4 par modèle en fonction de κ_1 et de κ_2. Les moyennes *a posteriori* de Λ_p et Δ_p sont, respectivement, fort semblables pour diverses valeurs de κ_1 et de κ_2. Notons qu'*a priori*

$$\tau_1 \sim \text{Gamma}\left(\frac{1}{\kappa_1^2\, c_0^2}, \frac{1}{\kappa_2\, \mu_0\, \kappa_1^2\, c_0^2}\right),$$

$E(\tau_1) = \kappa_2\, \mu_0$ et E.-T.$(\tau_1) = \kappa_1\, \kappa_2\, c_0\, \mu_0$; par conséquent, manifestement, $E(\tau_1)$ *a priori* augmente avec κ_2 et E.-T.(τ_1) augmente avec κ_1 ou κ_2, mais pas nécessairement *a posteriori*. Ces changements n'ont pas beaucoup d'effet sur l'inférence *a posteriori*. Pour presque toutes les combinaisons de κ_1 et κ_2, sous le modèle de non-réponse variable en fonction du temps, les écarts-types *a posteriori* de Λ_p sont plus faibles (mais pas considérablement) que sous le modèle de non-réponse non ignorable. Sous le modèle de non-réponse variable en fonction du temps, les écarts-types *a posteriori* de Δ_p sont considérablement plus petits que sous le modèle de non-réponse non ignorable pour toutes les combinaisons de κ_1 et κ_2.

Conclusion

La principale contribution du présent article comprend la construction et l'analyse d'un modèle de non-réponse non ignorable variable en fonction du temps et son application aux données des sondages électoraux en Ohio. À cela s'ajoutent deux autres accomplissements. Premièrement, nous avons comparé le modèle variable en fonction du temps à une version étendue (afin d'y inclure la variable de temps) des modèles de non-réponse ignorable et non ignorable de Nandram et coll. (2005). Deuxièmement, nous avons construit un nouveau paramètre pour faciliter la prédiction du gagnant; cependant, ce paramètre n'a pas changé énormément les choses, en partie parce que le modèle de non-réponse variable en fonction du temps ne comporte que trois points dans le temps.

Tableau 4
Sensibilité des moyennes *a posteriori* (MP) et des écarts-types *a posteriori* (ETP) de Λ_p et Δ_p en fonction des variations de κ_1 et κ_2 par modèle

		κ_2							
		1		5		25		50	
Modèle	κ_1	MP	ETP	MP	ETP	MP	ETP	MP	ETP
a. Λ_p									
Non ignorable	1	0,058	0,150	0,046	0,153	0,060	0,148	0,054	0,147
	2	0,051	0,153	0,046	0,146	0,062	0,151	0,054	0,145
	3	0,058	0,152	0,059	0,145	0,053	0,149	0,055	0,149
	4	0,055	0,151	0,057	0,148	0,049	0,148	0,043	0,154
Variable en fonction du temps	1	0,050	0,134	0,044	0,144	0,048	0,136	0,050	0,130
	2	0,049	0,136	0,052	0,140	0,056	0,129	0,047	0,137
	3	0,039	0,139	0,049	0,137	0,045	0,139	0,052	0,133
	4	0,037	0,138	0,042	0,138	0,041	0,141	0,051	0,129
b. Δ_p									
Non ignorable	1	0,663	0,200	0,650	0,194	0,666	0,186	0,670	0,182
	2	0,663	0,197	0,661	0,188	0,667	0,185	0,659	0,181
	3	0,663	0,199	0,647	0,196	0,666	0,184	0,669	0,180
	4	0,641	0,202	0,668	0,191	0,643	0,197	0,650	0,195
Variable en fonction du temps	1	0,648	0,155	0,642	0,123	0,657	0,099	0,661	0,095
	2	0,660	0,151	0,652	0,127	0,659	0,102	0,657	0,099
	3	0,622	0,153	0,636	0,137	0,649	0,120	0,648	0,115
	4	0,610	0,162	0,636	0,152	0,646	0,132	0,644	0,127

NOTA : Nous avons pris $\tau_1 \sim \text{Gamma}\,(1/\kappa_1^2 c_0^2,\ 1/\kappa_2 \mu_0 \kappa_1^2 c_0^2)$ et nous avons étudié la sensibilité en fonction de κ_1 et de κ_2. Voir la note du tableau 3.

Notre modèle variable en fonction du temps fournit des inférences *a posteriori* plus proches des valeurs réelles que les modèles de non-réponse ignorable et non ignorable, et que ceux de Chen et Stasny (2003). Il est naturel que les préférences des électeurs évoluent à mesure que de nouveaux renseignements, nuisibles ou favorables, sont révélés au public. Donc, notre modèle de non-réponse variable en fonction du temps, qui tient compte des changements au cours du temps et donne une meilleure précision, est celui auquel il faut donner la préférence. L'incertitude de la prédiction peut être réduite de deux manières. Premièrement, l'accroissement du nombre de sondages augmentera la précision des estimations des paramètres, ce qui à son tour pourra améliorer la prédiction. Deuxièmement, une plus grande quantité d'information *a priori* (par exemple, sondage à la sortie de l'isoloir) au sujet de l'élection de novembre permettra également d'améliorer la prédiction.

Nos intervalles de crédibilité à 95 % peuvent être raccourcis en utilisant des renseignements *a priori* sur la proportion d'électeurs votant pour Taft ou pour Fisher. Selon un examinateur, « dans les élections générales, les proportions de votes pour les grands partis sont comprises entre 35 % et 65 % et pour des états particuliers, un politicologue objectif pourrait généralement fournir un prior encore plus serré ». Cependant, le problème est complexe, parce que, quand les distributions *a priori* sont tronquées sur les $\boldsymbol{p}$, il existe une constante de normalisation qui est une fonction de τ_2. Donc, quand τ_2 est tiré de la densité *a posteriori* conditionnelle, nous devons procéder à une intégration de Monte Carlo pour calculer la constante de normalisation à chaque itération. Bien qu'il s'agisse d'une contribution utile, nous préférons ne pas poursuivre ce problème ici.

Le nombre de jours avant un élection a un effet important sur l'exactitude des sondages et cet effet peut varier considérablement selon le contexte de la campagne électorale (voir, par exemple, DeSart et Holbrook 2003). Il est donc fort difficile de prédire le résultat d'une élection plusieurs semaines avant qu'elle ait effectivement lieu, à moins qu'il existe une marge absolue. Une personne qui souhaite prédire le résultat d'une élection doit tenir compte de l'information supplémentaire aux approches de l'élection réelle. Notre prédiction est fondée sur l'hypothèse qu'aucun changement catastrophique ne surviendra à l'approche de l'élection; mais ce genre de changement soudain dans l'opinion publique peut avoir lieu. Par exemple, en 1988, Dukakis a perdu l'élection contre George Bush pour diverses raisons, dont le fait qu'il a passé la dernière semaine de la campagne au Massachusetts, que sa personnalité était jugée froide et que Bush a attaqué sa position libérale. En outre, une campagne efficace peut mobiliser les électeurs indécis à l'approche de l'élection (par exemple, Truman et Dewey en 1948). Un moyen de tenir compte d'une catastrophe éventuelle consiste à utiliser un mélange de distribution ou d'autres distribution à queues lourdes (à l'instar des chercheurs qui utilisent des distributions de Levy en mathématiques financières).

Remerciements

Les présents travaux ont été réalisés pendant que Balgobin Nandram était en congé sabbatique au National Center for Health Statistics, à Hyattsville, au Maryland, en 2003-2004.

Annexe A

Modèle variable en fonction du temps : densités a posteriori conditionnelles de $p_t, t = 0,\ldots,T$

Nous montrons comment tirer un échantillon de la densité *a posteriori* conditionnelle de $\boldsymbol{p}_0$ dans (17) en utilisant une méthode par grille, et comment tirer un échantillon des densités *a posteriori* conditionnelles des $\boldsymbol{p}_t, t = 1, ..., T$ dans (18), en utilisant des pas de Metropolis, chacun avec une chaîne indépendante.

En premier lieu, nous montrons comment tirer un échantillon d'une densité *a posteriori* conditionnelle de $\boldsymbol{p}_0$ dans (17) par une méthode de grille. En posant que $(q_{01}, ..., q_{0L}) = (p_{01}, ..., p_{0rc})$ et $(q_{11}, ..., q_{1L}) = (p_{11}, ..., p_{1rc})$ où $L = rc$, avec $\sum_{\ell=1}^{L-1} q_{0\ell} \le 1$, nous avons

$$\pi(q_{01},\ldots,q_{0L-1} \mid \text{sinon}, \boldsymbol{y}_1)$$

$$\propto \frac{q_{1L}^{(1-\sum_{\ell=1}^{L-1} q_{0\ell})\tau_2 - 1}}{\Gamma((1-\sum_{\ell=1}^{L-1} q_{0\ell})\,\tau_2)} \prod_{\ell=1}^{L-1} \frac{q_{1\ell}^{q_{0\ell}\tau_2 - 1}}{\Gamma(q_{0\ell}\tau_2)}, 0 \le q_{0\ell} \le 1,\ \ell = 1,\ldots,L-1,$$

et il est facile de montrer que

$$\pi(q_{0\ell} \mid \text{sinon}, \boldsymbol{y}_1) \propto \frac{q_{1L}^{(1-\sum_{\ell=1}^{L-1} q_{0\ell})\tau_2 - 1}}{\Gamma((1-\sum_{\ell=1}^{L-1} q_{0\ell})\tau_2)} \frac{q_{1\ell}^{q_{0\ell}\tau_2 - 1}}{\Gamma(q_{0\ell}\,\tau_2)},$$

$$0 \le q_{0\ell} \le 1 - \sum_{\ell'=1,\ \ell' \ne \ell}^{L-1} q_{0\ell'},\ \ell = 1,\ldots,L-1.$$

Pour chaque ℓ, nous divisons l'intervalle $0 \le q_{0\ell} \le 1 - \sum_{\ell'=1,\ \ell' \ne \ell}^{L-1} q_{0\ell'}$ en plusieurs sous-intervalles. Pour tirer une valeur aléatoire $q_{0\ell}$ de sa densité *a posteriori* conditionnelle, nous sélectionnons un intervalle proportionnel à son aire et tirons une valeur aléatoire uniforme de cet intervalle.

En deuxième lieu, nous montrons comment tirer un échantillon des densités *a posteriori* conditionnelles des $\boldsymbol{p}_t, t = 1, ..., T$ dans (18), en utilisant des pas de Metropolis, chacun avec une chaîne indépendante. Considérons $\boldsymbol{p}_t \mid \boldsymbol{p}_{t-1}, \tau_2, \boldsymbol{y}, t = 1,\ldots,T$. Nous utilisons la densité générant les valeurs candidates

$$\boldsymbol{p}_t \mid \boldsymbol{p}_{t-1}, \tau_2, \boldsymbol{y} \sim \text{Dirichlet}(\boldsymbol{a}_t),$$

où

$$a_{tjk} = y_{t\cdot jk} + \tau_2 p_{t-1jk},\ t = 1,\ldots,T,\ j = 1,\ldots,r,\ k = 1,\ldots,c.$$

Alors, la probabilité d'acceptation est $A_{s,s+1} = \min(1, \psi_{s+1}/\psi_s)$, où

$$\psi_s = \prod_{j=1}^{r} \prod_{k=1}^{c} p_{t+1jk}^{(p_{ijk}^{(s)}\tau_2 - 1)} / D(\boldsymbol{p}_t^{(s)}\tau_2).$$

Bibliographie

Chang, L.C., et Krosnick, J.A. (2001). Improving election forecasting. *Technical Report*, Department of Psychology, The Ohio State University.

Chen, Q.L., et Stasny, E.A. (2003). Handling undecided voters: Using missing data methods in election forecasting. *Technical Report*, Department of Statistics, The Ohio State University.

Chen, T., et Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.

DeSart, J., et Holbrook, T.M. (2003). Campaigns, polls, and the states: Assessing the accuracy of statewide presidential trial-heat polls. *Political Research Quarterly*, 56, 431-439.

Draper, D. (1995). Assessment and propagation of model uncertainty (avec discussion). *Journal of the Royal Statistical Society*, Séries B, 57, 45-97.

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Gelman, A., Roberts, G.O. et Gilks, W.R. (1996), Efficient Metropolis jumping rules. Dans *Bayesian Statistics*, (Éds., J.M. Bernardo, J.O. Berger, A.P. Dawid et A.F.M. Smith), Oxford, U.K. : Oxford University Press, 599-607.

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

Kalton, G., et Kasprzyk, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

Lavrakas, P.J. (1993). *Telephone Survey Methods*: *Sampling*, *Selection*, *and Supervision*. Newbury Park, CA : Sage Publications.

Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. 2ème édition, New York : John Wiley & Sons Inc.

Nandram, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*, 61, 97-126.

Nandram, B., et Choi, J.W. (2002a). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.

Nandram, B., et Choi, J.W. (2002b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.

Nandram, B., Cox, L.H. et Choi, J.W. (2005). Analyse bayésienne des données catégoriques manquantes non ignorables : une application à la densité minérale osseuse et au revenu familial. *Techniques d'enquête*, 31, 233-247.

Ritter, C., et Tanner, M.A. (1992). The Gibbs stopper and the Griddy Gibbs sampler. *Journal of the American Statistical Association*, 87, 861-868.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Techniques d'enquête, juin 2008
Vol. 34, N° 1, pp. 55-70
Statistique Canada, N° 12-001-X au catalogue

Une approche bayésienne non informative de l'échantillonnage d'une population finie en utilisant des variables auxiliaires

Radu Lazar, Glen Meeden et David Nelson [1]

Résumé

En échantillonnage de populations finies, on dispose souvent d'information *a priori* sous la forme de renseignements partiels sur une variable auxiliaire, dont on connaît, par exemple, la moyenne. Dans de tels cas, on utilise fréquemment l'estimateur par le ratio et l'estimateur par la régression pour estimer la moyenne de population de la caractéristique d'intérêt. La loi *a posteriori* de Pólya a été établie à titre d'approche bayésienne non informative de l'échantillonnage. Son application est indiquée quand on ne dispose que de peu d'information *a priori*, voire aucune, au sujet de la population. Ici, nous montrons que cette loi peut être étendue afin d'y intégrer divers types d'information *a priori* partielle au sujet des variables auxiliaires. Nous verrons que cette approche produit généralement des méthodes ayant de bonnes propriétés fréquentistes, même dans certains cas où les méthodes fréquentistes classiques sont difficiles à appliquer.

Mots clés : Échantillonnage; approche bayésienne non informative; variable auxiliaire; contraintes linéaires et loi *a posteriori* de Pólya.

1. Introduction

L'échantillonnage de populations finies est l'un des domaines de la statistique où l'information *a priori* est utilisée systématiquement pour faire des inférences. Dans la plupart des cas, cette information *a priori* n'est pas intégrée dans le problème de manière bayésienne, entre autres raisons, parce qu'elle ne mène habituellement pas, d'une manière évidente, à une loi *a priori* raisonnable. Dans l'approche prédictive (voir Valliant, Dorfman et Royall, 2000), on part d'un modèle hypothétique et l'on estime ses paramètres inconnus après avoir observé l'échantillon. Dans la théorie fréquentiste classique, l'information *a priori* est encapsulée dans le plan d'échantillonnage probabiliste ou plan de sondage. Basu a montré que, pour un tenant de la théorie bayésienne, une fois que l'échantillon a été observé, le plan d'échantillonnage ne joue plus aucun rôle dans la loi *a posteriori*. [Pour ce fait et d'autres réflexions de Basu sur l'échantillonnage de populations finies, voir Ghosh (1988).] Bien que cette conclusion radicale ne soit pas acceptée unanimement par les bayésiens, il est vrai que la théorie fréquentiste classique et l'approche bayésienne de l'échantillonnage ont peu de chose en commun.

La théorie classique de l'échantillonnage met l'accent sur l'estimation de la moyenne de population. Si celle-ci est connue *a priori* pour une variable auxiliaire, on utilise souvent l'estimateur par le ratio ou l'estimateur par la régression. Si l'on souhaite estimer un autre paramètre que la moyenne, disons un quantile de population ou la fonction de répartition de la population, ou que l'on possède de l'information *a priori* sur une autre variable auxiliaire que la moyenne, il faut élaborer de nouvelles méthodes. Des travaux récents allant dans ce sens sont décrits dans Chen et Qin (1993), Chen et Sitter (1999), Mak et Kuk (1993), Kuk et Mak (1989), Rao, Kovar et Mantel (1990), ainsi que Chambers et Dunstan (1986).

L'un des avantages de l'approche bayésienne est qu'une loi *a posteriori* raisonnable pour la population permet d'intégrer l'information *a priori* dans l'estimation de plusieurs paramètres de population. Même si la loi *a posteriori* n'a pas d'expression explicite pour un estimateur particulier, pour tout échantillon donné, nous pouvons trouver sa valeur approximative. Pour cela, nous échantillonnons la loi *a posteriori* pour simuler des répliques complètes de la population et nous employons des méthodes d'estimation de Monte Carlo. Si la loi *a posteriori* n'a pas de forme commode pour l'échantillonnage, il devrait être possible d'utiliser des méthodes Monte Carlo par chaînes de Markov pour mettre en œuvre le processus de simulation. Pour chacune des répliques simulées, nous calculons la valeur du paramètre d'intérêt. En simulant un grand nombre de répliques complètes de la population, nous pouvons trouver, approximativement, les estimations bayésiennes ponctuelles et par intervalle correspondantes du paramètre de population d'intérêt. Le problème se résume alors à trouver un modèle bayésien de population raisonnable qui utilise le genre d'information *a priori* disponible pour la variable auxiliaire.

Souvent, on peut fonder des modèles bayésiens raisonnables sur la loi *a posteriori* de Pólya. Cette dernière représente une méthode bayésienne non informative d'échantillonnage d'une population finie qui n'utilise que peu d'information *a priori*, voire aucune, au sujet de la

1. Radu Lazar, School of Statistics, University of Minnesota, Minneapolis, MN 55455. Courriel : radu@stat.umn.edu; Glen Meeden, School of Statistics, University of Minnesota, Minneapolis, MN 55455. Courriel : glen@stat.umn.edu; David Nelson, Center for Chronic Disease Outcomes Research, Minneapolis VA Medical Center, Minneapolis, MN 55417. Courriel : Dave.Nelson@med.va.gov.

population. Une bonne discussion de cette approche se trouve dans Ghosh et Meeden (1997). Son usage est approprié quand un spécialiste de l'échantillonnage classique est prêt à adopter un plan d'échantillonnage aléatoire simple. Ici, nous montrons comment l'approche peut être étendue aux cas où l'information *a priori* au sujet de la variable indépendante existe. Par exemple, nous pourrions connaître exactement la moyenne ou la médiane d'une variable auxiliaire, ou savoir qu'elle est comprise dans un intervalle donné de valeurs possibles.

L'idée fondamentale est de contraindre ou de restreindre la loi *a posteriori* de Pólya de telle façon que seules des probabilités positives soient appliquées aux populations simulées qui satisfont les contraintes découlant de l'information *a priori* sur les variables auxiliaires. Cette loi *a posteriori* de Pólya judicieusement contrainte peut alors être utilisée pour faire des inférences au sujet des paramètres de population d'intérêt. Nelson et Meeden (1998) ont envisagé plusieurs scénarios où une loi *a posteriori* de Pólya contrainte a produit des résultats fréquentistes raisonnables. Ils ont supposé dans cette étude que l'information sur les quantiles de population de la variable auxiliaire était connue *a priori*. Ici, nous supposons que nous disposons de formes plus générales d'information *a priori* au sujet de la moyenne ou des quantiles de population pour un ensemble de variables auxiliaires. Nous pouvons connaître exactement ces quantités ou savoir uniquement qu'elles appartiennent à un intervalle donné.

À la section 2, nous examinons la loi *a posteriori* de Pólya. À la section 3, nous présentons la loi *a posteriori* de Pólya contrainte et discutons du moyen d'utiliser des méthodes de Monte Carlo par chaînes de Markov pour trouver, approximativement, la valeur d'une estimation fondée sur cette loi. À la section 4, nous appliquons la loi *a posteriori* de Pólya à diverses situations et la comparons à des méthodes standard. À la section 5, nous discutons plus formellement de la façon dont cette loi se rapporte à certaines méthodes fréquentistes classiques. À la section 6, nous voyons que la loi *a posteriori* de Pólya contrainte peut être utilisée avec d'autres plans d'échantillonnage que l'échantillonnage aléatoire simple. À la section 7, nous présentons quelques conclusions. En annexe, nous prouvons un résultat d'admissibilité pour la loi *a posteriori* de Pólya contrainte qui donne une justification théorique des méthodes présentées ici.

2. La loi *a posteriori* de Pólya

Considérons une population finie constituée de N unités. Pour l'unité i, soit y_i, un nombre réel qui est la valeur inconnue d'une caractéristique d'intérêt. Nous supposons que l'état de la nature inconnu, $y = (y_1, ..., y_N)$, appartient à un sous-ensemble connu de l'espace euclidien à N dimensions. Supposons que nous voulions estimer une fonction $\gamma(y)$ de l'état inconnu. Dans une analyse bayésienne appropriée, l'étape suivante consisterait à spécifier une loi *a priori* sur l'espace des paramètres. Puis, étant donné un échantillon obtenu conformément au plan d'échantillonnage, on déterminerait la loi *a posteriori* des membres inobservés de la population, sachant les valeurs des unités observées dans l'échantillon. Dans la plupart des cas, la loi *a posteriori* ne dépendra pas du plan d'échantillonnage.

Nous pouvons utiliser la loi *a posteriori* de Pólya comme une loi *a posteriori* adéquate, même si elle n'émane pas d'un modèle bayésien approprié. Elle conviendrait dans le cas où l'on en sait fort peu au sujet de la population et que l'on suppose que l'échantillon est représentatif de la population. Le cas d'un plan d'échantillonnage aléatoire simple est un exemple de situation où son utilisation serait appropriée. Nous allons maintenant décrire brièvement cette loi et donner dans les grandes lignes sa justification théorique.

Compte tenu des données, la loi *a posteriori* de Pólya est une loi prédictive conjointe pour les unités inobservées de la population, sachant les valeurs dans l'échantillon. Étant donné un échantillon, nous montrons maintenant comment produire à partir de cette loi des valeurs positives pour les unités inobservées. Considérons deux urnes dont la première contient les n unités de l'échantillon ainsi que leurs valeurs y observées. La deuxième contient les $N - n$ unités non échantillonnées. Nous commençons par choisir une unité au hasard dans chacune des deux urnes. Puis, nous attribuons la valeur y observée de l'unité sélectionnée dans la première urne à l'unité sélectionnée dans la seconde, et nous les plaçons toutes les deux dans la première urne. Les urnes contiennent maintenant $n + 1$ et $N - n - 1$ boules, respectivement. Nous répétons ce processus jusqu'à ce que toutes les unités de la deuxième urne soient transférées dans la première et qu'une valeur leur soit attribuée. À chaque étape du processus, toutes les unités de la première urne ont la même probabilité d'être sélectionnées. Autrement dit, les unités auxquelles une valeur a été attribuée sont traitées exactement comme celles qui figuraient réellement dans l'échantillon. Une fois cet exercice terminé, nous avons produit une réalisation complète de la population sous la loi *a posteriori* de Pólya. Cette réplique complète, simulée, contient les n valeurs observées, ainsi que les $N - n$ valeurs simulées pour les membres inobservés de la population. Donc, un simple échantillonnage de Pólya donne une loi prédictive pour les unités inobservées, sachant les unités observées. Une bonne description de l'échantillonnage de Pólya figure dans Feller (1968). La loi *a posteriori* de Pólya est apparentée au bootstrap bayésien de Rubin (1981). Voir également Lo (1988) et Binder (1982).

Cette loi prédictive produit souvent des estimateurs semblables aux estimateurs fréquentistes classiques sous échantillonnage aléatoire simple. Considérons, par exemple, l'estimation de la moyenne de population. Avant de poursuivre, nous devons nous étendre un peu sur la notation.

Soit s un échantillon possible de taille $n(s)$. Il s'agit d'un sous-ensemble de $\{1, 2, ..., N\}$, l'ensemble d'étiquettes pour la population finie. Si $s = \{i_1, ..., i_{n(s)}\}$, alors $y_s = \{y_{i_1}, ..., y_{i_{n(s)}}\}$ est l'ensemble de valeurs observées pour y, la caractéristique d'intérêt. Soit $z = (s, y_s)$ un échantillon observé typique. Alors, sachant $z = (s, y_s)$, nous avons

$$\bar{z}_s = \sum_{j=1}^{n(s)} y_{i_j}/n(s)$$

et

$$\text{Var}(z) = \sum_{j=1}^{n(s)} (y_{i_j} - \bar{z}_s)^2/(n(s) - 1)$$

qui sont la moyenne et la variance d'échantillon. Soit $\gamma_{mn}(y) = \sum_{i=1}^{N} y_i/N$ la moyenne de population. Sous la loi *a posteriori* de Pólya,

$$E(\gamma_{mn}(y) \mid z) = \bar{z}_s$$

et

$$\text{Var}(\gamma_{mn}(y) \mid z) = (1 - f)\frac{\text{Var}(z)}{n(s)}\frac{n(s) - 1}{n(s) + 1}$$

où $f = n(s)/N$. Notons que, sauf pour le dernier facteur de la variance *a posteriori*, ces deux termes sont simplement la moyenne d'échantillon et sa variance sous échantillonnage aléatoire simple. Les probabilités données par le plan de sondage ne jouent aucun rôle explicite dans ces calculs. Néanmoins, pour que la loi *a posteriori* de Pólya soit jugée appropriée par le spécialiste de l'échantillonnage, il faut que les valeurs de la caractéristique d'intérêt pour les unités observées et non observées soient approximativement interchangeables. C'est dans ce genre de situation que l'échantillonnage aléatoire simple sans remise est utilisé.

Sous la loi *a posteriori* de Pólya, il n'est pas toujours possible de trouver explicitement l'intervalle bayésien de crédibilité pour la moyenne de population ou pour les estimations ponctuelles et par intervalle d'autres paramètres de population. Dans de tels cas, il est facile de trouver ces estimations approximativement en simulant répétitivement des répliques complètes de la population. Pour chaque réplique simulée, nous calculons le paramètre de population d'intérêt. L'expérience a montré que de 500 à 1 000 valeurs simulées donneront habituellement de bons résultats. La moyenne de ces valeurs calculées sera notre estimation ponctuelle et les quantiles 0,025 et 0,975 de ces valeurs calculées donneront notre estimation par intervalle.

Puisque, sous la loi *a posteriori* de Pólya, les seules valeurs de y qui figurent dans une réplique complète simulée de la population sont celles qui sont présentes dans l'échantillon, cette loi est simplement un moyen d'attribuer des poids aléatoires, c'est-à-dire des probabilités, aux unités comprises dans l'échantillon. Sous la loi *a posteriori* de Pólya, le poids moyen attribué à chaque unité échantillonnée est $1/n(s)$, de sorte que, comme nous l'avons vu, l'estimation de la moyenne de population soit simplement la moyenne d'échantillon. Ce sont cette relation et le caractère bayésien de la loi *a posteriori* de Pólya qui nous permettent de prouver l'admissibilité de la moyenne d'échantillon pour estimer la moyenne de population sous perte d'erreur quadratique. Cela donne à penser que les méthodes d'inférence fondées sur la loi *a posteriori* de Pólya auront tendance à concorder avec les méthodes fréquentistes et auront de bonnes propriétés fréquentistes.

À titre d'illustration supplémentaire de ce point, nous notons que récemment, deux des auteurs (Nelson et Meeden 2006) ont démontré que les intervalles de crédibilité bayésiens fondés sur la loi *a posteriori* de Pólya pour la médiane de population concordent asymptotiquement avec l'intervalle standard de Woodruff (Woodruff 1952). En guise d'autre exemple, considérons l'estimation de la moyenne ou du total d'une sous-population ou d'un domaine quand on utilise un échantillon aléatoire simple tiré de la population complète. Ici, le nombre d'unités échantillonnées qui appartiennent au domaine est une variable aléatoire. Donc, la moyenne des unités échantillonnées qui tombent dans le domaine est le ratio de deux variables aléatoires. Cette estimation est plus compliquée que celle de la moyenne de l'ensemble des unités dans l'échantillon. Pour obtenir une estimation de la variance pour cet estimateur, la méthode fréquentiste habituelle impose comme contrainte le nombre d'unités échantillonnées qui sont comprises dans le domaine. Cependant, quand on estime le total des domaines, cet argument conditionnel ne marche pas et il faut utiliser une méthode inconditionnelle pour obtenir une estimation de la variance. À cet égard, voir, par exemple, Cochran (1976). Récemment, l'un des auteurs (Meeden 2005) a montré que les inférences basées sur la loi *a posteriori* de Pólya concordent avec les réponses fréquentistes habituelles. Par conséquent, cette loi permet de traiter les deux situations en appliquant une seule théorie. Il est important de se rappeler que, dans l'approche fréquentiste, le conditionnement peut se faire sous l'échantillonnage aléatoire simple, mais que, pour des plans d'échantillonnage plus complexes, il n'est généralement pas faisable, puisque la loi aléatoire conditionnelle est inconnue. Enfin, notons que les méthodes d'échantillonnage par grappe à deux degrés fréquentistes habituelles peuvent être justifiées pour

une extension de la loi *a posteriori* de Pólya (Meeden 1999).

La loi *a posteriori* de Pólya est semblable en esprit aux méthodes du bootstrap pour l'échantillonnage d'une population finie. Dans l'un et l'autre cas, un genre d'argument d'échangeabilité est utilisé pour produire des pseudo-versions de la population. L'idée fondamentale du bootstrap est décrite dans Gross (1980). Supposons que nous ayons un échantillon aléatoire simple de taille $n(s)$ tiré de la population et supposons que $N/n(s) = m$ est un nombre entier. Partant de l'échantillon, nous créons une bonne approximation de la population en combinant m répliques de l'échantillon. Puis, nous tirons des échantillons aléatoires répétés de taille $n(s)$ de cette population créée pour étudier le comportement de l'estimateur d'intérêt. Les propriétés asymptotiques des estimateurs peuvent également être étudiées (voir Booth, Bulter et Hall 1994 pour des précisions). Cette situation diffère de l'utilisation de la loi *a posteriori* de Pólya qui, pour un échantillon fixe, produit des versions complètes de la population et permet d'examiner la distribution du paramètre d'intérêt dans la population, plutôt que les propriétés de l'estimateur du paramètre. Pour le paramètre de population d'intérêt, les propriétés de l'estimateur découlent directement de cette loi prédictive pour les valeurs de population.

La loi *a posteriori* de Pólya correspond au bootstrap bayésien de Rubin (1981) appliqué à l'échantillonnage d'une population finie. Le bootstrap bayésien original s'applique à un échantillon aléatoire tiré d'une population infinie. Rubin a montré que le bootstrap et le bootstrap bayésien sont opérationnellement fort semblables. Le même type d'analogie tient pour le cadre de population finie. Pour étudier la variabilité d'un estimateur, selon l'une et l'autre approche des poids sont appliqués aléatoirement aux unités sélectionnées dans l'échantillon. Toutefois, la logique suivie pour attribuer les poids diffère, ainsi que les justifications théoriques. Le bootstrap possède une justification asymptotique sous échantillonnage aléatoire répété, tandis que la loi *a posteriori* de Pólya possède une justification de décision théorique fondée sur sa nature bayésienne séquentielle (Ghosh et Meeden 1997).

Au lieu de produire une réplique complète de la population, il est souvent plus efficace d'utiliser une approximation bien connue de la loi *a posteriori* de Pólya. Supposons que la fraction d'échantillonnage f soit faible. Pour $j = 1, ..., n(s)$, soit p_j la proportion d'unités dans une réplique complète simulée de la population entière qui prend la valeur y_j. Alors, sous la loi *a posteriori* de Pólya, $p = (p_1, ..., p_{n(s)})$ suit approximativement une loi de Dirichlet avec un vecteur de paramètres qui ne contient que des valeurs 1, c'est-à-dire qui est uniforme sur le simplexe à $n(s) - 1$ dimensions, où $\sum_{j=1}^{n(s)} p_j = 1$. Cette approche sera très utile quand nous examinerons la loi *a posteriori* de Pólya contrainte.

3. La loi *a posteriori* de Pólya contrainte

3.1 L'idée fondamentale

Dans de nombreuses situations, en plus de la variable d'intérêt, y, l'échantillonneur dispose de variables auxiliaires, x, pour lesquelles existe de l'information *a priori*. Par exemple, la moyenne de population, μ_x, de x pourrait être connue. Étant donné une unité dans un échantillon aléatoire, nous observons sa paire de valeurs (y, x). En suivant notre notation antérieure, nous représentons l'échantillon par

$$z = (s, (y, x)_s) = (s, \{(y_{i_1}, x_{i_1}), ..., (y_{i_{n(s)}}, x_{i_{n(s)}})\}).$$

Dans cette situation, l'estimateur par la régression est souvent utilisé pour estimer la moyenne de population. Comment la loi *a posteriori* de Pólya devrait-elle être ajustée pour tenir compte du fait que la moyenne de population de x est connue? La réponse simple consiste à restreindre la loi prédictive uniquement aux populations compatibles avec l'information *a priori*. En pratique, cela reviendrait à ne produire que des répliques complètes de la population compatible avec l'information *a priori* connue. Pour l'illustrer, nous considérons la forme approximative de la loi *a posteriori* de Pólya décrite à la fin de la section précédente.

Pour $j = 1, ..., n(s)$, soit p_j la proportion d'unités dans une réplique complète de la population qui ont la valeur (y_{i_j}, x_{i_j}). Au lieu d'utiliser la loi uniforme pour $p = (p_1, ..., p_{n(s)})$ sur le simplexe pour générer les répliques simulées de la population, nous devrions utiliser la loi uniforme limitée au sous-ensemble du simplexe satisfaisant

$$\sum_{j=1}^{n(s)} p_j x_{i_j} = \mu_x. \qquad (1)$$

Avant de décrire comment nous générons les vecteurs de p à partir de la loi *a posteriori* de Pólya contrainte, nous examinons le lien entre l'estimateur résultant et l'estimateur par la régression.

De nombreux résultats de simulation (non présentés ici) montrent que le comportement de la loi *a posteriori* de Pólya contrainte ressemble fort à celui de l'estimateur par la régression sous échantillonnage aléatoire simple. Le simple argument qui suit montre pourquoi ces deux estimations ponctuelles devraient concorder fréquemment, même si la loi *a posteriori* de Pólya ne comporte aucune hypothèse au sujet de la relation entre y et x.

Supposons une population $y_i = a + bx_i + \varepsilon_i$, où ε_i est une erreur aléatoire d'espérance nulle. Soit $\bar{X}$ la moyenne

de population connue de x. Alors, étant donné un échantillon et les p_i satisfaisant $\sum_{i\in s} p_i x_i = \bar{X}$, nous avons

$$E\left(\sum_{i\in s} p_i y_i\right) = aE\left(\sum_{i\in s} p_i\right) + bE\left(\sum_{i\in s} p_i x_i\right) + \sum_{i\in s} E(p_i e_i)$$
$$\doteq a + b\bar{X}$$
$$\doteq \bar{y}_s - \hat{b}\bar{x}_s + \hat{b}\bar{X}$$
$$= \bar{y}_s + \hat{b}(\bar{X} - \bar{x}_s)$$

où $\hat{b}$ est l'estimation par les moindres carrés de b. Ici, les valeurs d'échantillon sont fixes, les p_i et les e_i sont aléatoires et l'espérance de p_i est calculée par rapport à la loi *a posteriori* de Pólya. La première approximation s'ensuit, puisque sous échantillonnage aléatoire simple, nous nous attendons à ce que les échantillons soient équilibrés, en moyenne, et que les p_i et les e_i soient approximativement indépendantes.

3.2 Constraintes linéaires et loi *a posteriori* de Pólya

L'information *a priori* concernant les variables auxiliaires peut avoir de multiples origines. Nous avons déjà discuté du cas où la moyenne de population d'une variable auxiliaire est connue. Un autre est celui où l'on connaît la médiane de population. De manière plus générale, on pourrait savoir uniquement qu'une moyenne ou une médiane de population est comprise dans un certain intervalle de nombres réels. Même si ce genre de cas est peu fréquemment discuté dans la littérature habituelle traitant de l'estimation fondée sur le plan de sondage, il semble assez raisonnable. Un autre cas est celui où une paire de variables auxiliaires décrit un tableau à double entrée dans lequel chaque unité doit appartenir à l'une des cellules et dont les totaux de population de ligne et de colonne pour les chiffres figurant dans chaque cellule sont connus. Avant de décrire l'approche de la loi *a posteriori* de Pólya contrainte pour résoudre ce genre de problème, nous devons mentionner un point technique mineur.

Supposons que nous sachions que la moyenne de la variable auxiliaire x est égale à $\mu(x)$. Il existera des échantillons pour lesquels la valeur de x est inférieure à $\mu(x)$ pour chaque unité échantillonnée. Dans de tels cas, il sera impossible d'utiliser la loi *a posteriori* de Pólya contrainte. Toutefois, en pratique, cette situation ne se présente presque jamais. Nous supposerons systématiquement que notre échantillon est « compatible » avec l'information *a priori*, ce que nous expliquons plus en détail plus loin. Dans nos études par simulation, nous rejetons systématiquement tout échantillon qui n'est pas compatible et nous en choisissons un autre. De nouveau, dans la plupart des cas, la probabilité de devoir rejeter un échantillon est très faible.

Chacun de nos exemples d'information *a priori* peut être représenté par une ou plusieurs contraintes linéaires d'égalité ou d'inégalité. Nous avons vu que connaître la moyenne de population donne une contrainte d'égalité linéaire. Si nous savons que la moyenne de population est comprise dans un certain intervalle, nous obtenons deux contraintes d'inégalité linéaires. Nous allons maintenant établir une notation qui nous permettra de considérer diverses situations où l'information *a priori* peut être décrite en utilisant des égalités et des inégalités linéaires comme contraintes.

Nous supposons qu'en plus de la caractéristique d'intérêt y, la population possède un ensemble de variables auxiliaires $x^1, x^2, \ldots, x^m$. Pour l'unité i, soit

$$(y_i, x_i) = (y_i, x_i^1, x_i^2, \ldots, x_i^m)$$

le vecteur de valeurs de y et de variables auxiliaires. Nous supposons que, pour toute unité échantillonnée, ce vecteur de valeurs est observé. Nous supposons aussi que l'information *a priori* au sujet de la population peut être exprimée à l'aide d'un ensemble de contraintes d'égalité et d'inégalité linéaires pondérées sur les valeurs auxiliaires distinctes dans la population, en prenant pour pondérations les proportions de la population qui prennent ces valeurs individuelles distinctes. Nous illustrons cet aspect plus en détail en expliquant comment nous transposons cette information *a priori* au sujet de la population aux valeurs observées dans l'échantillon, afin de pouvoir construire des pseudo-versions de la population compatibles avec l'information *a priori*.

Soit s un échantillon et, pour $j = 1, 2, \ldots, n(s)$, soit (y_{i_j}, x_{i_j}) les valeurs observées que nous supposons distinctes pour simplifier. Soit $p = (p_1, \ldots, p_{n(s)})$ la proportion d'unités qui sont affectées à la valeur (y_{i_j}, x_{i_j}) dans la réplique complète simulée de la population. Toute contrainte linéaire sur les valeurs de population d'une variable auxiliaire se traduira d'une manière évidente par une contrainte linéaire sur les valeurs simulées. Par exemple, si nous savons que la moyenne de population de x^1 est inférieure ou égale à une valeur donnée, disons b_1, pour la population simulée, cette information devient la contrainte

$$\sum_{j=1}^{n(s)} p_j x_{i_j}^1 \le b_1.$$

Si nous savons que la médiane de population de x^2 est égale à b_2, la contrainte pour la population simulée devient

$$\sum_{j=1}^{n(s)} p_j w_j = 0,5$$

où $w_j = 1$ si $x_{i_j}^2 \le b_2$ et est nul autrement. Si la moyenne de population de x^2 est inférieure ou égale à la moyenne de

population de x^3, alors la contrainte pour la population simulée devient

$$\sum_{j=1}^{n(s)} p_j(x_{i_j}^2 - x_{i_j}^3) \leq 0.$$

Donc, étant donné une famille de contraintes de population fondées sur l'information *a priori* et un échantillon, nous pourrons représenter les contraintes correspondantes sur le paramètre p simulé par deux systèmes d'équations

$$A_{1,s}\, p = b_1 \tag{2}$$

$$A_{2,s}\, p \leq b_2 \tag{3}$$

où $A_{1,s}$ et $A_{2,s}$ sont des matrices de dimensions $m_1 \times n(s)$ et $m_2 \times n(s)$, et b_1 et b_2 sont des vecteurs ayant les dimensions appropriées. Cela généralise l'argument menant à l'équation 1.

Nous supposons l'échantillon tel que le sous-ensemble du simplexe qu'il définit par les équations 2 et 3 n'est pas vide. Pour un tel échantillon, l'approximation asymptotique de la loi *a posteriori* de Pólya contrainte impose une loi uniforme sur ce sous-ensemble du simplexe. Avant d'aborder la question de la simulation à partir de cette loi, nous notons qu'elle possède une justification théorique. Il peut lui être donné une justification bayésienne séquentielle qui garantit qu'elle produira des méthodes admissibles. Des renseignements détaillés figurent en annexe.

3.3 Calculs

Soit P le sous-ensemble du simplexe qui est défini par les équations 2 et 3. P est un polytope de dimension non pleine. Nous aimerions produire des observations indépendantes à partir de la loi uniforme sur P. Malheureusement, nous ne savons pas comment le faire. Nous utilisons plutôt des méthodes de Monte Carlo par chaînes de Markov (MCMC) pour produire des échantillons dépendants.

En particulier, nous utilisons l'algorithme de Metropolis-Hastings qui dépend de l'emploi d'une chaîne de Markov pour produire une séquence dépendante de valeurs aléatoires pour $p \in P$. Le processus se déroule comme suit. Nous commençons par trouver un point de départ en p_0 dans l'intérieur relatif de P. Il s'agit de l'étape 1 énoncée plus bas. Ensuite, nous choisissons une direction aléatoire d dans P, ce qui pose un peu de difficulté parce que la dimension de P est strictement inférieure à $n(s) - 1$. Cela est accompli aux étapes 2 et 3 décrites plus bas. Ensuite, nous trouvons le segment de droite qui est l'intersection de la droite passant par p_0 dans la direction d avec P. Il s'agit de l'étape 4 énoncée plus bas. Puis, nous choisissons un point au hasard à partir de la loi uniforme au-dessus de ce segment de droite. Ce point est la première observation dans notre chaîne de Markov. Nous répétons alors le processus en remplaçant p_0 par ce point pour obtenir un deuxième point aléatoire. En substituant ce deuxième point aléatoire à p_0, nous en obtenons un troisième, et ainsi de suite. De manière plus formelle, notre algorithme est :

Étape 1. Choisir un vecteur de probabilités positives initiales p_0, tel que $A_{1,s}\, p_0 = b_1$ et $A_{2,s}\, p_0 < b_2$ et fixer $i = 0$.

Étape 2. Produire une direction aléatoire d_i uniformément distribuée sur la sphère unitaire dans R^n.

Étape 3. Soit d_i^* la projection normalisée de d_i sur l'espace nul de $A_{1,s}$.

Étape 4. Trouver le segment de droite $L_i = \{\alpha \in R \mid p_i + \alpha d_i^* \in P\}$ et générer α_i uniformément sur ce segment de droite.

Étape 5. Fixer $p_{i+1} = p_i + \alpha_i d_i^*$ et $i = i + 1$ et retourner à l'étape 2.

À première vue, le rôle joué par les contraintes dans ce processus pourrait ne pas être clair. Cependant, elles interviennent par la voie de la définition de P. La chaîne de Markov produite de cette façon converge en loi vers la loi uniforme sur le polytope. Le résultat de convergence des algorithmes de mélange de ce type a été prouvé par Smith (1984). Si nous voulons approximer la valeur espérée d'une fonction définie sur P, la moyenne de la fonction calculée aux valeurs simulées converge vers sa valeur réelle. Cela permet de calculer des estimations ponctuelles des paramètres de population. Trouver approximativement l'intervalle de crédibilité bayésien au niveau à 0,95 est plus difficile.

Une possibilité consiste à exécuter la chaîne pendant une longue période; par exemple, nous pouvons générer 4,1 millions de valeurs, écarter les 100 000 premières et trouver les quantiles 0,025 et 0,975 des valeurs restantes. Ces deux nombres détermineront notre intervalle de crédibilité à 0,95 approximatif. Ici, nous considérons uniquement les tailles d'échantillon inférieures à 100. Dans ces conditions, nous avons constaté que des chaînes de quelques millions de valeurs de long suffisent.

La vitesse de mélange d'une chaîne dépend des contraintes et du paramètre estimé. Il semble que l'obtention d'un bon mélange prend plus de temps quand on estime la médiane que la moyenne, ce qui n'est pas étonnant si l'on se souvient que, dans les méthodes bootstrap standard, beaucoup plus d'échantillons bootstrap sont requis quand on

estime les quantiles plutôt que les moyennes. Voir, par exemple, Efron et Tibshirani (1993).

Une autre approche qui peut donner de bons résultats consiste à exécuter la chaîne pendant une longue période et à n'utiliser que chaque m^{e} point, où m est un grand nombre entier. La méthode est inefficace, mais elle peut donner de bonnes réponses lorsqu'on veut trouver un intervalle de crédibilité à 95 % pour la médiane.

4. Applications

À la présente section, nous montrons comment divers types d'information partielle au sujet des variables auxiliaires peuvent être intégrés dans l'estimation des paramètres quand on utilise la loi *a posteriori* de Pólya contrainte. Dans de nombreux cas, l'information *a priori* utilisée dans l'estimation selon la loi *a posteriori* de Pólya contrainte ne peut pas l'être dans les méthodes fréquentistes classiques.

4.1 Stratification

La stratification est une forme d'information *a priori* qui est utilisée couramment en échantillonnage d'une population finie. Nous notons que l'estimateur stratifié habituel peut être considéré comme découlant de lois *a posteriori* de Pólya indépendantes dans chaque strate. Des renseignements détaillés figurent dans Vardeman et Meeden (1984). Quand, en plus de la stratification, il existe une variable auxiliaire, on peut obtenir une bonne estimation de la moyenne de population en combinant les estimations produites à l'aide de l'estimateur par régression dans chaque strate. Pour plus de précisions, voir Cochran (1976). Si l'on ne connaît que la moyenne de population de la variable auxiliaire, sous les approches classiques, il est difficile de combiner cette information avec la stratification, à moins que l'on émette l'hypothèse d'un modèle commun à toutes les strates. La Pólya contrainte peut intégrer les deux types d'information susceptibles d'améliorer les estimations, pourtant elle ne nécessite pas l'hypothèse d'un modèle commun.

Pour le démontrer, nous avons créé une population stratifiée de 900 valeurs comportant trois strates. Les tailles des strates étaient 300, 200 et 400. Nous avons considéré deux variables auxiliaires, disons x_1 et x_2. Dans la première strate, les $x_{1,i}$ étaient un échantillon aléatoire tiré d'une loi gamma $(10,1)$ et les $x_{2,i}$ étaient un échantillon aléatoire tiré d'une loi gamma $(2,1)$. Dans la deuxième strate, les $x_{1,i}$ et les $x_{2,i}$ étaient tirés des lois gamma $(15,1)$ et gamma $(7,1)$, respectivement. Dans la troisième strate, les $x_{1,i}$ et les $x_{2,i}$ étaient tirés des lois gamma $(5,1)$ et gamma $(3,1)$, respectivement. La caractéristique d'intérêt pour la population a été générée de la façon suivante :

strate 1 : $y_i = 1 + x_{1i}\, x_{2i} + \varepsilon_i$
strate 2 : $y_i = 3 + x_{1i} + x_{1i}\, x_{2i} + \varepsilon_i$
strate 3 : $y_i = 2 + x_{2i} + x_{1i}\, x_{2i} + \varepsilon_i$

où, dans la première strate, les ε_i suivaient une loi normale $(0,1)$, tandis que dans la deuxième strate, elles suivaient une loi normale $(0,1,5^2)$, et dans la troisième strate, elles suivaient une loi normale $(0,3,5^2)$. Toutes les ε_i étaient indépendantes.

En plus des tailles de strate, nous avons supposé que la médiane de population de x_1 et la moyenne de population de x_2 étaient connues. Nous avons produit 500 échantillons aléatoires conformément à notre plan d'échantillonnage en tirant 75 unités de telle sorte que 25 unités étaient dans la première strate, 20, dans la deuxième et 35, dans la troisième. Pour chaque échantillon, nous avons calculé la moyenne d'échantillon, l'estimation stratifiée habituelle qui est la somme des moyennes d'échantillon dans chaque strate corrigée de la taille de toutes les strates, l'estimation sous la loi *a posteriori* de Pólya contrainte, ainsi que les intervalles de confiance à 95 % et les intervalles de crédibilité à 95 % correspondants de ces estimations.

Les résultats des simulations sont présentés au tableau 1. L'examen de ce dernier montre que l'estimateur de Pólya contraint concorde, en moyenne, avec l'estimateur stratifié habituel et est essentiellement sans biais. Par contre, son erreur absolue moyenne est nettement plus faible que celle des deux autres estimateurs, ce qui n'est pas surprenant, puisqu'un estimateur devrait donner des résultats d'autant meilleurs qu'il utilise plus d'information et que les estimations sous la loi *a posteriori* de Pólya contrainte utilisent l'information provenant des variables auxiliaires dont il n'est pas tenu compte dans l'estimation basée uniquement sur la stratification. Notons que les estimations d'après la loi de Pólya contrainte ne s'appuient sur aucune hypothèse quant à la relation entre y, x_1 et x_2. En outre, la façon dont on pourrait utiliser dans les méthodes standard le fait que l'on connait la médiane de population de x_1 et la moyenne de population de x_2. est une question qui reste à éclaircir. Si l'on ne dispose d'information qu'au sujet des moyennes, les méthodes fondées sur la vraisemblance empirique de Chen et Sitter (1999) et de Zhong et Rao (2000) pourraient être utilisées. Les résultats montrent clairement que la loi *a posteriori* de Pólya contrainte utilise cette information supplémentaire de manière sensée.

Dans cet exemple, les estimations au moyen de la loi de Pólya contrainte ont été obtenues en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Tableau 1
Résultats des simulations pour l'exemple stratifié où la médiane de la première variable auxiliaire et la moyenne de la seconde sont connues

Méthode	Estimation ponctuelle		Intervalles de confiance ou de crédibilité à 95 %		
	Estimation moyenne	Erreur absolue moyenne	Borne inférieure moyenne	Longueur moyenne	Fréquence de couverture
Moyenne	47,978	4,821	36,44	23,09	1,000
Stratification	43,395	2,072	38,22	10,35	0,942
Pólya	43,355	1,516	40,19	6,75	0,936

4.2 Variables auxiliaires catégoriques

Supposons que les éléments d'une population de taille connue N sont associés aux éléments de k variables auxiliaires catégoriques. Pour simplifier, nous posons que $k = 2$, mais la théorie s'applique à plus de deux variables catégoriques. Si une variable auxiliaire prend r valeurs distinctes et que l'autre prend c valeurs distinctes, elles permettent de classer les éléments de la population dans un tableau à double entrée contenant $r \times c$ cellules. Soit N_{ij} le nombre d'éléments de la population qui appartiennent à la cellule ij, pour i dans {1, ..., r} et j dans {1, ..., c}, alors $\sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij} = N$. Si les N_{ij} sont connus et que s est un échantillon aléatoire contenant n_{ij} éléments provenant de la cellule ij, alors une bonne estimation de la moyenne de population est donnée par

$$\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{N_{ij}}{N}\, \bar{y}_{ij}^{s},$$

où $\bar{y}_{ij}^{s}$ est la moyenne des n_{ij} éléments provenant de la cellule ij dans l'échantillon. Il s'agit de l'estimateur stratifié habituel, où les cellules du tableau sont considérées comme étant les strates.

Un problème plus ardu est celui de l'estimation de la moyenne de population quand les dénombrements, N_{ij}, ne sont pas connus, mais que les dénombrements totaux de marge le sont. Soit $N_{i\cdot} = \sum_{j=1}^{c} N_{ij}$ les totaux de marge des lignes, pour i dans {1, ..., r} et $N_{\cdot j} = \sum_{i=1}^{r} N_{ij}$, les totaux de marge des colonnes, pour j dans {1, ..., c}. Dans de tels cas, un moyen d'estimer la moyenne de population consiste à appliquer la méthode fréquentiste appelée calage ou méthode itérative du quotient. Selon cette méthode, étant donné un échantillon s, l'estimateur est donné par $\sum_{k\in s} \hat{w}_k y_k$, où les $\hat{w}_k$ ne sont pas les poids de sondage, mais de nouveaux poids attribués aux unités de l'échantillon. Un bon ensemble de poids doit satisfaire deux conditions. La première est que les poids doivent préserver les totaux de marge connus, par exemple, $\sum_{k\in s(\cdot,j)} \hat{w}_k = N_{\cdot j}$ où $s(\cdot, j)$ est la portion de l'échantillon se trouvant dans la j^{e} colonne du tableau à double entrée. La deuxième est que les poids devraient, dans un certain sens, être proches des poids d'échantillonnage $1/\pi_k$, où $\pi_k = P(k \in s)$. Suivant la fonction utilisée pour mesurer la distance, divers estimateurs par calage peuvent être obtenus. Bien qu'il s'agisse d'une idée raisonnable, la sélection de la bonne mesure de distance et l'obtention d'une estimation raisonnable de la variance pour l'estimateur résultant n'ont pas de réponse fréquentiste standard. Pour des détails, voir Deville et Särndal (1992).

La loi *a posteriori* de Pólya donne une approche alternative de résolution du problème, puisque l'information fournie par les totaux de marge connus détermine un ensemble de contraintes linéaires sur les poids aléatoires attribués aux unités échantillonnées. S'il existe des variables auxiliaires continues pour lesquelles nous disposons d'information *a priori*, nous pouvons ajouter des contraintes supplémentaires. Pour voir comment cette approche pourrait fonctionner en pratique, nous avons envisagé l'exemple simple de deux variables dichotomiques, afin que chaque unité puisse être classée dans une cellule d'un tableau 2×2, et d'une troisième caractéristique auxiliaire continue. Nous avons émis l'hypothèse de quatre niveaux distincts d'information *a priori*.

1. Les totaux de marge pour le tableau à double entrée sont connus.
2. Les totaux de marge et la moyenne de la variable auxiliaire continue sont connus.
3. Les totaux de marge et la médiane de la variable auxiliaire sont connus.
4. Les totaux de marge sont connus et l'on sait que la moyenne de la variable auxiliaire continue est comprise entre deux bornes. Nous avons choisi les 45^{e} et 65^{e} quantiles de sa population de valeurs pour spécifier ces bornes.

Pour chaque cas, nous avons formé une population en utilisant le modèle qui suit, où toutes les variables aléatoires sont indépendantes.

Cellule 1,1 $x_i \sim \text{gamma}(8,1)$, $\varepsilon_i \sim \text{normale}(0, 7^2)$ et $y_i = 25 + 3x_i + \varepsilon_i$ pour i dans {1, ..., 150}.

Cellule 1,2 $x_i \sim \text{gamma}(10,1)$, $\varepsilon_i \sim \text{normale}(0, 7^2)$ et $y_i = 25 + 3x_i + \varepsilon_i$ pour i dans {1, ..., 350}.

Cellule 2,1 $x_i \sim \text{gamma}(6,1)$, $\varepsilon_i \sim \text{normale}(0, 4^2)$ et $y_i = 25 + 2x_i + \varepsilon_i$ pour i dans {1, ..., 250}.

Cellule 2,2 $x_i \sim \text{gamma}(4,1)$, $\varepsilon_i \sim \text{normale}(0, 4^2)$ et $y_i = 25 + 2x_i + \varepsilon_i$ pour i dans {1, ..., 250}.

Pour chacun des cas, nous avons généré une population et tiré 500 échantillons aléatoires de taille 80 avec 20 unités

provenant de chaque cellule. Pour chaque échantillon, nous avons calculé la moyenne d'échantillon et l'estimation par stratification, en supposant que les effectifs de population réels des cellules étaient connus, ainsi que les intervalles de confiance à 95 % correspondants. Nous avons également calculé les estimations à partir de la loi *a posteriori* de Pólya contrainte, ainsi que leurs intervalles de crédibilité à 0,95. Nous avons obtenu ces dernières estimations à partir des derniers 4 000 000 de points d'une chaîne de Markov de longueur égale à 4 100 000. Les résultats des simulations sont présentés dans les tableaux 2 à 5. Ils révèlent que les estimations d'après la loi *a posteriori* de Pólya contrainte fondée sur les totaux de marge et une moyenne ou une médiane connue ou un intervalle connu autour de la moyenne sont meilleures que les estimations stratifiées fondées sur les effectifs de cellules connus. Les estimations stratifiées ne sont meilleures que les estimations d'après la loi *a posteriori* de Pólya contrainte que quand cette dernière n'utilise que les totaux de marge connus.

Tableau 2
Résultats des simulations pour l'exemple catégorique quand seuls les totaux de marge sont supposés connus

Méthode	Estimation ponctuelle		Intervalles de confiance ou de crédibilité à 95 %		
	Estimation moyenne	Erreur absolue moyenne	Borne inférieure moyenne	Longueur moyenne	Fréquence de couverture
Moyenne	43,805	0,919	41,107	5,396	0,976
Stratification	44,355	0,846	42,259	4,191	0,940
Pólya	43,909	0,896	41,863	4,197	0,922

Tableau 3
Résultats des simulations pour l'exemple catégorique quand les totaux de marge et la moyenne de la variable auxiliaire sont supposés connus

Méthode	Estimation ponctuelle		Intervalles de confiance ou de crédibilité à 95 %		
	Estimation moyenne	Erreur absolue moyenne	Borne inférieure moyenne	Longueur moyenne	Fréquence de couverture
Moyenne	43,804	0,922	41,063	5,482	0,964
Stratification	44,399	0,862	42,272	4,256	0,942
Pólya	44,506	0,510	43,257	2,497	0,960

Tableau 4
Résultats des simulations pour l'exemple catégorique quand les totaux de marge et la médiane de la variable auxiliaire sont supposés connus

Méthode	Estimation ponctuelle		Intervalles de confiance ou de crédibilité à 95 %		
	Estimation moyenne	Erreur absolue moyenne	Borne inférieure moyenne	Longueur moyenne	Fréquence de couverture
Moyenne	43,439	0,877	40,783	5,312	0,986
Stratification	43,927	0,884	41,804	4,244	0,940
Pólya	43,784	0,785	42,032	3,640	0,920

Tableau 5
Résultats des simulations pour l'exemple catégorique quand les totaux de marge sont supposés connus et que l'on sait que la moyenne de la variable auxiliaire est comprise entre ses 45^e et 65^e quantiles connus

Méthode	Estimation ponctuelle		Intervalles de confiance ou de crédibilité à 95 %		
	Estimation moyenne	Erreur absolue moyenne	Borne inférieure moyenne	Longueur moyenne	Fréquence de couverture
Moyenne	43,463	0,840	40,789	5,348	0,978
Stratification	43,948	0,865	41,825	4,245	0,948
Pólya	43,519	0,829	41,555	4,029	0,938

4.3 Un exemple

À la présente section, nous considérons les données provenant de la Veterans Health Administration. En 1998, le VA Upper Midwest Health Care Network a administré une enquête sur l'état fonctionnel des anciens combattants utilisateurs des établissements de la VA dans le réseau (Singh, Borowsky, Nugent, Murdoch, Zhao, Nelson, Petzel et Nichol 2005). Les anciens combattants admissibles à cette enquête étaient ceux qui avaient reçu des soins au service des consultations externes ou qui avaient été hospitalisés entre octobre 1997 et mars 1998 dans l'un des cinq établissements de la VA du réseau. En plus des mesures démographiques de base, telles que l'âge et le sexe, la principale composante de l'enquête était le questionnaire SF36-V (Kazis, Miller, Clark, Skinner, Lee, Rogers, Spiro, Payne, Fincke, Selim et Linzer 1998). Ce questionnaire sur la qualité de la vie liée à l'état de santé comprend huit sous-échelles pour le fonctionnement physique, les limitations des rôles dus à des problèmes physiques, la douleur corporelle, l'état de santé général, l'énergie/vitalité, le fonctionnement social, les limitations des rôles dues à des problèmes émotionnels et la santé mentale. Ces échelles sont combinées pour produire les scores sommaires de la composante physique (PCS) et de la composante mentale (MCS). L'état de santé est d'autant meilleur que ces scores sont élevés. Des données administratives de la VHA mesurant les principaux états comorbides présents l'année qui a précédé l'enquête ont également été recueillies.

À partir de la population de l'un des cinq établissements, nous avons sélectionné une sous-population comprenant toutes les femmes et un sous-ensemble aléatoire des hommes pour créer une population de 2 500 personnes. Pour les besoins du présent exemple, les nombres de comorbidités ont été regroupés en trois catégories correspondant à une bonne santé, une santé passable et une mauvaise santé. Puis, nous avons tiré à partir de la population 200 échantillons aléatoires stratifiés de taille égale à 100. Les tailles de strate, ainsi que les tailles d'échantillon sont présentées au tableau 6. Notre plan d'échantillonnage prévoyait le suréchantillonnage des femmes. Ce genre de plan d'échantillonnage non équilibré est utilisé fréquemment en pratique.

Tableau 6
Tailles des strates et tailles des échantillons pour les données de la Veterans Administration

	Bonne	Passable	Mauvaise
F	353(20)	155(10)	117(10)
H	890(30)	493(20)	492(10)

Nous avons comparé trois estimateurs différents du score PCS moyen pour cette population de 2 500 personnes, à savoir la moyenne d'échantillon calculée sans tenir compte de la stratification, l'estimateur stratifié habituel en supposant que les tailles de strate sont connues et un estimateur fondé sur la loi *a posteriori* de Pólya contrainte en supposant que les totaux de marge des lignes et des colonnes du tableau 6 étaient connus, de même que l'âge moyen des membres de la population. La corrélation de population entre le score PCS et l'âge est de -0,22. Les corrélations du score PCS avec le sexe et avec la catégorie d'état de santé fondée sur la comorbidité sont égales à -0,13 et -0,28. Les résultats du tableau 7 nous montrent que l'estimateur fondé sur la loi *a posteriori* de Pólya contrainte donne à peu près le même résultat que l'estimateur stratifié et que tous deux sont un peu meilleurs que la moyenne d'échantillon. Afin de calculer l'estimateur fondé sur la loi *a posteriori* de Pólya contrainte, nous avons produit des chaînes de Markov de longueur égale à 7 000 000.

Tableau 7
Résultats de l'estimation du score PCS d'après les données de la Veterans Administration. L'estimateur de la loi *a posteriori* de Pólya contrainte suppose que les totaux de ligne et de colonne sont connus, ainsi que l'âge moyen des membres de la population

Méthode	Estimation ponctuelle		Intervalles de confiance ou de crédibilité à 95 %		
	Estimation moyenne	Erreur absolue moyenne	Borne inférieure moyenne	Longueur moyenne	Fréquence de couverture
Moyenne	37,235	1,040	34,907	4,650	0,938
Stratification	36,648	0,925	34,322	4,651	0,948
Pólya	36,644	0,925	34,344	4,605	0,958

5. Relation avec les méthodes de vraisemblance empiriques

Nous allons maintenant examiner certaines méthodes fréquentistes pour la résolution de problèmes comportant des contraintes et discuter de leur relation avec la loi *a posteriori* de Pólya contrainte.

Chen et Qin (1993) ont étudié une approche de vraisemblance empirique de l'estimation par sondage lorsque l'on dispose d'information *a priori* au sujet d'une caractéristique auxiliaire. Pour construire des estimateurs après que l'échantillon a été observé, les unités échantillonnées sont pondérées de façon à refléter l'information *a priori*. Par exemple, supposons que la moyenne d'échantillon est inférieure à la moyenne de population connue des valeurs de x. Alors, nous choisissons pour les unités échantillonnées des poids positifs, dont la somme est égale à 1, de sorte que la moyenne des valeurs x_s sous la loi de probabilité donnée par les poids satisfasse la contrainte connue. Bien qu'il ne soit pas possible de trouver explicitement ces poids, il est facile de les calculer. Quand nous estimons la moyenne de population de y, nous obtenons un estimateur qui a été mentionné pour la première fois dans Hartley et Rao (1968) et dont l'équivalence asymptotique à l'estimateur par la régression a été démontrée. Si la médiane de population de x est connue, nous attribuons aux unités de l'échantillon dont la valeur est inférieure à cette médiane des poids égaux dont la somme est égale à 0,5 et nous en faisons de même pour les unités échantillonnées dont les valeurs de x sont supérieures. Quand nous estimons la médiane de population, l'estimateur résultant est l'un de ceux proposés par Kuk et Mak (1989).

Un avantage de la loi *a posteriori* de Pólya contrainte et plus généralement de l'approche bayésienne tient au fait qu'il est facile d'estimer un grand nombre de paramètres de population, outre la moyenne, sans devoir élaborer une nouvelle théorie ou de nouvelles méthodes. Étant donné une réplique simulée de la population complète qui satisfait les contraintes, on calcule simplement les paramètres de population d'intérêt. Ensuite, on utilise ces valeurs simulées exactement de la même façon que quand on estime la moyenne.

Pour comparer l'estimateur de Chen et Qin de la médiane de population de y à l'estimateur de la loi *a posteriori* de Pólya contrainte lorsque la moyenne de population de x est connue, nous avons construit huit populations différentes. Pour la moitié de ces populations, on pourrait s'attendre à ce que l'estimateur par la régression donne de bons résultats pour l'estimation de la moyenne de population, tandis que l'autre moitié n'a pas satisfait les hypothèses habituelles de modélisation d'une superpopulation associées à l'estimateur par la régression. Pour chaque population, nous avons tiré 500 échantillons aléatoires de taille égale à 30 et à 50, sous la contrainte que l'échantillon contienne les valeurs de x supérieures et inférieures à la moyenne connue. Dans tous les cas, les deux estimateurs utilisant l'information *a priori* ont donné de meilleurs résultats que la médiane d'échantillon. Ces résultats concordaient avec les résultats de simulation de Chen et Qin. Nous avons calculé l'erreur absolue moyenne pour les deux estimateurs utilisant la contrainte de moyenne. Dans chacun des 16 ensembles distincts de simulation, nous avons alors calculé le ratio de l'erreur absolue de l'estimateur de la loi *a posteriori* de Pólya contrainte à celle de l'estimateur de Chen et Qin. L'intervalle de ces 16 valeurs allait de 0,85 à 1,00 avec une

moyenne de 0,91. Donc, si l'on s'en tient à l'erreur absolue, la performance de l'estimateur de la loi *a posteriori* de Pólya contrainte était supérieure d'environ 10 %, en moyenne, à celle de l'estimateur de Chen et Qin.

Supposons maintenant que la médiane de population de x est connue. Pour simplifier, posons qu'aucune des valeurs réelles n'est égale à la médiane de population de x. Soit n_l le nombre d'unités de l'échantillon dont les valeurs de x sont inférieures à la médiane de population connue de x. Alors, $n_u = n(s) - n_l$ est le nombre d'unités de l'échantillon qui sont situées de l'autre côté de la médiane connue. Soit $p_l = (p_1, ..., p_{n_l})$ et $p_u = (p_1, ..., p_{n_u})$ les deux vecteurs de probabilités. Intuitivement, une loi *a posteriori* raisonnable, étant donné l'échantillon et la médiane de population connue, serait que p_l et p_u suivent des lois de Dirichlet indépendantes dont tous les paramètres sont égaux à un et affectés chacun d'un poids égal à un demi, afin que leur somme soit égale à un. Il découle du théorème prouvé en annexe que, sous notre plan d'échantillonnage, ces lois *a posteriori* sont des bayésiennes séquentielles. Notons que sous ces lois *a posteriori*, les valeurs espérées des proportions affectées à chaque unité de l'échantillon sont les poids affectés à l'échantillon par Chen et Qin. Cela prouve l'admissibilité de leur estimateur de la médiane de population et, conséquemment, de l'estimateur de Kuk et Mak. Les résultats des simulations montrent que les intervalles de crédibilité à 95 % de cette loi *a posteriori* de Pólya contrainte couvrent environ 95 % du temps, sauf dans un cas particulier. Si la taille d'échantillon est faible et que y et x sont fortement corrélés, les médianes des populations simulées sous la loi *a posteriori* de Pólya contrainte ne varient pas suffisamment et les intervalles résultants sont trop courts et leur fréquence de couverture pourrait être considérablement inférieure à 95 %.

Cette relation étroite entre l'approche de vraisemblance empirique et celle de la loi *a posteriori* de Pólya n'est pas étonnante si l'on souligne que, dans le cas non contraint, la séquence des lois *a priori* menant à la loi *a posteriori* de Pólya peut être utilisée pour prouver l'admissibilité de l'estimateur du maximum de vraisemblance pour le vecteur de probabilités d'une loi multinomiale.

6. Autres plans d'échantillonnage

Tous les résultats des simulations présentés jusqu'à présent ont été obtenus en utilisant un plan d'échantillonnage aléatoire simple (stratifié) sans remise (EAS). L'un des examinateurs d'une version antérieure du manuscrit souhaitait savoir dans quelle mesure le comportement des estimateurs fondés sur la loi *a posteriori* de Pólya contrainte dépendent de l'utilisation de ce plan d'échantillonnage. La réponse est qu'il existe une certaine dépendance, mais qu'elle n'est pas aussi prononcée qu'on pourrait l'imaginer au départ.

Nous avons constaté, sous l'EAS, que l'estimateur fondés sur la loi *a posteriori* de Pólya contrainte (PPC) se comporte de façon fort semblable à l'estimateur par la régression (REG). Formellement, l'estimateur par la régression dépend uniquement de la connaissance de la moyenne de population de la variable auxiliaire. Ces propriétés sont habituellement étudiées sous échantillonnage aléatoire simple et l'estimateur de sa variance n'est valide que pour les grands échantillons.

Pour un plan général, l'estimateur d'Horvitz-Thompson (HT) est fréquemment utilisé. Il est sans biais, mais le calcul des probabilités d'inclusion exactes peut être difficile. Il en est ainsi, par exemple, si l'échantillonnage est fait avec probabilités de sélection proportionnelles à la taille pour une variable auxiliaire x, disons PPS(x). En pratique, on suppose simplement que la probabilité d'inclusion d'une unité est proportionnelle à sa valeur de x et que l'estimateur résultant sera approximativement sans biais.

Nous avons exécuté plusieurs études par simulation en vue de comparer ces trois méthodes pour l'estimation d'un total de population. Afin d'abréger, nous présentons les résultats de deux de ces études. Dans le cadre de celles-ci, nous avons construit deux populations de taille égale à 500. La variable auxiliaire est la même dans les deux populations et est un échantillon aléatoire tiré d'une loi gamma dont le paramètre de forme est égal à 5 et le paramètre d'échelle, à 1. Les tracés des deux populations sont présentés à la figure 1. Nous ne sous-entendons pas qu'en pratique, on utiliserait vraisemblablement l'estimateur par la régression dans la deuxième population. Il est présenté ici simplement pour illustrer ce qui peut arriver.

Pour chaque population, nous avons tiré 400 échantillons aléatoires de tailles égales à 30 et à 60 sous deux plans d'échantillonnage différents, à savoir PPS(x) et PPS($1/(x + 5)$). Nous avons supposé que la moyenne de population de x était connue. Pour chaque échantillon, nous avons calculé les trois estimations du total de population. Les résultats pour le premier plan d'échantillonnage sont donnés au tableau 8. Nous voyons que l'estimateur PPC l'emporte clairement sur les autres. La fréquence de couverture observée de l'estimateur HT de l'intervalle est celle qui s'approche le plus du niveau nominal de 95 %. Mais l'intervalle est ridiculement long, parce que les inverses des probabilités d'inclusion varient fortement. Pour la première population, qui est approximativement linéaire, les estimateurs REG et PPC se comportent de la même manière. Cependant, pour la deuxième population, PPC donne de meilleurs résultats que REG. Son seul défaut ici est que sa couverture est insuffisante pour la plus petite taille d'échantillon. Sous PPS($1/(x + 5)$), l'histoire est en grande

partie la même, quoique la différence entre REG et PPC soit nettement plus petite pour la deuxième population. Par exemple, quand la taille d'échantillon est égale à 30, l'erreur absolue moyenne et la fréquence de couverture sont égales à 131,9 et 0,875, respectivement, pour REG comparativement à 124,3 et 0,908 pour PPC. Quand la taille d'échantillon est égale à 60, les chiffres pour REG sont 88,4 et 0,905 comparativement à 90,1 et 0,958 pour l'estimateur PPC. Les longueurs moyennes de leurs intervalles sont égales à 384 et à 560, respectivement.

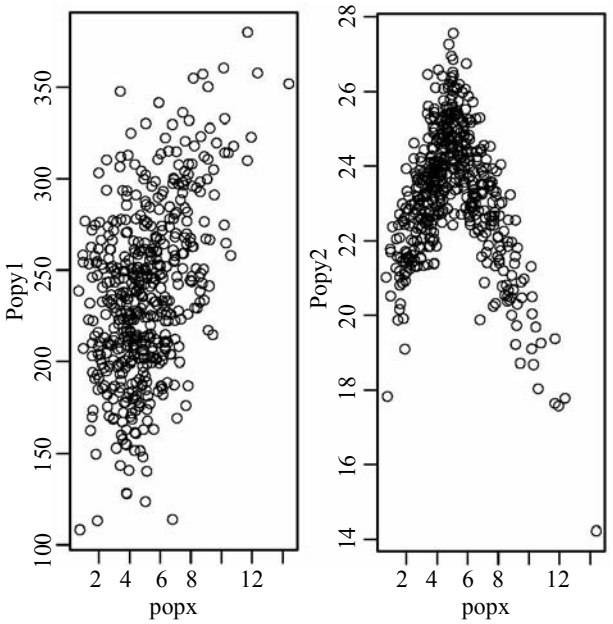

Figure 1 Tracés des deux populations utilisées dans les simulations au tableau 8. Les corrélations pour les deux populations sont égales à 0,47 et -0,22 et leurs totaux sont 118 210 et 11 648,7

Pour la deuxième population, nous avons réalisé un deuxième ensemble de simulations en utilisant le plan PPS(x) pour les tailles d'échantillon de 30 et de 60. Cette fois-ci, nous avons supposé que les moyennes de population de x et x^2 étaient connues. Nous avons alors comparé l'estimateur PPC qui intègre les contraintes sur x ainsi que sur x^2 à l'estimateur par la régression reposant sur l'hypothèse d'une fonction quadratique de x comme modèle. Ces estimateurs sont désignés par PPC2 et REG2 dans le tableau. À première vue, il pourrait paraître étonnant que les résultats pour PPC et PPC2 soient essentiellement les mêmes. Cependant, si l'on y réfléchit, ce résultat est celui auquel on devrait s'attendre. La loi *a posteriori* de Pólya contrainte simule des répliques entières de la population, qui sont « équilibrées » en ce qui concerne x, autrement dit qui concordent avec sa moyenne connue de population. La contrainte supplémentaire qu'une réplique simulée de la population doive aussi être « équilibrée » par rapport à x^2 ajoute peu d'information. Par ailleurs, avec un modèle raisonnable, l'estimateur par la régression peut exploiter l'information supplémentaire, ce qui aboutit à un estimateur ponctuel amélioré, mais la couverture des estimations des intervalles reste insuffisante.

Tableau 8
Résultats des simulations pour les deux populations de la figure 1 quand on estime le total de population. Dans chaque cas, elles sont fondées sur 400 échantillons qui ont été sélectionnés en utilisant PPS(x) comme plan d'échantillonnage. Notons que abserr est la valeur absolue de l'écart entre les valeurs estimée et réelle du total de population

Méthode	Estimation moyenne	Erreur absolue moyenne	Longueur moyenne	Fréquence de couverture
Pour la population 1 avec le total = 118 210,2				
pour une taille d'échantillon de 30				
HT	118 803,1	8 095,3	38 696,6	0,905
REG	116 838,1	3 355,3	14 136,4	0,905
PPC	117 515,7	3 277,3	14 330,7	0,905
pour une taille d'échantillon de 60				
HT	119 139,2	5 395,6	28 233,3	0,952
REG	117 041,4	2 213,2	9 561,0	0,910
PPC	118 041,4	2 195,3	11 836,5	0,938
Pour la population 2 avec le total = 11 648,7				
pour une taille d'échantillon de 30				
HT	11 737,2	783,5	4 012,0	0,945
REG	11 800,3	179,7	533,0	0,745
PPC	11 689,9	122,4	535,4	0,900
REG2	11 660,0	97,2	382,3	0,862
PPC2	11 689,9	122,4	535,4	0,900
PPCbd	11 683,2	116,5	537,0	0,918
pour une taille d'échantillon de 60				
HT	11 774,2	564,8	2 908,2	0,955
REG	11 795,8	155,2	373,1	0,635
PPC	11 647,9	80,4	524,4	0,978
REG2	11 663,1	66,7	266,2	0,895
PPC2	11 651,2	88,4	523,6	0,962
PPCbd	11 644,6	83,9	552,1	0,978

Pour la deuxième population, nous avons exécuté un troisième ensemble de simulations en utilisant PPS(x) comme plan pour les tailles d'échantillon de 30 et de 60. Dans ce cas, nous avons supposé que la moyenne de population de x était contenue dans l'intervalle (4,45, 5,53). Ces valeurs correspondent au 45^e et au 65^e quantiles de la population x. La moyenne de cette population est égale à 5,02. Les résultats sont présentés au tableau 8 sous l'étiquette PPCbd. Nous voyons que les résultats sont fort semblables à ceux obtenus en supposant que la moyenne de population de x était connue.

Les trois estimateurs utilisent l'information contenue dans la variable auxiliaire x, mais l'estimateur HT est le seul qui dépend de la connaissance du plan d'échantillonnage. Comme nous l'avons mentionné, il est bien connu que les estimateurs bayésiens ne font pas entrer les probabilités d'échantillonnage dans les calculs. Dans les exemples présentés ici, nous voyons que l'estimateur PPC utilise efficacement l'information contenue dans la variable auxiliaire. En général, la loi *a posteriori* de Pólya et ses

variantes, telles que la loi *a posteriori* de Pólya contrainte, ne dépendent pas directement de l'échantillonnage aléatoire simple, de l'échantillonnage aléatoire stratifié ou de tout autre plan d'échantillonnage. Leur caractère approprié et leur performance dépendent de la concordance entre la structure sous-jacente de la population et celle spécifiée dans la loi prédictive choisie.

La notion fondamentale qui étaye la PPC est que l'on devrait utiliser l'échantillon et l'information auxiliaire disponibles pour simuler des répliques complètes, représentatives, de la population. Dans de simples exemples tels que ceux donnés plus haut, nous voyons que son estimateur ponctuel devrait avoir d'excellentes propriétés fréquentistes pour une vaste classe de plans d'échantillonnage et que la performance de son estimateur par intervalle est adéquate si la taille d'échantillon n'est pas trop faible. Cela signifie-t-il que la PPC peut ajuster automatiquement de « mauvais » échantillons afin d'obtenir de bonnes estimations? Pas vraiment, puisque si nous avons un très mauvais échantillon, c'est-à-dire un échantillon qui est médiocrement compatible avec l'information *a priori* connue, deux choses indésirables peuvent se produire. Premièrement, des échantillons très déséquilibrés ou biaisés induiront un certain biais dans l'estimation ponctuelle. Deuxièmement, ils restreindront sévèrement les valeurs possibles de p sous la PPC et donneront lieu à une variance *a posteriori* trop faible, ce qui donnera des intervalles estimés qui sont trop courts et qui couvrent insuffisamment le paramètre d'intérêt. Dans les situations plus compliquées, il faudra poursuivre l'étude afin de découvrir quand l'utilisation de la loi *a posteriori* de Pólya contrainte peut être avantageuse.

7. Remarques finales

L'un des inconvénients des méthodes fréquentistes classiques est que chaque problème demande sa propre solution. Estimer la médiane de population de y quand la moyenne de population x est connue n'est pas le même problème qu'estimer la moyenne de y quand la moyenne de x est connue. En outre, si la moyenne de population de x n'est pas connue exactement et que l'on sait seulement qu'elle est comprise dans un certain intervalle de valeurs, les méthodes fréquentistes classiques ne peuvent pas utiliser cette information. L'une des forces d'une approche bayésienne tient au fait que, si l'on dispose d'une loi *a posteriori* qui combine raisonnablement l'échantillon avec l'information *a priori*, les inférences peuvent être faites pour de nombreux paramètres de population d'intérêt en simulant simplement des répliques complètes de la population.

Ici, nous avons soutenu que la loi *a posteriori* de Pólya contrainte est une méthode raisonnable d'introduction d'information *a priori* objective au sujet des variables auxiliaires dans une approche bayésienne non informative de l'échantillonnage d'une population finie. Les estimateurs ponctuels résultants ont une justification bayésienne séquentielle qui garantit leur admissibilité. Leurs intervalles de crédibilité à 95 % sont habituellement des intervalles de confiance à 95 % approximatifs et ils fournissent des réponses raisonnables pour des problèmes pour lesquels aucune méthode fréquentiste classique n'est disponible. Cela démontre une force importante de la loi *a posteriori* de Pólya. Une fois que l'on peut simuler des répliques raisonnables de la population complète, l'inférence pour toute une gamme de problèmes devient simple. Par contre, on doit utiliser des méthodes MCMC pour leur calcul. Tous nos calculs ont été effectués en R (R Development Core Team 2005). Deux des auteurs ont diffusé récemment un progiciel R appelé Polyapost qui permet à d'autres d'utiliser facilement nos méthodes. Ici, nous nous sommes limités à des échantillons de taille inférieure à 100, uniquement pour des raisons pratiques, afin de pouvoir exécuter nos simulations en un temps raisonnable. En pratique, pour un échantillon spécifique plus grand, il suffit d'exécuter une chaîne plus longue. Puis, on peut utiliser certains diagnostics standard pour décider si elle semble ou non avoir convergé.

Annexe

Une preuve d'admissibilité

La justification théorique fondamentale pour les estimateurs ponctuels découlant de la loi *a posteriori* de Pólya est qu'ils sont admissibles. Les preuves d'admissibilité s'appuient sur la nature bayésienne séquentielle de la loi *a posteriori* de Pólya. Nous présentons ici une preuve pour les estimateurs ponctuels fondés sur la loi *a posteriori* de Pólya contrainte.

Dans ces arguments bayesiens séquentiels, nous sélectionnons une séquence finie de sous-ensembles disjoints de l'espace des paramètres, où l'ordre est important. Nous définissons une loi *a priori* différente pour chacun de ces sous-ensembles. Ensuite, nous trouvons l'estimateur de Bayes pour chaque point échantillon qui reçoit une probabilité positive sous la première loi *a priori*. Puis, nous trouvons l'estimateur de Bayes pour tout point échantillon qui reçoit une probabilité positive sous la deuxième loi *a priori* et qui n'est pas considéré sous la première loi *a priori*. Ensuite, nous considérons la troisième loi *a priori*, et ainsi de suite. Pour un point échantillon particulier, la valeur de l'estimation bayesienne séquentielle est la valeur de l'estimateur bayesien provenant de l'étape à laquelle ce point a été considéré. La nature bayesienne séquentielle de la loi *a posteriori* de Pólya explique ses propriétés quelque peu paradoxales. Pour un échantillon donné, elle se

comporte exactement comme une loi *a posteriori* bayesienne correcte, mais on ne doit jamais spécifier explicitement de loi *a priori*. Pour plus de détails et la discussion de ces points, consulter Ghosh et Meeden (1997).

Pour prouver l'admissibilité des estimateurs découlant de la loi *a posteriori* de Pólya pour l'espace des paramètres $[0, \infty)^N$, la partie principale de l'argument bayésien séquentiel commence par l'hypothèse que l'espace des paramètres est Λ^N, où Λ est un ensemble fini arbitraire de nombres réels positifs. Une fois que l'admissibilité a été démontrée pour un espace général Λ, l'admissibilité pour l'espace des paramètres $[0, \infty)^N$ s'ensuit facilement. Un argument semblable sera utilisé pour la loi *a posteriori* de Pólya contrainte.

Le traitement des contraintes sur des populations finies pose certaines problèmes techniques qu'il est difficile de régler. Par conséquent, nous supposerons que la population est suffisamment grande comparativement à la taille de l'échantillon pour que la forme approximative de la loi *a posteriori* de Pólya faisant intervenir la loi de Dirichlet est appropriée. Pour simplifier, nous supposons que la population U est infinie.

Nous supposons que pour tout j dans U, $(y_j, X_j) = a_i$ pour un i dans $\{1, ..., k\}$, où $a_i = (a_{i1}, ..., a_{i(m+1)})$ sont des vecteurs distincts dans R^{m+1} et où k peut être très grand. Autrement dit, les vecteurs (y_j, X_j) ne peuvent prendre qu'un nombre fini de valeurs. Si p_i est la proportion de valeurs (y_j, X_j) dans la population qui sont égales à a_i, pour i dans $\{1, ..., k\}$, la moyenne de population de Y est $\sum_{i=1}^{k} p_i a_{i1}$.

Nous supposons que de l'information *a priori* est disponible au sujet des variables auxiliaires $X^i := \{x_j^i \mid j \in U\}$ pour i dans $\{1, ..., m\}$, ce qui donne lieu à des égalités et à des inégalités linéaires faisant intervenir les proportions p de la forme

$$A_1 p = b_1 \tag{4}$$

$$A_2 p \leq b_2 \tag{5}$$

où A_1, A_2 sont les matrices de dimensions $m_1 \times k$ et $m_2 \times k$, respectivement, et b_1, b_2, les vecteurs de dimensions appropriées. Dans ces conditions, par exemple, nous pouvons vouloir estimer

$$\mu(p) = \sum_{i=1}^{k} p_i a_{i1},$$

sous les contraintes des équations 4 et 5 et où $\sum_{i=1}^{i=k} p_i = 1$ avec $p_i \geq 0$, pour tout i dans $\{1, ..., k\}$.

Considérons un échantillon s de taille n que, pour faciliter la notation, nous supposerons être composé de n a_i distincts. Soit a_s cet ensemble de valeurs. Nous posons alors que $A_{1,s}$ et $A_{2,s}$ sont les matrices de dimensions $m_1 \times n$ et $m_2 \times n$ qui sont simplement A_1 et A_2 restreintes aux colonnes correspondant aux membres de a_s. Soit p_s la probabilité p restreinte aux membres de a_s. Alors, les contraintes sur la population données par les équations 4 et 5 se traduisent dans les contraintes suivantes

$$A_{1,s} p_s = b_1 \tag{6}$$

$$A_{2,s} p_s \leq b_2 \tag{7}$$

pour les poids aléatoires affectés aux membres de l'échantillon. Autrement dit, étant donné un échantillon, la loi *a posteriori* de Pólya contrainte est simplement la loi uniforme sur le sous-ensemble du simplexe défini par les équations 6 et 7.

Une difficulté technique lorsque l'on prouve l'admissibilité sous des contraintes est que, même si la population satisfait les contraintes énoncées, il est toujours possible d'obtenir un échantillon qui ne les satisfait pas. Il existe plusieurs moyens de traiter ce genre de cas. L'un d'eux consiste à supposer que les contraintes sont incorrectes et de les ignorer. Cette tactique a été utilisée dans Nelson et Meeden (1998). Un autre consiste à utiliser l'information *a priori* pour augmenter l'échantillon de sorte qu'il satisfasse les contraintes. Cela peut être un exercice délicat et la réponse peut dépendre fortement de la façon dont l'échantillon est ajusté. Nous adoptons ici une troisième approche.

Nous supposons que nous suivons un plan d'échantillonnage aléatoire simple et que notre information *a priori* doit être correcte. Dans ces conditions, il pourrait être sensé de rejeter tout échantillon qui ne satisfait pas les contraintes, puisqu'il s'agit clairement d'un échantillon non représentatif. Plus précisément, supposons que nous tirons un échantillon aléatoire simple de taille n de la population et que nous observons tous les x_j^i dans l'échantillon. Soit $p^s = (p_1^s, ..., p_k^s)$ les proportions des vecteurs possibles pour les x_j qui sont observés dans l'échantillon. L'élément p_i^s est nul chaque fois que le vecteur $a_i = (a_{i2}, ..., a_{i(m+1)})$ ne figure pas dans l'échantillon. Si p^s satisfait les équations 6 et 7, nous gardons l'échantillon, sinon nous l'écartons et essayons de nouveau. Nous donnons à ce plan le nom d'échantillonnage aléatoire sous contraintes. En pratique, pour des contraintes typiques, il ne sera jamais nécessaire d'écarter un échantillon. Bien que ce plan d'échantillonnage ne soit jamais utilisé, il ne s'agit pas d'une mauvaise approximation de ce qui est fait en réalité.

Plus formellement, soit Z_i le nombre de (y_j, X_j) dans l'échantillon qui sont égales à a_i, pour i dans $\{1, ..., k\}$, alors $(Z_1, ..., Z_k)$ est une loi Multinomiale$(n, p_1, ..., p_k)$ où les valeurs des paramètres appartiennent à

$$P := \left\{ (p_1, ..., p_k) \middle| \begin{array}{c} A_1 p = b_1, A_2 p \leq b_2, \sum_{i=1}^{k} p_i = 1, \\ \text{et} \quad p_i \geq 0 \forall i \in \{1, ..., k\} \end{array} \right\}. \tag{8}$$

Pour un échantillon donné $s = (z_1, \ldots, z_k)$, soit

$$P^s := \{p \mid p \in P \text{ et } p_i = 0 \text{ quand } p_i^s = 0 \text{ pour } i = 1, \ldots, k\}. \quad (9)$$

Nous voyons que nous gardons un échantillon si, et uniquement si, P^s n'est pas vide.

Dénotons le simplexe à $k-1$ dimensions par

$$F := \left\{(p_1, .., p_k) \middle| \sum_{i=1}^{k} p_i = 1, p_i \geq 0 \forall i \in \{1, \ldots, k\}\right\}.$$

Pour $i = 1, \ldots, k$, soit e_i les sommets de F. Les e_i sont les vecteurs unitaires dont la i^e valeur est 1 et les autres sont 0.

Or, P est un polytope convexe qui est l'intersection de F avec l'espace

$$G := \{(p_1, \ldots, p_k) \mid A_1 p = b_1, A_2 p \leq b_2\}.$$

Une partition de l'espace des paramètres P peut être obtenue de la manière suivante. Soit F_j l'ensemble de facettes de dimension j du simplexe F, $j = 0, 1 \ldots, k-1$. Alors, F_0 est l'ensemble de ses sommets, F_j est la série d'enveloppes convexes de toutes les combinaisons de $j+1$ sommets, pour $j = 1, \ldots, k-2$ et F_{k-1} est le simplexe F. Si $\text{int}(F_j)$ est l'ensemble des intérieurs des facettes de dimension j, pour $j = 1, \ldots, k-1$, alors $\{F_0, \text{int}(F_1), \ldots, \text{int}(F_{k-1})\}$ détermine une partition du simplexe F. Si $G_0 := F_0 \bigcap G$ et $G_j := \text{int}(F_j) \bigcap G$ pour $j = 1, \ldots, k-1$, alors $\{G_0, G_1, \ldots, G_{k-1}\}$ est une partition de l'espace des paramètres P. Notons que certains des G_j pourraient être vides. Les étapes de l'argument bayésien séquentiel suivent les membres non vides des G_j.

Si Z est l'espace échantillon des dénombrements $(Z_1, \ldots, Z_k)$, alors pour $p \in P$, la distribution des dénombrements, disons $f_P(z \mid p)$ est Multinomiale$(n, p_1, \ldots, p_k)$, quand la taille d'échantillon est n. Soit P_F la restriction de l'espace des paramètres P à F, où ce dernier est un sous-ensemble de P et Z_{P_F} est la restriction de l'espace échantillon Z déterminé par P_F.

Nous sommes maintenant prêts à prouver l'admissibilité de l'estimateur fondé sur la loi *a posteriori* de Pólya contrainte de $\mu(p)$ sur P. Supposons que nous sommes à l'étape où nous considérons G_j, pour un certain $j \geq 0$. Supposons que $G_j = G \bigcap F$ n'est pas vide pour un certain sous-ensemble. Deux cas sont possibles.

Cas 1. Si la dimension de G_j est nulle, c'est-à-dire s'il est constitué d'un vecteur, disons p^0, alors nous prenons la loi *a priori* qui applique une masse unitaire à ce vecteur. Dans ces conditions, la loi *a posteriori* applique aussi une masse unitaire à ce vecteur et, si z est le membre unique de Z_{P_F} alors l'estimateur bayésien est $\delta_{\pi_{P_F}}(z) = E(\mu(p) \mid z) = \mu(p^0)$.

Cas 2. Si la dimension de G_j est supérieure à zéro, alors la distribution de $(Z_1, \ldots, Z_k)$ restreinte à Z_{P_F} est

$$f_{P_F}(z \mid p) = \frac{f_P(z \mid p)}{\sum_{z \in Z_{P_F}} f_P(z \mid p)}.$$

La loi *a priori* que nous considérons sur P_F est

$$\pi_{P_F}(p) \frac{\sum_{z \in Z_{P_F}} f_P(z \mid p)}{\prod_{\{i \mid p \in P_F, p_i > 0\}} p_i},$$

qui peut être normalisée afin d'être une loi *a priori* correcte, puisque $\sum_{z \in Z_{P_F}} f_P(z \mid p)$ peut s'écrire $g(p) \prod_{\{i \mid p \in P_F p_i > 0\}} p_i$ où $g(p)$ est une fonction bornée de p. Avec cette loi *a priori*, la loi *a posteriori* est un noyau de la densité de Dirichlet restreint à P_F,

$$f_{P_F}(p \mid z) \propto f_{P_F}(z \mid p) \pi_{P_F}(p) \propto \prod_{\{i \mid p \in P_F, p_i > 0\}} p_i^{z_i - 1}.$$

L'estimateur bayésien de $\mu(p)$ en fonction de π_{P_F}, où p appartient à P_F, est $\delta_{\pi_{P_F}}(z) = E(\mu(p) \mid z)$ pour tout z dans Z_{P_F}. Donc, si nous utilisons la série de priors

$$\{\{\pi_{P_F \mid F \in G_0}\}, \{\pi_{P_F \mid F \in G_1}\}, \{\pi_{P_F \mid F \in G_2}\}, \ldots, \{\pi_{P_F \mid F \in G_{\gamma-1}}\}\},$$

en ignorant à chaque étape les ensembles vides, alors l'estimateur $\delta(z)$ défini par

$$\delta(z) = \delta_{\pi_{P_F}}(z) \text{ pour } z \in Z_{P_F}, F \in G_i, i = 1, \ldots, \gamma - 1, \quad (10)$$

où $\gamma = k$ si $k < n$ et n si $k \geq n$, est un estimateur admissible de $\mu(p)$. Ceci conclut la preuve du théorème suivant.

Théorème 1. *Sous le plan d'échantillonnage aléatoire sous contraintes défini par les équations 2 et 3 avec l'espace des paramètres défini par l'équation 8, l'estimateur fondé sur la loi a posteriori de Pólya contrainte donné par l'équation 10 pour estimer la moyenne de population est bayésien séquentiel et donc admissible sous une perte d'erreur quadratique.*

Remerciements

Les travaux de recherche de Glen Meeden et de Radu Lazar ont été financés en partie par la NSF Grant DMS 0406169. Les auteurs remercient les examinateurs et le rédacteur adjoint de plusieurs commentaires constructifs.

Bibliographie

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

Booth, J.G., Bulter, R.W. et Hall, P. (1994). Bootstrap methods for finite population sampling. *Journal of the American Statistical Association*, 89, 1282-1289.

Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chen, J., et Sitter, S.S. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9, 385-406.

Chen, J., et Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.

Cochran, W. (1976). *Sampling Techniques*. New York : John Wiley & Sons, Inc., 3ème Édition.

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

Efron, B., et Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, Londres.

Feller, W. (1968). *An Introduction of Probability Theory and its Applications*. New York : John Wiley & Sons, Inc., Volume I.

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

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

Gross, S. (1980). Median estimation in survey sampling. Dans *Proceedings of the Section of Section on Survey Research Methods*, American Statistical Association, 181-184.

Hartley, H.O., et Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 159-167.

Kazis, L.E., Miller, D.R., Clark, J., Skinner, K., Lee, A., Rogers, W., Spiro, A. 3ème, Payne, S., Fincke, G., Selim, A. and Linzer, M. (1998). Health related quality of life in patients served by the department of veterans affairs. *Archives of Internal Medicine*, 158, 626-632.

Kuk, A.Y.C., et Mak, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society*, Séries B, 51, 261-269.

Lo, A. (1988). A Bayesian bootstrap for a finite population. *Annals of Statistics*, 16, 1684-1695.

Mak, T.K., et Kuk, A. (1993). A new method for estimating finitepopulation quantiles using auxiliary information. *Canadian Journal of Statistics*, 21, 29-38.

Meeden, G. (1999). A noninformative Bayesian approach for two-stage cluster sampling. *Sankhyā*, Séries A, 61, 133-144.

Meeden, G. (2005). A noninformative bayesian approach to domain estimation. *Journal of Statistical Planning and Inference*, 129, 85-92.

Nelson, D., et Meeden, G. (1998). Using prior information about population quantiles in finite population sampling. *Sankhyā*, Séries A, 60, 426-445.

Nelson, D., et Meeden, G. (2006). Noninformative nonparametric finite population quantile estimation. *Journal of Statistical planning and Inference*, 136, 53-67.

R Development Core Team (2005). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rao, J.N.K., Kovar, J.G. et Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.

Rubin, D. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.

Singh, J.A., Borowsky, S.J., Nugent, S., Murdoch, M., Zhao, Y., Nelson, D., Petzel, R. et Nichol, K.L. (2005). Health related quality of life, functional impairment, and health care utilization by veterans: Veterans' quality of life study. *Journal of American Geriatric Society*, 53, 108-113.

Smith, R.L. (1984). Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32, 1296-1308.

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

Vardeman, S., et Meeden, G. (1984). Admissible estimators for the total of a stratified population that employ prior information. *Annals of Statistics*, 12, 675-684.

Woodruff, R.S. (1952). Confidence intervals for the median and other positive measures. *Journal of the American Statistical Association*, 47, 635-646.

Zhong, B., et Rao, J.N.K. (2000). Empirical likelihood inference under stratified random sampling using auxiliary population infomation. *Biometrika*, 87, 920-938.

Techniques d'enquête, juin 2008
Vol. 34, N° 1, pp. 71-86
Statistique Canada, N° 12-001-X au catalogue

Répartition optimale de l'échantillon pour une régression convergente sous le plan dans une enquête sur les services aux cancéreux quand les variables du plan d'échantillonnage sont connues pour des agrégats

Alan M. Zaslavsky, Hui Zheng et John Adams [1]

Résumé

Nous considérons les taux d'échantillonnage optimaux dans des plans d'échantillonnage par élément, quand l'analyse prévue est la régression linéaire pondérée par les poids de sondage et que les paramètres à estimer sont des combinaisons linéaires des coefficients de régression provenant d'un ou de plusieurs modèles. Nous commençons par élaborer des méthodes en supposant que des renseignements exacts sur les variables du plan existent dans la base de sondage, puis nous les généralisons à des situations où l'information pour certaines variables du plan n'est disponible que sous forme d'agrégat pour des groupes de sujets éventuels ou provient de données inexactes ou périmées. Nous envisageons également un plan d'échantillonnage pour l'estimation de combinaisons de coefficients provenant de plus d'un modèle. Une généralisation supplémentaire permet d'utiliser des combinaisons flexibles de coefficients choisies pour améliorer l'estimation d'un effet tout en en contrôlant un autre. Les applications éventuelles comprennent l'estimation des moyennes pour plusieurs ensembles de domaines chevauchants, ou l'amélioration des estimations pour des sous-populations telles que les races minoritaires par échantillonnage non proportionnel des régions géographiques. Dans le contexte de la conception d'un sondage sur les soins reçus par les cancéreux (l'étude CanCORS) qui a motivé nos travaux, l'information éventuelle sur les variables du plan d'échantillonnage comprenait des données de recensement au niveau de l'îlot sur la race/ethnicité et la pauvreté, ainsi que des données au niveau individuel. Pour un emplacement de l'étude, un plan d'échantillonnage avec probabilités inégales en utilisant les adresses résidentielles des sujets et des données de recensement réduirait la variance de l'estimateur d'un effet du revenu de 25 %, ou de 38 % si la race des sujets avait été connue également. Par pondération flexible des contrastes du revenu selon la race, la variance de l'estimateur serait réduite de 26 % en utilisant les adresses résidentielles seulement et de 52 % en utilisant les adresses et les races. Nos méthodes seraient utiles dans les études où l'on considère un suréchantillonnage géographique selon la race-ethnicité ou les caractéristiques socioéconomiques, ou dans toute étude où les caractéristiques pour lesquelles des données sont disponibles dans les bases de sondage sont mesurées avec une erreur.

Mots clés : Grandeur descriptive de population; erreur de mesure; répartition de Neyman; modèles de régression; plan d'échantillonnage; sondage.

1. Introduction

Un sondage doit être conçu afin d'obtenir des données qui serviront à estimer les coefficients d'un ou de plusieurs modèles de régression. On dispose d'information sur la distribution des covariables dans la population et la base de sondage contient également certains renseignements sur les covariables. Comment cette information peut-elle être utilisée pour rendre le plan de sondage plus efficace ? De combien un tel plan de sondage peut-il réduire la variance comparativement à l'échantillonnage aléatoire simple et comment variera cette réponse si la base de sondage fournit les distributions des covariables agrégées sur des groupes, mais non pour les sujets pris individuellement ?

Ces questions ont été soulevées au moment de la conception d'un sondage sur les processus de fourniture de soins de santé (tels que la fourniture d'une chimiothérapie quand il convient) et les résultats (tels que la qualité de vie après le traitement) auprès d'un grand échantillon de cancéreux à sept emplacements aux États-Unis mis sur pied dans le cadre de l'étude CanCORS (Cancer Care Outcomes Research and Surveillance) (Ayanian, Chrischilles, Wallace, Fletcher, Fouad, Kiefe, Harrington, Weeks, Kahn, Malin, Lipscomb, Potosky, Provenzale, Sandler, Vanryn et West 2004). L'un des principaux objectifs de l'étude était d'estimer les effets conjugués de la race et du revenu sur ces mesures, en utilisant des modèles de régression incluant ces deux caractéristiques des patients. Toutefois, on ne disposait que de données limitées au moment de l'échantillonnage des patients en vue de leur participation à l'étude. L'expérience acquise donnait à penser que la race et l'adresse résidentielle pourraient être déterminées de manière raisonnablement exacte au moment de la confirmation des cas en vue de leur recrutement éventuel pour l'étude, mais que le revenu ne pourrait pas être déterminé tant que le sujet n'aurait pas été recruté et interviewé, et que les renseignements ne pourraient pas être recueillis de manière commode durant une entrevue de sélection. Nous avons entrepris les travaux de recherche présentés ici afin de déterminer comment les données disponibles sur les patients pourraient

1. Alan M. Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, États-Unis; Hui Zheng, Biostatistics Center, Massachusetts General Hospital, 50 Staniford Street, Boston, MA 02114, États-Unis; John Adams, RAND Corporation, 1776 Main Street, Santa Monica, CA 90401.

être combinées à des données de recensement sur les distributions race-revenu dans les îlots de recensement pour procéder à un échantillonnage non proportionnel des patients et, donc, améliorer les estimations des effets de la race et du revenu.

Ce genre de préoccupations surviennent souvent quand des données de sondage sont utilisées pour estimer les coefficients d'un ou de plusieurs modèles de régression. Par exemple, la National Health Interview Survey (NHIS) comporte un suréchantillonnage géographique conjugué à une interview de sélection pour suréchantillonner les répondants noirs et hispaniques de manière à améliorer l'estimation par domaine (Botman, Moore, Moriarity et Parsons 2000, page 12); les données de la NHIS ont été utilisées à grande échelle dans des analyses par régression, dont l'estimation par domaine est un cas particulier. Sastry, Ghosh-Dastidar, Adams et Pebley (2005, pages 1 013-1 014) ont suréchantillonné des secteurs de recensement en fonction de leur composition en groupes minoritaires, en recourant à des simulations pour évaluer la puissance de divers plans d'échantillonnage pour les analyses par régression d'intérêt. Le Youth Risk Behavior Surveillance System suréchantillonne les *écoles* dans les UPE où la prévalence des minorités est forte afin d'améliorer la précision des estimations pour les groupes raciaux/ethniques minoritaires (Eaton, Kann, Kinchen, Ross, Hawkins, Harris, Lowry, McManus, Chyen, Shanklin, Lim, Grunbaum et Wechsler 2006, pages 2-3).

La littérature sur la conception optimale des expériences est abondante. Toutefois, les objectifs de conception des sondages diffèrent de manière importante de ceux des expériences dans lesquelles le chercheur peut attribuer arbitrairement *a priori* des unités identiques aux traitements. Une approche de l'estimation des coefficients de régression fortement fondée sur un modèle dicterait la sélection d'un ensemble approprié d'observations à fort effet de levier, en grande partie comme dans la conception d'une expérience (Royall 1970), mais l'application de ces principes à la conception des sondages est controversée (voir la section 4). L'approche fondée sur le plan de sondage requiert un échantillon qui est représentatif, grâce à un mécanisme probabiliste connu, d'une population définie; des positions intermédiaires sont également possibles (sections 2.5, 3.4). De ce point de vue, l'échantillonneur n'est pas libre de sélectionner, par exemple, 100 répondants blancs dans un quartier dont, de façon commode, la population est principalement blanche et 100 répondants hispaniques dans un quartier dont, de façon commode, la population est principalement hispanique et dire que l'échantillon est « représentatif » pour l'estimation des différences entre les Blancs et les Hispaniques. Les objets de l'inférence fondée sur le plan de sondage sont des grandeurs qui décrivent la population; dans le cas de la régression, nous parlerons de régression d'une « grandeur descriptive de la population » (*descriptive population quantity*) ou GDP (Pfeffermann 1993, pages 319 à 321).

Depuis Neyman (1934), l'abondante littérature sur la conception optimale des sondages (recensée dans des ouvrages classiques tels que Cochran 1977 ou Särndal, Swensson et Wretman 1992) a porté principalement sur l'estimation d'une grandeur simple, telle qu'une moyenne ou un ratio, ou de plusieurs grandeurs de ce genre (Kish 1974; Bellhouse 1984; Chromy 1987). Bien que l'estimation de la variance des estimations fondées sur le plan de sondage des coefficients de régression ait fait couler beaucoup d'encre (Fuller 1975; Fuller 1984; Binder 1981; Binder 1983), assez peu d'attention a été accordée aux plans d'échantillonnage optimaux correspondants. [L'estimation *assistée* par la régression d'une moyenne (Cassel, Särndal et Wretman 1976; Särndal, Swensson et Wretman 1992, section 12.2) est un problème distinct.]

De surcroît, les caractéristiques qui pourraient être utilisées pour définir un plan d'échantillonnage à probabilités inégales sont vraisemblablement enregistrées avec une erreur dans les bases de sondage, parce qu'elles sont fondées sur des données agrégées ou que les caractéristiques associées à une unité (comme une adresse ou un ménage) changent au cours du temps. Ce genre d'erreur peut affecter fortement l'efficacité d'un plan d'échantillonnage qui est censé être optimal; voir Morris, Newhouse et Archibald (1979, section III) au sujet de l'échantillonnage stratifié pour l'estimation par domaine et Thomsen, Tesfu et Binder (1986) au sujet de l'échantillonnage avec probabilité proportionnelle à la taille. Waksberg (1973, 1995) considère la stratification par îlot de recensement sur une seule caractéristique agrégée pour l'estimation des moyennes de domaines tels que des groupes raciaux/ethniques ou les pauvres, avec ou sans interview de sélection subséquente.

Notre objectif ici est de décrire des plans optimaux pour des échantillons qui seront utilisés en analyse par régression (pondérée par les poids de sondage) d'une grandeur descriptive de la population (GDP), au sens de la minimisation de la somme pondérée des variances de certaines combinaisons linéaires présélectionnées de coefficients de régression. Nous examinons aussi certaines classes de grandeurs à estimer et les estimateurs correspondants qui s'écartent de l'approche GDP afin d'accroître l'efficacité. À la section 2, nous établissons la notation et dérivons les taux d'échantillonnage optimaux pour la régression GDP sous des scénarios représentatifs de l'information au niveau individuel et au niveau du domaine dont on pourrait disposer dans des sondages de population quand la base de sondage est imparfaite. Nous supposons d'abord que nous disposons d'information exacte sur les variables du plan de

sondage dans la base de sondage, puis nous généralisons à des situations où l'information sur certaines variables n'est disponible que sous forme d'agrégat pour des sous-domaines ou provient de données inexactes. Ensuite, nous envisageons l'estimation optimale de combinaisons de coefficients provenant de plus d'un modèle et de combinaisons flexibles de coefficients. À la section 3, nous estimons les avantages possibles de ces méthodes pour un sondage réalisé aux emplacements choisis pour l'étude CanCORS, en utilisant des données de recensement au niveau de l'îlot sur la race/ethnicité et sur la pauvreté. Enfin, à la section 4, nous examinons la pertinence de l'approche GDP et les extensions possibles de la méthodologie.

2. Calcul du plan de sondage optimal

2.1 Notation

Supposons que la population cible est divisée en cellules dénotées par les indices $b = 1, 2, ..., B$, avec les éléments de la cellule b dénotés par les indices $k = 1, 2, ..., K_b$. À chaque élément est associé un vecteur de covariables $\mathbf{x}_{bk}$ avec $\mathbf{x}'_{bk} = (\mathbf{u}'_{bk}, \mathbf{t}'_{bk})$, où $\mathbf{u}_{bk}$ est la composante observée pour les individus identifiables. La distribution de $\mathbf{t}_{bk}$ est connue dans chaque cellule, mais les valeurs individuelles ne sont pas observées; donc, la cellule est l'unité d'agrégation pour certaines dipour toutes les variables du plan. Par conséquent, nous connaissons les valeurs de population finie $\mathbf{T}_b = (\mathbf{t}_{b1}, \mathbf{t}_{b2}, ..., \mathbf{t}_{bK_b})'$, mais nous ne pouvons pas associer les lignes avec les individus. Définissons $\bar{\mathbf{t}}_b = 1'\mathbf{T}_b / K_b$, la moyenne de $\mathbf{t}$ dans la cellule b.

Un coût c_{bk} est associé à l'échantillonnage de chaque élément. Un plan d'échantillonnage est défini en attribuant une probabilité de sélection π_{bk} à chaque élément. Supposons une contrainte sur le coût attendu,

$$\sum_{b,\,k} c_{bk}\,\pi_{bk} \le C. \qquad (1)$$

Afin de simplifier l'exposé, nous supposons aussi que le taux d'échantillonnage est faible et que les avantages éventuels de la stratification sont minimes, si bien que le plan peut être décrit approximativement comme un échantillonnage avec probabilités inégales non stratifié avec remise. Nous supposons également qu'il s'agit d'un échantillonnage par élément à un seul degré. La population est $U = \{(b, k): b = 1, 2, ..., B; k = 1, 2, ..., K_b\}$ et un échantillon est $S \subset U$.

Le coefficient de régression par les moindres carrés ordinaires (MCO) descriptif de la population, qui correspond au modèle $y_{bk} = \boldsymbol{\beta}'\mathbf{x}_{bk} + \varepsilon_{bk}$ avec $\varepsilon_{bk} \sim [0, \sigma^2]$, est $\boldsymbol{\beta}_U = (\mathbf{X}'_U\,\mathbf{X}_U)^{-1}\mathbf{X}'_U\,\mathbf{y}_U$, où l'indice inférieur U indique que les matrices ou les vecteurs correspondent à l'ensemble de la population. (Ici $[0, \sigma^2]$ signifie une loi de moyenne 0 et de variance σ^2, mais de forme non spécifiée.) Alors

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'_S\,\mathbf{W}_S\,\mathbf{X}_S)^{-1}\mathbf{X}'_S\,\mathbf{W}_S\,\mathbf{y}_S \qquad (2)$$

est l'estimateur sous le plan de sondage habituel de $\boldsymbol{\beta}$, où S signifie que seules les lignes correspondant à l'échantillon sont incluses, et $\mathbf{W}$ est la matrice diagonale des poids $1/\pi_{bk}$.

Pour *concevoir* le sondage, nous devons émettre certaines hypothèses quant à la distribution des résultats y_{bk}, même si nous ne nous appuyons pas sur les mêmes hypothèses pour *analyser* les données. Plus précisément, nous supposons que les résultats sont générés par un modèle $\xi: y_{bk} = \mathbf{x}'_{bk}\boldsymbol{\beta} + \varepsilon_{bk}$ dont les $\varepsilon_{bk} \sim [0, \sigma^2_{bk}]$ sont indépendantes et les σ^2_{bk} sont connues (jusqu'à un facteur constant). Notons que les distributions des variables du plan d'échantillonnage $\mathbf{x}_{bk}$ et des résidus importent pour l'optimisation du plan de sondage, mais non la valeur de $\boldsymbol{\beta}$, parce qu'elle n'a pas d'incidence sur la variance des estimateurs par la régression. En outre, l'hypothèse d'indépendance des résidus d'un modèle de régression pourraient être plus raisonnable que celle de l'indépendance des valeurs de données. Nous tenons compte de l'hétéroscédasticité, même lorsque nous ajustons un modèle MCO. Les coefficients obtenus par les MCO (y compris les cas particuliers tels que la moyenne globale ou les moyennes de domaine) sont souvent des statistiques descriptives utiles, même si le modèle MCO ne tient effectivement pas; toutefois, si l'information au sujet de l'hétéroscédasticité est disponible, elle peut être utilisée pour rendre le plan de sondage plus efficace.

2.2 Plan de régression GDP optimal avec variables au niveau individuel uniquement

Considérons d'abord le cas où $\mathbf{t}$ est vide, de sorte que $\mathbf{x}_{bk} = \mathbf{u}_{bk}$, traduisant un scénario où le chercheur possède l'information sur toutes les variables du plan d'échantillonnage pertinentes (race et revenu dans notre plan CanCORS) avant l'échantillonnage. Puisque, dans ces conditions, les cellules correspondent à des cas uniques, nous laissons tomber l'indice inférieur b et écrivons $\hat{\boldsymbol{\beta}} = (\sum_S w_k\,\mathbf{x}_k\,\mathbf{x}'_k)^{-1}(\sum_S w_k\,\mathbf{x}_k\,y_k)$. Alors, pour toute combinaison linéaire fixe de coefficients avec les poids $\mathbf{a}$, en supposant que le premier facteur est un estimateur convergent sous le plan (après mise à l'échelle) de $N(\mathbf{X}'_U\,\mathbf{X}_U)^{-1}$, nous obtenons l'espérance sous échantillonnage de la variance fondée sur le modèle (White 1980) de l'estimateur,

$$\begin{aligned} V_a &= E_\pi \operatorname{Var}_\xi \mathbf{a}'\hat{\boldsymbol{\beta}} \\ &\approx \mathbf{a}'(\mathbf{X}'_U\mathbf{X}_U)^{-1}\left(E_\pi \operatorname{Var}_\xi \sum_{k\in S} \mathbf{x}_k y_k/\pi_k\right)(\mathbf{X}'_U\mathbf{X}_U)^{-1}\mathbf{a} \\ &= \mathbf{a}'(\mathbf{X}'_U\mathbf{X}_U)^{-1}\left(\sum_{k\in U}(\sigma^2_k/\pi_k)\,\mathbf{x}_k\,\mathbf{x}'_k\right)(\mathbf{X}'_U\mathbf{X}_U)^{-1}\mathbf{a}. \qquad (3) \end{aligned}$$

Dans le contexte de l'inférence fondée sur le plan de sondage, la mesure pertinente est la variance moyenne sous le plan d'échantillonnage sur toutes les populations possibles obtenues sous le modèle ξ, $E_\xi \mathrm{Var}_\pi \mathbf{a}'\hat{\boldsymbol{\beta}}$ (la « variance attendue » d'Isaki et Fuller 1982; voir aussi Bellhouse 1984, section 1); cette grandeur est approximativement égale à la variance sous un modèle prévue (voir l'annexe pour la preuve et les conditions asymptotiques).

En vertu de l'argument type du multiplicateur de Lagrange pour les problèmes de répartition optimale (par exemple, Valliant, Dorfman et Royall 2000, pages 169-170), V_a est minimisée sachant la contrainte de coût attendu (1) quand $\partial V_a / \partial \pi_k = c_k \lambda$ pour une constante λ et toutes les valeurs de k, de sorte que $\pi_k \propto \sigma_k (\mathbf{a}' \mathbf{X}_U' \mathbf{X}_U)^{-1} \mathbf{x}_k) / \sqrt{c_k}$. Donc, le taux d'échantillonnage optimal est plus élevé pour les cas où la variance sous le modèle est plus grande et le coût est plus faible (comme dans le cas habituel d'estimation d'une moyenne), ainsi que pour les cas dont l'effet levier est plus grand dans la régression. Ce résultat diffère des calculs fondés sur un modèle standard pour obtenir un plan expérimental optimal, qui donnerait la répartition de l'échantillon complet entre quelques points du plan à effet levier élevé. L'estimateur convergent sous le plan de la régression GDP ne s'appuie pas sur l'hypothèse que le modèle est correct et, par conséquent, requiert que chaque cas ait une probabilité de sélection positive. Donc, pour l'estimation d'un ratio β sous un modèle homoscédastique $y_k = \beta x_k + \varepsilon_k$, l'estimation fondée sur un modèle semblerait dicter la sélection des unités ayant les plus grande valeur de x, mais nos probabilités de sélection sont proportionnelles à x.

Habituellement, une étude a pour but d'estimer plus d'une grandeur; ainsi, l'étude CanCORS est conçue pour estimer les effets de la race ainsi que du revenu. Nous généralisons (3) à l'estimation simultanée de plusieurs combinaisons linéaires de coefficients en optimisant une somme pondérée des variances $V = \sum_i d_i V_{\mathbf{a}_i}$, où l'indice i désigne les grandeurs à estimer. Suivant les mêmes arguments, les probabilités d'échantillonnage optimales pour cet objectif sont

$$\pi_k \propto \sigma_k \left(\sum_i d_i (\mathbf{a}_i' (\mathbf{X}_U' \mathbf{X}_U)^{-1} \mathbf{x}_k)^2 / c_k \right)^{1/2}. \qquad (4)$$

Pour certains choix de $\{\mathbf{a}_i\}$, l'adhésion stricte à l'expression (4) pourrait produire des probabilités π_{bk} arbitrairement faibles (et, donc, des poids arbitrairement grands) pour les cas dont l'effet levier est presque nul. Afin d'éviter cette situation, nous pourrions imposer une valeur plancher positive à π_k. Ou bien, en incluant l'estimation de la moyenne de population parmi les objectifs (section 2.4), nous garantissons un échantillonnage avec probabilité positive sur l'ensemble de la population. L'une et l'autre méthode rend le plan plus robuste à l'erreur dans le calcul approximatif de l'effet levier et crée de meilleures conditions pour faire face à des décisions a posteriori éventuelles d'estimer des grandeurs qui n'ont pas été prévues dans le plan de sondage original (section 2.6). En outre, une estimation raisonnablement bonne des moyennes est nécessaire pour s'assurer de la convergence sous le plan du premier facteur de (2).

2.3 Plan optimal avec variables aux niveaux individuel et agrégé

Supposons maintenant que le vecteur de covariables $\mathbf{t}_{bk}$ n'est pas vide et que $\mathbf{u}_{bk}$ est constant dans chaque cellule, comme quand l'information agrégée sur le plan est disponible pour les cellules correspondant à des classes de covariables de $\mathbf{u}$ à l'intérieur des îlots. Dans l'étude CanCORS, si nous connaissons la race ($\mathbf{u}$), mais non le revenu ($\mathbf{t}$) pour les sujets pris individuellement et que nous connaissons la distribution du revenu pour chaque race dans chaque îlot de recensement, nous définirons des cellules constituées de personnes d'une seule race dans un seul îlot de recensement.

Puisqu'il est impossible de faire la distinction entre les cas appartenant à une même cellule d'après les covariables, nous supposons en outre que $\sigma_{bk} = \sigma_b$ et $c_{bk} = c_b$ sont a priori constants dans toute la cellule, de sorte que le plan optimal rend aussi $\pi_{bk} = \pi_b$ constante dans chaque cellule.

Nous pouvons maintenant réécrire (3) sous la forme

$$V_a \approx \mathbf{a}'(\mathbf{X}_U' \mathbf{X}_U)^{-1} \left(E_\pi \mathrm{Var}_\xi \sum_S \mathbf{x}_{bk} y_{bk} / \pi_b \right) (\mathbf{X}_U' \mathbf{X}_U)^{-1} \mathbf{a}$$

$$= \mathbf{a}'(\mathbf{X}_U' \mathbf{X}_U)^{-1} \left(\sum_{b,k} (\sigma_b^2 / \pi_b) \mathbf{S}_b \right) (\mathbf{X}_U' \mathbf{X}_U)^{-1} \mathbf{a}, \qquad (5)$$

où

$$\mathbf{S}_b = \begin{pmatrix} \mathbf{u}_b \mathbf{u}_b' & \mathbf{u}_b \bar{\mathbf{t}}_b' \\ \bar{\mathbf{t}}_b \mathbf{u}_b' & \mathbf{S}_{Tb} \end{pmatrix}$$

est la matrice des carrés moyens et des produits croisés dans la cellule b, avec $\mathbf{S}_{Tb} = \mathbf{T}_b' \mathbf{T}_b / K_b$. Les probabilités optimales d'échantillonnage correspondant à (4) sont alors

$$\pi_b \propto \sigma_b \left(\sum_i d_i \mathbf{a}_i' (\mathbf{X}_U' \mathbf{X}_U)^{-1} \mathbf{S}_b (\mathbf{X}_U' \mathbf{X}_U)^{-1} \mathbf{a}_i / K_b c_b \right)^{1/2}. \qquad (6)$$

Si $\mathbf{t}$ est mesuré par un recensement dans chaque cellule, alors $\bar{\mathbf{t}}_b$ et S_{Tb} sont connus exactement. Les mêmes principes s'appliquent toutefois si, au lieu d'être observé directement, S_{Tb} est estimé sous un modèle ζ. Nous remplaçons alors $\bar{\mathbf{t}}_b$ et $\mathbf{S}_b$ dans (5) par les espérances prédictives $\tilde{\bar{\mathbf{t}}}_b = E_\zeta \bar{\mathbf{t}}_b$ et $\tilde{\mathbf{S}}_b = E_\zeta \mathbf{S}_b$. Les situations qui suivent pourraient être des exemples : 1) pour chaque cellule, les

données sont disponibles uniquement pour un échantillon, 2) les données sur le plan d'échantillonnage sont anciennes et la distribution des variables du plan d'échantillonnage dans la cellule pourrait avoir changé au cours du temps ou 3) les données sur les éléments individuels sont mesurées avec une erreur. De même, la distribution de $\mathbf{t}$ pourrait n'être disponible que pour une supercellule qui contient plusieurs valeurs de $\mathbf{u}$ (par exemple, la race et l'îlot de recensement de résidence sont connus pour chaque individu, mais la distribution du revenu est connue pour l'îlot dans son ensemble, mais non pour chaque race dans l'îlot), si bien que $\bar{\mathbf{t}}_b$ et $\mathbf{S}_{Tb}$ doivent être estimés sous un modèle.

2.4 Plus d'un modèle

Dans l'exposé qui précède, nous avons supposé que toutes les grandeurs que l'on veut estimer sont des combinaisons des paramètres d'un seul modèle. De manière plus générale, les analyses envisagées pourraient comprendre l'ajustement de plusieurs modèles, et V pourrait représenter la somme des variances de diverses combinaisons de paramètres provenant de ces modèles. Un cas particulier évident est l'estimation d'une moyenne de population (comme il est suggéré à la section 2.2), le coefficient du modèle $y_k = \beta_0 \cdot 1 + \varepsilon_{bk}$, ainsi que certains coefficients de régression. Un autre exemple simple est l'estimation des moyennes de divers domaines définis, c'est-à-dire les coefficients de modèles de la forme $y_k = \boldsymbol{\beta}'_{(m)}\mathbf{x}_{(m)bk} + \varepsilon_{(m)bk}$, où $\mathbf{x}_{(m)bk}$ est un vecteur d'indicateurs d'appartenance à un domaine, les diverses définitions de domaine étant désignées par l'indice $m = 1, ..., M$, ou les contrastes de ces moyennes. Par exemple, nous pourrions souhaiter estimer les moyennes selon la race ainsi que selon l'âge.

Si chacune des combinaisons d'intérêt ne comprend que des paramètres en provenance d'un seul modèle, alors chaque combinaison possède sa propre matrice de plan, de sorte que l'indice de modèle m peut être identifié à l'indice de la grandeur à estimer i. Donc, dans (5) et (6), nous remplaçons $\mathbf{X}_U$ par $\mathbf{X}_{U(i)}$, ainsi que $\mathbf{S}_b$ par $\mathbf{S}_{(i)b}$.

Si certaines grandeurs à estimer combinent des paramètres provenant de modèles différents, nous empilons les estimateurs $\hat{\boldsymbol{\beta}}_{(m)}$ pour les divers modèles. Puis, dans (5) et (6), nous remplaçons $\mathbf{X}'_U\mathbf{X}_U$ par $\text{diag}(\mathbf{X}'_{U(m)}\mathbf{X}_{U(m)}, m = 1, ..., M)$ et nous redéfinissons $\mathbf{S}_b$ comme étant les sommes combinées de la matrice des carrés et des produits croisés pour l'ensemble des modèles, avec les blocs

$$\mathbf{S}_{b(m,\,m')} = \begin{pmatrix} \mathbf{u}_{b(m)}\mathbf{u}'_{b(m')} & \mathbf{u}_{b(m)}\bar{\mathbf{t}}'_{b(m')} \\ \bar{\mathbf{t}}_{b(m')}\mathbf{u}'_{b(m)} & \mathbf{S}_{Tb(m,\,m')} \end{pmatrix}.$$

La suite de l'optimisation est la même qu'à la section 2.3.

2.5 Poids de contraste flexibles

Dans l'étude CanCORS, nous voulons estimer l'effet du revenu en tenant compte de l'effet de la race et de l'effet moyen pour l'ensemble des races. La façon dont les races sont pondérées dans cette moyenne est moins importante pour nous, car les régions étudiées ne sont pas représentatives des proportions nationales selon la race. Nous pourrions alors estimer l'effet du revenu sur la pauvreté/non-pauvreté pour chaque race et combiner les résultats obtenus en utilisant des poids choisis de façon à minimiser la variance de l'estimateur de la moyenne pondérée des effets intra-race du revenu.

En général, nous examinons des situations dans lesquelles l'intérêt scientifique est d'estimer ou de tester *toute* combinaison $\mathbf{a}_i = \mathbf{A}_i\mathbf{f}_i$, où $\mathbf{A}_i$ est fixe et tous les $\mathbf{f}_i$ sont arbitraires (et ne sont pas nécessairement tous de la même dimension) sous la contrainte $1'\mathbf{f}_i = 1$, $f_{ij} \geq 0$. Dans notre exemple motivé par l'étude CanCORS, le modèle sous-jacent comprend huit variables indicatrices pour chacun des groupes définis par quatre groupes raciaux croisés avec un niveau de pauvreté dichotomique, et $\mathbf{A}_1$ est une matrice 8×4 dont chaque colonne contient une valeur 1 et une valeur -1 pour le contraste entre pauvre et non pauvre au sein d'une race. Alors, $\mathbf{f}_1$ contient les poids appliqués au contraste dans chaque race, et $\mathbf{a}'_1 = (f_{11}, -f_{11}, f_{12}, -f_{12}, f_{13}, -f_{13}, f_{14}, -f_{14})$ est le contraste pondéré des coefficients des huit indicateurs.

En substituant dans (5) et (6), nous optimisons sur les probabilités d'échantillonnage $\boldsymbol{\pi} = \{\pi_k\}$ ainsi que les poids de combinaison $\mathbf{f} = \{\mathbf{f}_i\}$. Dans le cas de modèles multiples, nous utilisons l'une ou l'autre des formules de la section 2.4, selon que les combinaisons d'intérêt comprennent les coefficients d'un ou de plusieurs modèles. La définition de $\mathbf{a}_i$ est donc déterminée en partie par des considérations scientifiques et en partie par l'information disponible provenant de la population étudiée.

Une approche naturelle de l'optimisation conjointe de $\boldsymbol{\pi}$ et $\mathbf{f}$ consiste à minimiser alternativement V par rapport à $\boldsymbol{\pi}$ en utilisant l'expression (6) modifiée et par rapport à $\mathbf{f}$, en observant les contraintes sur $\mathbf{f}$. Dans l'optimisation, $\mathbf{f}_i$ apparaît dans une expression de la forme $\mathbf{f}'_i\mathbf{D}_i(\pi)\mathbf{f}_i$. Par minimisation sous la contrainte $\mathbf{f}'_i \cdot \mathbf{1} = 1$ en utilisant les multiplicateurs de Lagrange, nous obtenons $\hat{\mathbf{f}}_i = \mathbf{D}_i^{-1}(\pi)\mathbf{1}/(\mathbf{1}'\mathbf{D}_i^{-1}(\pi)\mathbf{1})$ à condition que $\pi_{bk} > 0$ et que les contraintes de non-négativité ne soient pas obligatoires. Si elles le sont, des méthodes de programmation quadratique peuvent être utilisées.

2.6 Précision des analyses non prévues

Un plan d'échantillonnage destiné à être optimal pour un coefficient de régression pourrait être très inefficace pour un

autre, dans le même modèle ou dans des modèles différents. Faire de la moyenne de population l'une des grandeurs à estimer facilite le contrôle de ce risque. Nous illustrons ce point à l'aide d'un exemple où les variables de plan d'échantillonnage x_k, z_k, qui suivent une loi conjointe

$$\zeta\colon (X, Z) \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

sont entièrement observées pour les individus (dénotés par l'indice k à la section 2.2) et les modèles de régression constante et de régression univariée :

Modèle 0 : $y_k = \alpha_0 + \varepsilon_k^0,\ \varepsilon_k^0 \sim [0, \sigma_0^2]$

Modèle 1 : $y_k = \alpha_X + \beta_X x_k + \varepsilon_k^X,\ \varepsilon_k^X \sim [0, \sigma_X^2]$

Modèle 2 : $y_k = \alpha_Z + \beta_Z z_k + \varepsilon_k^Z,\ \varepsilon_k^Z \sim [0, \sigma_Z^2]$. Pour simplifier la notation, nous supposons que $\sigma_0^2 \approx \sigma_X^2 \approx \sigma_Z^2 \approx 1$, $\bar{x}_U = \bar{z}_U = 0$ et que les coûts c_k sont constants.

Considérons le plan d'échantillonnage optimisé pour $V = dV(\hat{\alpha}_0) + V(\hat{\beta}_X),\ d \geq 0$. En vertu de (4), le plan optimal a pour probabilités $\pi_k \propto \sqrt{d + x_k^2}$. Sous ce plan, la variance prévue est approximée par $V(\hat{\beta}_Z) \approx n^{-1}\sigma_Z^2 (Z_U' Z_U)^{-1} (Z_U' W_U Z_U)(Z_U' Z_U)^{-1}$ où $Z_U = (z_1, \ldots, z_N)'$ et $W_U = \text{diag}(\pi_1^{-1}, \ldots, \pi_N^{-1})$. Alors, $E_\zeta nV(\hat{\beta}_Z) \approx c_0 E_\zeta (Z^2/\sqrt{d + X^2})$, où c_0 dépend uniquement de d, de sorte que $E_\zeta nV(\hat{\beta}_Z)$ dépend de ρ et d. Si $d = 0$ (aucun poids n'est lié à l'estimation de la moyenne), $E_\zeta nV(\hat{\beta}_Z) = \infty$, à moins que $\rho = \pm 1$. Donc, sauf si l'objectif accorde un certain poids à la variance de l'estimateur de la moyenne, le plan pourrait être très médiocre pour les coefficients liés aux covariables qui ne se trouvent pas dans l'étendue des variables des modèles optimisés. Par contre, si $d > 0$, nous pouvons décomposer Z en composantes parallèles et perpendiculaires à X, $Z = Z_1 + Z_2$ où $Z_1 = \rho X$ et $Z_2 = Z - \rho X$, de sorte que $Z_1 \perp Z_2$, $Z_2 \perp X$ et $E_\zeta Z_2 = 0$. Alors,

$$E_\zeta nV(\hat{\beta}_Z) = \rho^2 c_0 E_\zeta (X^2/\sqrt{d + X^2}) + (1 - \rho^2) c_0 E_\zeta (1/\sqrt{d + X^2})$$
$$= \rho^2 E_\zeta V_{\text{opt}}(\hat{\beta}_X) + (1 - \rho^2) E_\zeta V_{\text{opt}}(\hat{\alpha}_0).$$

Exprimé en mots, la variance du coefficient du nouveau modèle est une combinaison des deux variances qui étaient contrôlées dans l'optimisation. Cela donne à penser qu'un plan d'échantillonnage qui comprend l'estimation de la moyenne globale de population dans l'optimisation offre une certaine protection contre l'inefficacité extrême pour d'autres modèles contenant des variables qui n'ont pas été prises en compte dans le plan original, quoique les résultats simples présentés ici ne se généralisent pas nécessairement de manière à couvrir chaque cas.

3. Application : Régressions sur la race et l'état de pauvreté

3.1 Description des emplacements et des données

Le projet CanCORS (Ayanian et coll. 2004) comprend cinq emplacements définis géographiquement (Nord de la Californie, Los Angeles, Alabama/Georgie, Caroline du Nord et Iowa) et deux emplacements définis sur une base organisationnelle. Le Nord de la Californie comprend neuf comtés qui s'étendent de la région de la baie de San Francisco au comté semi-rural Placer County sur la frontière avec le Nevada. Cet emplacement est diversifié ethniquement et géographiquement et, par conséquent, convient bien pour illustrer les méthodes. Nous décrivons les résultats pour cet emplacement en détail, puis nous résumons les résultats pour les autres.

Nos données sont fondées sur l'échantillon qui a reçu le « questionnaire détaillé » du Recensement décennal des États-Unis de 2000 et ont été extraites pour les neuf comtés de notre région cible (Alameda, Contra Costa, Placer, Sacramento, San Francisco, San Joaquin, San Mateo, Santa Clara et Solano) de SF-3, tableaux 159a à 159i, « Poverty Status in 1999 by Age ». Nous avons produit des tableaux croisés pour les résidents d'au moins 65 ans échantillonnés de chaque groupe d'îlots de recensement (une petite région continue à peu près équivalente à plusieurs îlots urbains, appelée ci-après un îlot) selon la race/ethnicité et le revenu, en utilisant les poids d'échantillonnage du recensement. La contrainte d'âge correspond approximativement aux âges pour lesquels sont relevés le plus de nouveaux cas de cancer admissibles pour l'étude. Le revenu du ménage a été dichotomisé comme étant supérieur ou inférieur au seuil de pauvreté standard. Le recensement comportait des questions distinctes sur l'ethnicité hispanique et la race; nous avons classifié la population comme étant hispanique ou blanche non hispanique, noire, ou asiatico-américaine. Une catégorie hétérogène « Autre » regroupe les 3 % restant de la population de personnes âgées. (Par souci de concision, dans la suite de l'exposé, nous parlerons de « catégorie raciale ».) La région étudiée comptait 844 560 personnes de plus de 65 ans réparties dans 5 098 groupes d'îlots, soit une moyenne de 166 par groupe d'îlots.

Le tableau 1 résume la répartition selon la race et le revenu dans le Nord de la Californie. Les Noirs sont les individus dont le taux global de pauvreté est le plus élevé et chez lesquels la ségrégation est la plus forte (coefficient de variation le plus élevé du pourcentage de Noirs par îlot), résultats qui confirment les profils nationaux de ségrégation résidentielle (Denton et Massey 1993). Les Hispaniques sont ceux pour lesquels la variation géographique relative du taux de pauvreté est la plus prononcée (coefficient de variation le plus élevé du taux de pauvreté par îlot).

Tableau 1
Distributions des races et de la pauvreté chez les personnes de 65 ans et plus, selon le groupe d'îlots de recensement dans le Nord de la Californie (CV = coefficient de variation)

	Blancs	Noirs	Asiatiques	Hispaniques	Autre	Total
Pourcentage de la population	65,70	6,40	16,80	8,20	3,00	100,00
Pourcentage de pauvres	5,20	14,20	10,10	10,60	11,30	7,20
CV pourcentage par îlot de la race	0,46	2,94	1,21	1,73	2,28	-
CV pourcentage par îlot de pauvres	1,53	1,37	1,58	1,89	2,30	1,16

3.2 Contraintes du plan d'échantillonnage : information disponible et objectifs du plan

Nous avons calculé l'efficacité, comparativement à l'échantillonnage aléatoire simple (EAS), du plan optimal pour des scénarios définis par deux contraintes, à savoir 1) le choix de la fonction objectif et 2) les hypothèses au sujet de l'information disponible pour déterminer les probabilités d'échantillonnage.

Nous avons considéré six hypothèses possibles au sujet de l'information disponible pour la race (non disponible, ou disponible au niveau individuel) et pour le revenu (non disponible, disponible uniquement par îlot, ou disponible au niveau individuel). Comme la race est plus souvent consignée dans les dossiers d'hospitalisation que le revenu, nous avons exclu le cas où la catégorie de revenu est connue au niveau individuel, mais la race n'est connue que par groupe d'îlots. Chaque hypothèse correspond à une définition de la cellule utilisée pour l'élaboration de la section 2.3 et à une définition correspondante des variables $\mathbf{t}$ et $\mathbf{u}$:

1. Aucune information sur les variables du plan d'échantillonnage n'est disponible : la cellule correspond à l'ensemble de la population et $\mathbf{u}$ inclut la race et le revenu. (Colonnes intitulées « EAS » dans le tableau 2.)
2. Race uniquement : la cellule est une catégorie raciale, $\mathbf{u}$ contient les variables de race, et $\mathbf{t}$ est le revenu. (Colonnes intitulées « Race ».)
3. Données agrégées pour l'îlot uniquement : la cellule est un groupe d'îlots de recensement, $\mathbf{u}$ est vide et $\mathbf{t}$ comprend la race et le revenu. (Colonnes intitulées « Îlot ».)
4. Données individuelles sur la race, données agrégées au niveau de l'îlot sur le revenu selon la race : la cellule correspond à la population d'une race dans un groupe d'îlots, $\mathbf{u}$ est la race et $\mathbf{t}$ est le revenu. (Colonnes intitulées « Race+Îlot ».)
5. Données individuelles sur le revenu, aucune donnée sur la race : la cellule est une catégorie de revenu, $\mathbf{u}$ est le revenu et $\mathbf{t}$ est vide. (Colonnes intitulées « Revenu ».)
6. Données sur la race et sur le revenu disponibles pour chaque individu : la cellule est une catégorie raciale selon le revenu, $\mathbf{u}$ comprend la race et le revenu et $\mathbf{t}$ est vide. (Colonnes intitulées « Race+ Revenu ».)

Nous avons calculé les taux d'échantillonnage optimaux sous chaque hypothèse au sujet de l'information disponible, en utilisant diverses fonctions objectif. Dans chacune de ces fonctions, nous avons pondéré ensemble les variances des estimations des coefficients dans certains de quatre modèles de régression ou tous : 1) le modèle à « terme constant uniquement » dont le seul paramètre est la moyenne de population, 2) un modèle de race paramétrisé comme une moyenne pour les Blancs et des contrastes pour les différences entre les Blancs et chacun des autres grands groupes raciaux (Noirs, Hispaniques et Asiatiques), 3) un modèle de revenu paramétrisé comme une moyenne pour les non-pauvres et un contraste entre les pauvres et les non-pauvres, et 4) un modèle additif conjoint comprenant des effets de race et de revenu. Chaque fonction objectif comprend la pondération $d_{\text{moyenne}} > 0$, qui garantit que toutes les probabilités $\pi_{bk} > 0$, ce qui évite des problèmes numériques dans l'optimisation. Donc, au moins deux modèles sont représentés dans chaque fonction objectif (section 2.4). Quand la fonction objectif pondère à la fois les effets de revenu et de race, l'effet de revenu unique reçoit la pondération $d_{\text{revenu}} = 3$ pour qu'elle concorde avec les trois effets de race ayant une pondération de 1.

Nous avons étudié un choix de poids objectifs qui mettent l'accent sur l'estimation des effets de race, des effets de revenu ou des deux. Chaque partie du tableau 2 représente un choix particulier de poids objectifs d_i (troisième colonne) pour les coefficients des contrastes $\mathbf{a}_i$ (deuxième colonne) d'une série de modèles (premières colonnes). La quatrième colonne donne la variance (normalisée pour une taille d'échantillon unitaire) $nV_{\mathbf{a}_i}$ pour l'estimation de ce coefficient sous EAS en supposant que la variance résiduelle est $\sigma^2 = 1$. Dans les autres colonnes, nous présentons les effets de plan, les ratios de la variance normalisée $nV_{\mathbf{a}_i}$ pour le plan optimisé sous diverses hypothèses au sujet de l'information disponible sur les variables du plan par rapport à la variance sous EAS. Les lignes dont le poids objectif est $d_i = 0$ n'ont pas d'incidence sur l'optimisation, mais sont incluses pour illustrer l'effet de chaque plan d'échantillonnage sur l'efficacité de l'estimation d'un coefficient qui n'est pas inclus dans la fonction objectif. La dernière ligne résume l'effet de plan pondéré correspondant à la fonction de perte, c'est-à-dire la combinaison pondérée des variances.

3.3 Efficacité sous modèles fixes

Les deux premières fonctions objectives optimisent l'estimation des contrastes de race et de la moyenne globale. L'utilisation de données individuelles sur la race améliore considérablement l'efficacité de l'estimation des effets des races noires et hispaniques. Les améliorations les plus importantes sont observées pour l'effet de la race noire (la plus petite des trois grandes minorités raciales, dont la variance est réduite à 43 % de sa valeur sous EAS. Inversement, nous n'obtenons aucune amélioration pour les Asiatico-américains, dont la représentation dans la population est proche du taux d'échantillonnage optimal. Dans cette fonction objectif, une fois que l'information sur la race est disponible, des renseignements supplémentaires sur les variables du plan (revenu au niveau de l'îlot ou de l'individu) n'a aucune importance pour l'optimisation. Si la race au niveau individuel est inconnue, l'utilisation de l'îlot de résidence peut faciliter le suréchantillonnage des Noirs (le groupe pour lequel la ségrégation résidentielle est la plus importante), ce qui réduit la variance de l'effet estimé de la race noire à environ 65 % de sa valeur sous EAS, mais le suréchantillonnage par îlot uniquement ne réduit que légèrement la variance de l'effet estimé de la race hispanique. Connaître le revenu est, en soi, de peu d'utilité en ce qui concerne l'amélioration de l'échantillonnage pour l'estimation des effets de race.

L'échantillonnage non proportionnel, ajusté pour optimiser l'estimation des effets de la race, accroît les variances des autres estimateurs de paramètre. Si l'on applique un poids minimal à la moyenne dans l'objectif d'optimisation (objectif 1), cette augmentation peut être assez importante, soit un facteur de 181 % pour la moyenne et les effets du revenu. Accorder plus de poids à la moyenne (objectif 2) modère cet effet, réduisant l'accroissement de la variance à 136 % pour la moyenne et 121 % pour l'effet du revenu, tout en n'accroissant que légèrement les variances pour les effets de race.

La variance normalisée minimale possible pour l'estimation de l'effet du revenu (objectif 3) est de 4 (27 % de la variance sous EAS), valeur obtenue quand le revenu est connu au niveau individuel sous un plan d'échantillonnage dans lequel l'échantillon est réparti de manière égale entre les pauvres et les non-pauvres. Dans le cas d'information au niveau de l'îlot, la variance peut être réduite à 74 % de sa valeur sous EAS. Alors que ne connaître que la race offre peu d'avantages pour cet objectif, l'ajout de données individuelles sur la race à l'information au niveau de l'îlot donne une réduction supplémentaire de la variance de l'effet estimé du revenu, laquelle atteint 60 % de sa valeur sous EAS. Les variances des estimations de la moyenne et des effets de race augmentent considérablement sous ces plans d'échantillonnage, mais l'accroissement de la pondération de la moyenne (objectif 4) donne des résultats nettement meilleurs en ce qui concerne l'amélioration de la variance de la moyenne et des effets de race, et n'accroît que légèrement la variance de l'effet estimé du revenu.

L'introduction des effets de la race ainsi que du revenu dans l'objectif 5 produit des plans d'échantillonnage qui ne sont pas aussi bons que les plans optimaux pour les uns et les autres de ces effets pris individuellement, mais malgré tout nettement meilleurs que sous EAS. Par exemple, les variances des effets de race lorsque l'on connaît la race et l'îlot de résidence sont de 10 % à 24 % plus élevées que sous les plans d'échantillonnage utilisant la même information sur les variables du plan, mais optimisés séparément pour la race ou le revenu. Si l'on connaît uniquement l'information individuelle sur la race ou sur le revenu, le plan est essentiellement optimisé pour les effets correspondant à la variable disponible, en accroissant la variance des effets estimés de l'autre variable.

Le plan d'échantillonnage optimisé pour les effets conjugués de la race et du revenu dans le modèle additif à deux facteurs (objectif 6) donne des résultats assez approchés de l'optimisation pour les effets de race et de revenu dans des modèles marginaux distincts (objectif 5). Quand nous optimisons des effets distincts, les variances de ces derniers sont un peu plus faibles que celles obtenues pour les effets correspondants dans le modèle à deux facteurs. Quand nous optimisons les effets dans le modèle conjoint, leurs variances sont réduites, mais, dans la plupart des cas, demeurent un peu plus élevées que celles des effets correspondants dans les modèles marginaux de race et de revenu, à cause de la confusion partielle des effets de race et de revenu.

De même, l'optimisation de l'effet du revenu dans le modèle à deux facteurs (objectif 7) est assez semblable à l'optimisation de l'effet univarié du revenu (objectif 4) lorsqu'on ne dispose d'aucune donnée sur la race. Cependant, la disponibilité de données sur la race, ainsi que sur le revenu au niveau de l'îlot ou de l'individu réduit considérablement les variances des effets de race sous le plan d'échantillonnage pour le modèle à deux facteurs. À cause de la confusion partielle des effets de la race et du revenu sous ce modèle, ce plan d'échantillonnage est adapté afin d'estimer les premiers plus efficacement, en accumulant plus de données aux points du plan d'échantillonnage qui sont essentiels à l'élimination de la confusion entre ces effets.

Tableau 2
Variances normalisées et fonctions objectif pour les plans d'échantillonnage optimaux pour divers poids objectifs et hypothèses concernant l'information disponible sur les variables du plan

Objectif 1 : Optimisation pour les effets de race

				Variance en pourcentage de la variance sous EAS (selon l'information disponible sur les variables du plan)				
Modèle	Effet	Pondération (d_i)	Variance sous EAS	Race	Îlot	Race+ Îlot	Revenu	Race+ Revenu
Constante	Moyenne	0,1	1,0	181	119	181	100	181
Race	Noirs	1	17,2	43	65	43	99	43
	Asiatiques	1	7,5	100	106	100	100	100
	Hispaniques	1	13,7	55	90	55	100	55
Revenu	Pauvres	0	15,0	182	104	182	81	182
Race+Revenu	Noirs	0	17,4	44	65	44	99	44
	Asiatiques	0	7,5	100	106	100	100	100
	Hispaniques	0	13,8	55	90	55	100	55
	Pauvres	0	15,2	182	104	182	81	182
Total = $nV = n\sum d_i V_{a_i}$			38,6	59	82	59	99	59

Objectif 2 : Optimisation pour les effets de race et la moyenne globale

				Variance en pourcentage de la variance sous EAS (selon l'information disponible sur les variables du plan)				
Modèle	Effet	Pondération (d_i)	Variance sous EAS	Race	Îlot	Race+ Îlot	Revenu	Race+ Revenu
Constante	Moyenne	3	1,0	136	115	136	100	136
Race	Noirs	1	17,2	44	66	44	99	44
	Asiatiques	1	7,5	100	104	100	100	100
	Hispaniques	1	13,7	56	90	56	100	56
Revenu	Pauvres	0	15,0	121	101	121	82	121
Race+Revenu	Noirs	0	17,4	45	66	45	99	45
	Asiatiques	0	7,5	100	104	100	100	100
	Hispaniques	0	13,8	56	90	56	100	56
	Pauvres	0	15,2	122	102	122	82	122
Total = $nV = n\sum d_i V_{a_i}$			41,5	65	84	65	100	65

Objectif 3 : Optimisation pour les effets du revenu

				Variance en pourcentage de la variance sous EAS (selon l'information disponible sur les variables du plan)				
Modèle	Effet	Pondération (d_i)	Variance sous EAS	Race	Îlot	Race+ Îlot	Revenu	Race+ Revenu
Constante	Moyenne	0,001	1,0	103	154	173	173	173
Race	Noirs	0	17,2	75	119	152	163	163
	Asiatiques	0	7,5	90	144	173	170	170
	Hispaniques	0	13,7	86	142	196	168	168
Revenu	Pauvres	0	15,0	97	74	60	27	27
Race+Revenu	Noirs	0	17,4	75	119	153	164	164
	Asiatiques	0	7,5	90	144	174	171	171
	Hispaniques	0	13,8	86	143	197	169	169
	Pauvres	0	15,2	97	75	63	29	29
Total = $nV = n\sum d_i V_{a_i}$			45,0	97	74	60	27	27

Objectif 4 : Optimisation pour les effets du revenu et la moyenne globale

				Variance en pourcentage de la variance sous EAS (selon l'information disponible sur les variables du plan)				
Modèle	Effet	Pondération (d_i)	Variance sous EAS	Race	Îlot	Race+ Îlot	Revenu	Race+ Revenu
Constante	Moyenne	3	1,0	103	134	147	151	151
Race	Noirs	0	17,2	76	107	128	142	142
	Asiatiques	0	7,5	91	127	145	148	148
	Hispaniques	0	13,7	86	125	161	147	147
Revenu	Pauvres	3	15,0	97	75	61	27	27
Race+Revenu	Noirs	0	17,4	76	107	129	143	143
	Asiatiques	0	7,5	91	127	146	149	149
	Hispaniques	0	13,8	86	125	162	147	147
	Pauvres	0	15,2	97	75	63	29	29
Total = $nV = n\sum d_i V_{a_i}$			48,0	97	79	66	35	35

Tableau 2 (suite)
Variances normalisées et fonctions objectif pour les plans d'échantillonnage optimaux pour divers poids objectifs et hypothèses concernant l'information disponible sur les variables du plan

Objectif 5 : Optimisation pour des effets de race distincts, l'effet du revenu et la moyenne globale

				Variance en pourcentage de la variance sous EAS (selon l'information disponible sur les variables du plan)				
Modèle	Effet	Pondération (d_i)	Variance sous EAS	Race	Îlot	Race+ Îlot	Revenu	Race+ Revenu
Constante	Moyenne	3	1,0	111	117	135	114	150
Race	Noirs	1	17,2	54	74	55	109	53
	Asiatiques	1	7,5	95	106	109	112	116
	Hispaniques	1	13,7	67	96	69	112	67
Revenu	Pauvres	3	15,0	101	82	72	38	37
Race+Revenu	Noirs	0	17,4	55	74	55	109	52
	Asiatiques	0	7,5	95	106	109	113	115
	Hispaniques	0	13,8	67	96	69	112	66
	Pauvres	0	15,2	101	82	72	39	35
Total = $nV = n\sum d_i V_{a_i}$		0	86,4	86	86	74	73	56

Objectif 6 : Optimisation pour les effets de race et l'effet du revenu dans un modèle à deux facteurs et pour la moyenne globale

				Variance en pourcentage de la variance sous EAS (selon l'information disponible sur les variables du plan)				
Modèle	Effet	Pondération (d_i)	Variance sous EAS	Race	Îlot	Race+ Îlot	Revenu	Race+ Revenu
Constante	Moyenne	3	1,0	111	119	138	114	156
Race	Noirs	0	17,2	55	74	55	108	53
	Asiatiques	0	7,5	95	107	109	112	114
	Hispaniques	0	13,7	67	96	69	111	67
Revenu	Pauvres	0	15,0	101	82	72	38	37
Race+Revenu	Noirs	1	17,4	55	74	55	109	52
	Asiatiques	1	7,5	95	107	109	113	113
	Hispaniques	1	13,8	67	96	69	112	66
	Pauvres	3	15,2	101	81	71	39	35
Total = $nV = n\sum d_i V_{a_i}$			87,2	86	86	73	73	54

Objectif 7 : Optimisation pour l'effet du revenu dans un modèle à deux facteurs et pour la moyenne globale

				Variance en pourcentage de la variance sous EAS (selon l'information disponible sur les variables du plan)				
Modèle	Effet	Pondération (d_i)	Variance sous EAS	Race	Îlot	Race+ Îlot	Revenu	Race+ Revenu
Constante	Moyenne	3	1,0	103	135	149	148	156
Race	Noirs	0	17,2	77	100	98	139	97
	Asiatiques	0	7,5	91	124	132	145	132
	Hispaniques	0	13,7	86	122	135	144	122
Revenu	Pauvres	0	15,0	97	75	62	28	28
Race+Revenu	Noirs	0	17,4	77	100	98	140	96
	Asiatiques	0	7,5	91	124	132	146	132
	Hispaniques	0	13,8	86	122	135	144	121
	Pauvres	3	15,2	97	75	62	29	27
Total = $nV = n\sum d_i V_{a_i}$			48,6	97	79	67	36	35

3.4 Efficacité sous des poids de contraste flexibles

Nous examinons maintenant les avantages éventuels de l'estimation des effets du revenu sous un scénario de pondération flexible (tableau 3). La fonction objectif prend en considération les coefficients de deux modèles, le modèle constant dont le paramètre est la moyenne de population et un modèle contenant des variables indicatrices pour chaque cellule race par revenu. Nous estimons l'effet du revenu à l'intérieur de chaque race comme étant la différence entre les coefficients pour les pauvres et les non-pauvres à l'intérieur de la race, puis nous combinons ces estimations avec des poids flexibles pour estimer un effet global du revenu. Cette stratégie s'approche plus de l'objectif 7, qui comporte aussi l'estimation des effets du revenu en tenant compte de l'effet de la race. L'analyse des contrastes souples dépend moins du modèle que l'option du modèle à deux facteurs en ce sens qu'elle ne s'appuie pas sur

l'hypothèse d'additivité de ce modèle. Par ailleurs, la façon dont les races sont combinées ne reflète pas nécessairement les proportions dans la population. Les poids appliqués au contraste du revenu dans chaque race, estimés comme nous l'avons décrit à la section 2.5, sont présentés à la partie inférieure du tableau 3 pour démontrer comment cette approche nous permet de modifier la grandeur estimée afin d'exploiter l'information disponible sur les variables du plan. (L'algorithme d'optimisation alternée a convergé vers une précision adéquate en sept itérations.)

Sous EAS, la variance de l'effet du revenu sous le modèle à pondération flexible est un peu plus grande que sous le modèle à deux facteurs (15,91 contre 14,99). Le poids accordé au contraste de la race blanche sous ce plan (51 %) est plus faible que la proportion de Blancs dans la population (66 %), parce que relativement peu de Blancs sont pauvres et que, par conséquent, le contraste du revenu parmi les Blancs est relativement imprécis. Inversement, le poids appliqué au contraste du revenu chez les Noirs (12 %) est presque le double de la part de ce groupe dans la population, à cause du taux disproportionnellement élevé de pauvreté dans ce groupe.

L'utilisation de données sur la race de niveau individuel dans le plan d'échantillonnage accentue cette disproportion, une plus grande part de l'échantillon et une plus forte pondération (75 %) étant données au groupe des Noirs, avec le pourcentage le plus élevé de pauvreté. Donc, la pondération flexible permet de réduire considérablement la variance de l'effet estimé du revenu (à 63 % de sa valeur sous EAS) en utilisant uniquement la race, ce qui était impossible sous le modèle GDP à deux facteurs plus restrictif.

L'information au niveau de l'îlot est un peu moins utile pour ce plan d'échantillonnage que celle sur la race. La combinaison de l'information sur l'îlot et sur la race est toutefois très puissante, réduisant la variance de l'effet du revenu à 48 % de sa valeur sous EAS. Sous ce plan d'échantillonnage, beaucoup plus de poids (46 %) est accordé au contraste du revenu des Hispaniques, qui peut être estimé efficacement à cause de la ségrégation plus forte du revenu chez ce groupe (tableau 1). Si l'on dispose de données individuelles sur le revenu (avec ou sans la race), les poids des contrastes correspondent approximativement aux proportions par race, puisque des contrastes de revenu efficaces peuvent être obtenus pour toute race et que l'inclusion de la moyenne globale dans l'objectif déplace le plan d'échantillonnage vers l'échantillonnage proportionnel. Donc, le plan d'échantillonnage diffère radicalement sous diverses hypothèses concernant la disponibilité de l'information sur les variables du plan.

3.5 Comparaisons entre les emplacements

Au tableau 4, nous comparons les améliorations pour l'échantillonnage non proportionnel à quatre emplacements de l'étude CanCORS, étant exclus les emplacements non géographiques et un emplacement (Iowa) dont la population était presque entièrement blanche. Pour chaque emplacement, nous avons procédé à l'optimisation, pour la pondération unitaire ($d_i = 1$), des variances de la moyenne globale et de l'effet du revenu dans le modèle à deux facteurs (proportionnel à l'objectif 7), sous diverses hypothèses concernant l'information disponible sur les variables du plan d'échantillonnage. Le minimum théorique pour cet objectif avec une population équilibrée est 5 ($V_{\text{moyenne}} = 1$, $V_{\text{revenu}} = 4$). L'EAS est inefficace à chaque emplacement, surtout en Alabama et dans le Nord de la Californie, et l'information sur la race serait, à elle seule, de peu d'utilité. Inversement, la meilleure variance qui peut être obtenue en utilisant l'information complète sur la race et sur le revenu au niveau individuel est comprise entre 5,60 et 5,72 à chaque emplacement. Le suréchantillonnage fondé sur l'information sur le revenu au niveau de l'îlot réduirait considérablement les variances, les progrès étant nettement plus importants en Alabama et dans le Nord de la Californie que pour les autres emplacements.

4. Discussion

Afin d'élaborer divers plans d'échantillonnage pour une étude sur les services de santé, nous avons étendu les méthodes antérieures d'élaboration de plans optimaux pour l'estimation par domaine afin de montrer comment un plan d'échantillonnage avec probabilités inégales optimal doit être conçu pour estimer les coefficients de régression d'un ou de plusieurs modèles. Dans notre application, nous avons pu réaliser des réductions importantes de la variance même quand l'information sur certaines variables n'était disponible que pour des agrégats géographiques. Nous avons obtenu des améliorations particulièrement importantes pour les variables explicatives catégoriques (état de pauvreté, race) présentant des distributions très asymétriques.

Essentiellement, notre approche d'établissement du plan d'échantillonnage lorsque les variables du plan sont mesurées de manière imprécise s'appuie sur la distribution prédictive des variables du plan pour chaque unité échantillonnée, plus précisément les espérances des variables, ainsi que leur carré et leur produit croisé. Ce concept unit le plan d'échantillonnage utilisant des agrégats de cellules (estimés d'après des données de recensement ou d'échantillon), utilisant des variables mesurées avec erreur, ou utilisant une base de sondage dont les caractéristiques des unités pourraient avoir changé au cours du temps.

Tableau 3
Variances normalisées et poids des contrastes pour des plans d'échantillonnage GDP optimaux avec pondération flexible des contrastes du revenu selon la race. Des lignes pour les contrastes fixes sont incluses afin de démontrer l'effet de divers choix de pondération flexible, en vue de la comparaison à des scénarios d'objectif avec pondération fixe.

Variances pour l'estimation « sous pondération flexible » et pour les contrastes présentés au tableau 2

				Variance en pourcentage de la variance sous EAS (selon l'information disponible sur les variables du plan)				
Modèle	Effet	Pondération (d_i)	Variance sous EAS	Race	Îlot	Race+ Îlot	Revenu	Race+ Revenu
Constante	Moyenne	3	1,0	233	139	206	152	152
Contraste flexible	Revenu	3	15,9	63	74	48	28	26
Race	Noirs	0	17,2	34	90	61	143	139
	Asiatiques	0	7,5	197	124	172	149	146
	Hispaniques	0	13,7	172	124	61	148	144
Revenu	Pauvres	0	15,0	209	77	124	27	39
Race+Revenu	Noirs	0	17,4	35	90	61	144	139
	Asiatiques	0	7,5	197	124	173	150	147
	Hispaniques	0	13,8	172	124	61	148	144
	Pauvres	0	15,2	210	77	125	29	39
Total = $\sum d_i V_{a_i}$			50,7	73	78	57	35	33

Poids optimaux (en pourcentage, $100\,\% \times f_i$) pour chaque contraste du revenu à l'intérieur de la race dans le calcul de l'estimation combinée de l'effet du revenu, sous chaque hypothèse concernant l'information disponible sur les variables du plan (Les totaux des colonnes pourraient ne pas être égaux à 100 % à cause de l'erreur d'arrondissement.)

	Hypothèses concernant l'information sur les variables du plan					
Contraste	EAS	Race	Îlot	Race+Îlot	Revenu	Race+Revenu
Noirs	12	75	17	25	9	6
Asiatiques	24	9	26	16	21	17
Hispaniques	12	5	13	46	11	8
Blancs	51	12	45	13	59	68

Tableau 4
Fonction objectif normalisée pour les plans GDP optimaux sous pondération égale ($d_i = 1$) des variances des moyennes globales et de l'effet du revenu dans le modèle à deux facteurs, à quatre emplacements de l'étude CanCORS

	Variance sous	Variance en pourcentage de la variance sous EAS				
Emplacement	EAS	Race	Îlot	Race+Îlot	Revenu	Race+Revenu
Alabama	16,2	97	79	67	36	35
Los Angeles	11,8	98	85	76	49	47
Caroline du Nord	10,2	97	89	86	59	55
Nord de la Californie	16,2	97	79	67	36	35

Les méthodes décrites ici pour optimiser les probabilités d'échantillonnage par élément peuvent être combinées à la stratification et à la mise en grappes ou à l'échantillonnage à plusieurs degrés. (Aucune de ces caractéristiques d'échantillonnage ne figure dans l'étude CanCORS qui a motivé nos travaux de recherche. La stratification n'était pas commode étant donné l'identification séquentielle des sujets et le fait que l'on disposait de peu d'information a priori pour orienter la construction de strates homogènes. L'interview par téléphone a rendu opérationnellement inutile la mise en grappes de nos sujets.) Comme ces caractéristiques du plan d'échantillonnage peuvent avoir une incidence sur les distribution d'échantillonnage des variables du plan d'échantillonnage ainsi que des variables de résultat, et que les objectifs du plan d'échantillonnage englobaient à la fois le modèle de population postulé et le modèle scientifique d'intérêt, le nombre de combinaisons possibles est encore plus grand que dans le cas d'un plan d'échantillonnage pour l'estimation d'une moyenne de population. Par conséquent, nous nous limitons à suggérer quelques idées qui pourraient être poursuivies dans le cadre de futurs travaux.

La stratification peut améliorer un plan d'échantillonnage pour une analyse par régression d'au moins trois façons : 1) pour mettre en œuvre un échantillonnage non proportionnel (en utilisant des probabilités égales proches de celles calculées selon notre méthode), 2) pour contrôler la distribution des variables du plan d'échantillonnage afin qu'elles soient plus proches du plan d'échantillonnage optimal que dans le cas d'un plan avec probabilités inégales non stratifiées et 3) pour réduire la variation intra-strate de la

statistique d'influence des cas et, par conséquent, réduire la variance des estimations des coefficients (Fuller, 1975). Puisque l'efficacité du plan est insensible à de faibles écarts par rapport à la situation optimale, certains plans stratifiés avec probabilités égales dans les strates pourraient approcher l'efficacité du plan optimal. Des stratifications *ponctuelles* pourraient être moins efficaces, même sous répartition optimale entre les strates. Par exemple, la stratification des îlots en fonction du groupe race-revenu représenté dont la prévalence est la plus faible produit un plan d'échantillonnage dont le gain d'efficacité est égal à environ la moitié de celui de notre plan d'échantillonnage avec utilisation de données agrégées au niveau de l'îlot.

En ce qui concerne le dernier point, il convient de souligner que la conception de strates homogènes pour l'estimation des coefficients de régression est vraisemblablement plus difficile que pour l'estimation d'une moyenne. L'influence d'une observation dépend du résidu de cette dernière produit par le modèle de régression et non de sa valeur brute, de sorte que pour réduire l'homogénéité, la stratification devrait comprendre des variables prédictives qui ne sont pas incluses dans le modèle. L'influence dépend aussi de l'effet levier de l'observation pour chaque coefficient, fait qui pourrait être une fonction complexe des covariables.

Dans le cas de l'échantillonnage par grappe, l'équivalence de $E_\pi \mathrm{Var}_\xi \mathbf{a}'\hat{\boldsymbol{\beta}}$ et $E_\xi \mathrm{Var}_\xi \mathbf{a}'\hat{\boldsymbol{\beta}}$ pourrait ne pas tenir, sauf sous des hypothèses contraignantes, telles que l'indépendance des résidus; donc, les termes du facteur du milieu de (5) prendraient une forme plus complexe. Plusieurs cas sont possibles pour l'échantillonnage par grappe, selon la relation entre les cellules et les grappes, et il conviendrait de les élaborer dans le cadre de futurs travaux.

Une autre extension naturelle est celle aux modèles de régression non linéaire et à d'autres grandeurs à estimer définies par des équations d'estimation. La formulation des moindres carrés pondérés de l'étape de Newton-Raphson (McCullagh et Nelder 1989, section 2.5) pour un modèle linéaire généralisé peut être appliquée en définissant de manière appropriée σ_{bk}^2 dans (3) et donc dans (4) à (6); une procédure semblable peut être appliquée à d'autres équations d'estimation (Binder 1981; Binder 1983; Morel 1989). Comme les variances sont des fonctions des prédictions du modèle, la mise en œuvre de ces modifications nécessite des hypothèses de conception du plan d'échantillonnage concernant le modèle ajusté, ainsi que la distribution des covariables.

Toute optimisation a son prix, lequel, dans le cas de nos méthodes, peut être pratique et statistique.

Dans l'étude CanCORS, les nouveaux cas de cancer étudiés ont été identifiés en temps réel grâce à une opération sur le terrain (« confirmation rapide des cas »); une prise de contact a alors eu lieu très rapidement avec les patients afin de commencer à communiquer avec eux pour les interviewer dans l'intervalle souhaité (trois mois après le diagnostic). Donc, les problèmes pratiques de mise en œuvre d'un sondage étaient exacerbés. Les considérations qui, en dernière analyse, nous ont menés à appliquer le plan d'échantillonnage GDP ont été, notamment 1) la difficulté de procéder à un géocodage précis des patients dans le délai permis, 2) la détermination incomplète et inexacte de la race dans les données de confirmation des cas et 3) les taux de participation plus faibles que prévu qui ont rendu tout type d'échantillonnage problématique.

Ce genre de situation pose moins de problèmes dans les sondages à l'aide d'une base de sondage statique qui peut être traitée dans des délais moins contraignants, particulièrement dans les sondages à grande échelle et (ou) répétés dans lesquels même une réduction modeste de la variance justifie une complexité supplémentaire. Ils pourraient être utilisés, par exemple, pour évaluer les améliorations éventuelles grâce au suréchantillonnage à base géographique dans les sondages pour lesquels des estimations nationales selon la race sont requises.

Les préoccupations statistiques concernant d'autres stratégies d'élaboration du plan d'échantillonnage découlent du fait que l'optimisation pour un ensemble d'objectifs statistiques prédéterminés réduira vraisemblablement l'efficacité pour d'autres ensembles. À part pour les études dont les objectifs sont cernés le plus étroitement, il est généralement difficile de prévoir toutes les analyses possibles. L'optimisation simultanée pour un ensemble raisonnablement complet d'analyses, et l'étude de la sensibilité du plan d'échantillonnage à la variation des poids relatifs des divers objectifs devrait offrir une certaine protection contre un plan d'échantillonnage surspécialisé. Cependant, cette approche ne peut être utilisée qu'avec des variables pour lesquelles il existe certaines données avant l'étude. Les résultats de la section 2.6 donnent à penser que la surveillance de l'effet de l'échantillonnage non proportionnel sur la précision de la moyenne de population offre une certaine protection contre les plans d'échantillonnage excessivement inefficaces pour les analyses et les variables non prévues, quoique les bornes ne soient pas très générales.

De manière plus générale, nous pourrions nous demander quand l'analyse GDP correspond à la grandeur à estimer scientifiquement pertinente. Les modèles de régression sont souvent utilisés dans des analyses destinées à être généralisables à des populations plus générales, plutôt qu'à décrire la population finie dont on dispose, tout comme les emplacements de l'étude CanCORS ont été choisis à dessein pour étudier les profils et les variations des soins qui pourraient refléter les tendances nationales plus générales. Bien que l'utilisation de poids d'échantillonnage dans des

études énumératives suscite relativement peu de controverse, leur utilisation dans les études analytiques a provoqué un vif débat (Hansen, Madow et Tepping, 1983 et discussion; DuMouchel et Duncan 1983; Bellhouse 1984; Pfeffermann 1993, Fuller 2002, section 5). Une analyse descriptive de la population offre une certaine robustesse à la sélection éventuelle de l'échantillon d'une manière qui fausse les relations typiques. Donc, même si l'on ne peut pas justifier une analyse GDP pure pour des raisons de représentativité énumérative, un échantillon tiré pour optimiser l'estimation non pondérée des coefficients de régression pourrait avoir une valeur scientifique limitée. Ainsi, supposons que les données CanCORS soient analysées par une régression *non pondérée* pour estimer un simple effet du revenu (un contraste des moyennes) en utilisant l'information sur les variables du plan d'échantillonnage au niveau de l'îlot provenant du recensement. De manière optimale, l'échantillon serait tiré à partir d'un ensemble d'îlots pour lesquels, collectivement, environ la moitié des résidents vivraient dans la pauvreté. Puisque les taux de pauvreté sont rarement aussi élevés, cette situation requiert effectivement un échantillonnage uniquement dans les îlots où les taux de pauvreté sont les plus élevés. Pareil échantillonnage ne serait représentatif d'aucune des catégories de revenu. De même, un échantillon qui surreprésente les résidents noirs par échantillonnage d'îlots habités principalement par des Noirs serait (s'il est analysé sans pondération) non représentatif de la population noire en général, parce que les services offerts dans les régions où la ségrégation est forte diffèrent vraisemblablement de ceux offerts dans des régions mixtes.

Il est nécessaire d'établir des formules plus générales, avec les hypothèses et les objectifs clairement énoncés, dans lesquelles « les paramètres du modèle sont considérés comme les paramètres cibles ultimes, mais se concentrent en même temps sur les grandeurs descriptives de la population (GDP) comme moyen de rendre l'inférence robuste » (Pfeffermann 1993), en tenant compte des objectifs scientifiques de l'étude. Il a été proposé antérieurement de tester l'hypothèse nulle que les poids n'ont pas d'effet sur la régression (DuMouchel et Duncan 1983; Fuller 1984) ou d'inclure les variables du plan d'échantillonnage (Nathan et Holt, 1980; Little 1991) ou les poids proprement dits (Rubin 1985) comme variables de contrôle dans la régression. Ces approches posent néanmoins des problèmes quand les poids sont des fonctions des covariables d'intérêt scientifique primordial. Nous avons essayé, grâce à la pondération flexible des contrastes (section 2.5), de faire un pas vers une formule générale, en étendant l'approche GDP afin de mettre l'accent sur une gamme d'inférences valides pour des objectifs scientifiques particuliers, plutôt qu'exclusivement sur l'inférence pour des populations finies. À partir de cette gamme, un chercheur peut choisir un objectif inférentiel et un plan d'échantillonnage adapté à la structure de la population et aux aspects pratiques de la conception de l'étude.

Annexe

Équivalence des variances sous le plan d'échantillonnage et sous le modèle

Nous montrons que $E_\xi \mathrm{Var}_\pi \hat{\boldsymbol{\beta}} \approx E_\pi \mathrm{Var}_\xi \hat{\boldsymbol{\beta}}$ sous les contraintes suivantes :

1. $1/N\ \mathbf{X}'_U \mathbf{X}_U \to \Sigma$ pour une certaine Σ définie positive. Cette condition minimale relie la série hypothétique de populations.
2. L'estimateur par la régression sous le plan de sondage peut s'écrire $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_U + R_{n,S}$, où $\mathrm{Var}_\xi E_\pi R_{n,S} = o(n^{-1})$ et $\mathrm{Var}_\pi E_\xi R_{n,S} = o(n^{-1})$. Notons que $\hat{\boldsymbol{\beta}}$ ne peut pas être défini strictement comme dans (2), parce que la matrice inverse est indéfinie quand les valeurs d'échantillon de x ne couvrent pas l'espace du plan d'échantillonnage et que son espérance et sa variance sont également indéfinies. Un estimateur par le ratio scalaire pourrait de même être indéfini avec une probabilité non nulle mais à l'ordre $o(n^{-1})$, parce que l'échantillon pourrait n'avoir que des valeurs 0 pour la variable du dénominateur. Si nous attribuons une valeur arbitraire à cet événement, l'estimateur pourrait néanmoins avoir de bonnes propriétés asymptotiques. Un argument semblable nous permet de supposer qu'un $\hat{\boldsymbol{\beta}}$ approprié peut être défini. Nous ne spécifions pas comment (2) doit être modifié pour satisfaire techniquement la contrainte, puisque celle-ci dépend des caractéristiques particulières de ξ et de la série de plans d'échantillonnage.
3. $\max(\pi_i) = O(n/N)$ et $n = o(N)$, essentiellement notre hypothèse que les corrections pour population finie peuvent être ignorées.
4. Homoscédasticité, $\mathrm{Var}_\xi\, y_k = \sigma^2$; cela n'est pas contraignant, puisque cette condition peut toujours être rendue vraie par une transformation appropriée de x et de y.

Sous ces contraintes,

$$\begin{aligned}\mathrm{Var}_{\pi\xi}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}_\pi E_\xi \hat{\boldsymbol{\beta}} + E_\pi \mathrm{Var}_\xi \hat{\boldsymbol{\beta}} \\ &= \mathrm{Var}_\pi \boldsymbol{\beta} + E_\pi \mathrm{Var}_\xi \hat{\boldsymbol{\beta}} \\ &= o(n^{-1}) + E_\pi \mathrm{Var}_\xi \hat{\boldsymbol{\beta}}\end{aligned}$$

Par ailleurs

$$\mathrm{Var}_{\xi\pi}\,\hat{\boldsymbol{\beta}} = \mathrm{Var}_{\xi}E_{\pi}\,\hat{\boldsymbol{\beta}} + E_{\xi}\mathrm{Var}_{\pi}\,\hat{\boldsymbol{\beta}}$$

Le premier terme de l'équation qui précède est

$$\begin{aligned}\mathrm{Var}_{\xi}E_{\pi}(\boldsymbol{\beta}_U + R_{n,S}) &= \mathrm{Var}_{\xi}(\boldsymbol{\beta}_U + E_{\pi}R_{n,S}) \\ &= (\mathbf{X}'_U\mathbf{X}_U)^{-1}\sigma^2 + o(n^{-1}) + o(N^{-1/2}n^{-1/2}) \\ &= O(N^{-1}) + o(n^{-1}) + o(n^{-1/2}N^{-1/2}) = o(n^{-1})\end{aligned}$$

La preuve est une élaboration d'une preuve donnée par Isaki et Fuller (1982), résumée dans Pfeffermann (1993, page 321).

Remerciements

La présente étude a été financée par les subventions U01-CA93344 (Zaslavsky et Zheng), U01-CA93324 (Zaslavsky) et U01-CA093348 (Adams) du National Cancer Institute. Les auteurs remercient Nat Schenker et Van Parsons de leurs commentaires constructifs concernant une version antérieure, ainsi que le rédacteur adjoint et deux examinateurs de leurs commentaires judicieux.

Bibliographie

Ayanian, J.Z., Chrischilles, E.A., Wallace, R.B., Fletcher, R.H., Fouad, M., Kiefe, C.I., Harrington, D.P., Weeks, J.C., Kahn, K.L., Malin, J.L., Lipscomb, J., Potosky, A.L., Provenzale, D.T., Sandler, R.S., Vanryn, M. et West, D.W. (2004). Understanding cancer treatment and outcomes: The Cancer Care Outcomes Research and Surveillance Consortium. *Journal of Clinical Oncology*, 2, 2992-2996.

Bellhouse, D.R. (1984). A review of optimal designs in survey sampling. *The Canadian Journal of Statistics*, 12, 53-65.

Binder, D.A. (1981). Les variances d'estimateurs asymptotiquement normaux bases sur des enquêtes complexes. *Techniques d'enquête*, 7, 162-178.

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

Botman, S.L., Moore, T.F., Moriarity, C.N. et Parsons, V.L. (2000). *Design and Estimation for the National Health Interview Survey, 1995-2004*. Vital and Health Statistics, 2(130). Washington, DC : National Center for Health Statistics.

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.

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

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

Denton, N.A., et Massey, D.S. (1993). *American Apartheid: Segregation and the Making of the Underclass*. Cambridge, MA : Harvard University Press.

Dumouchel, W.H., et Duncan, G.J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.

Eaton, D.K., Kann, L., Kinchen, S., Ross, J., Hawkins, J., Harris, W.A., Lowry, R., Mcmanus, T., Chyen, D., Shanklin, S., Lim, C., Grunbaum, J.A. et Wechsler, H. (2006). Youth risk behavior surveillance - United States, 2005. *Morbidity and Mortality Weekly Report*, 55(SS-5), 1-108.

Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhyā*, C37, 117-132.

Fuller, W.A. (1984). Application de la méthode des moindres carrées et de techniques connexes aux plans de sondage complexes. *Techniques d'enquête*, 10, 107-130.

Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.

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.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-97.

Kish, L. (1974). Optimal and proximal multipurpose allocation. Alexandria, VA. In *Proceedings of the Social Statistics Section*, American Statistical Association, 111-118.

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

McCullagh, P., et Nelder, J.A. (1989). *Generalized Linear Models* (Deuxième édition). London : Chapman & Hall Ltd.

Morel, J.G. (1989). Régression logistique selon des plans de sondage complexes. *Techniques d'enquête*, 15, 213-233.

Morris, C., Newhouse, J.P. et Archibald, R. (1979). On the theory and practice of obtaining unbiased and efficient samples in social surveys and experiments. *Research in Experimental Economics*, 1, 199-220.

Nathan, G., et Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society*, Séries B, *Methodological*, 42, 377-386.

Neyman, J. (1934). On the two different aspects of representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

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

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Rubin, D.B. (1985). The use of propensity scores in applied Bayesian inference. Dans *Bayesian Statistics 2*, (Éds., J.M. Bernardo, M.H. DeGroot, D.V. Lindley et A.F.M. Smith). Amsterdam : Elsevier/North-Holland, 463-472.

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

Sastry, N., Ghosh-Dastidar, B., Adams, J. et Pebley, A.R. (2005). The design of a multilevel survey of children, families, and communities: The Los Angeles Family and Neighborhood Survey. *Social Science Research*, 35, 1000-1024.

Thomsen, I., Tesfu, D. et Binder, D.A. (1986). Estimation of design effects and intraclass correlations when using outdated measures of size. *Revue Internationale de Statistique*, 54, 343-349.

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

Waksberg, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. Dans *ASA Proceedings of the Social Statistics Section*, American Statistical Association. Alexandria, VA.

Waksberg, J. (1995). Distribution of poverty in census block groups (BGs) and implications for sample design. Dans *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association. Alexandria, VA, 497-502.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817-838.

Techniques d'enquête, juin 2008
Vol. 34, N° 1, pp. 87-98
Statistique Canada, N° 12-001-X au catalogue

Estimateurs par la régression généralisée d'un total de population finie utilisant la méthode de Box-Cox

Yan Li [1]

Résumé

Un nouvel estimateur par la régression généralisée d'un total de population finie basé sur la méthode de transformation de Box-Cox et son estimateur de la variance sont proposés sous un plan général de sondage à probabilités inégales. En étant convergent par rapport au plan de sondage, l'estimateur proposé retient la propriété de robustesse de l'estimateur GREG, même si le modèle sous-jacent est défaillant. En outre, la méthode de Box-Cox permet de trouver automatiquement une transformation raisonnable de la variable dépendante en se servant des données. La robustesse et l'efficacité du nouvel estimateur sont évaluées analytiquement et par des études en simulation de Monte Carlo.

Mots clés : Estimateur par la régression généralisée (GREG); méthode de Box-Cox; pseudo maximum de vraisemblance (PMV).

1. Introduction

Les estimateurs par la régression généralisée (GREG) de totaux et de moyennes de population finie sont établis au moyen de modèles de régression appropriés. Bien que l'on se serve de modèles pour construire ces estimateurs, la sélection des échantillons et l'évaluation des propriétés statistiques des stratégies d'estimation résultantes doivent se faire sous randomisation. Des exemples peuvent être consultés dans Särndal, Swensson et Wretman (1992), Estevao, Hidiroglou et Särndal (1995), Fuller, Loughin et Baker (1994), ainsi que Jayasuriya et Valliant (1996). Le choix d'un bon modèle est essentiel si l'on veut limiter la variabilité d'un estimateur *assisté par modèle* tel que l'estimateur GREG. Si le modèle hypothétique décrit bien la population finie, l'utilisation de l'estimateur GREG au lieu de l'estimateur d'Horvitz-Thompson (Horvitz et Thompson 1952) peut réduire considérablement la variance. Une discussion générale de l'estimation par la régression est présentée dans Fuller (2002). Särndal et coll. (1992) offrent une description détaillée du cadre de travail assisté par modèle pour la construction d'estimateurs de sondage.

Les études portant sur l'estimateur GREG ont été réalisées pour la plupart dans le contexte de la modélisation par la régression linéaire. Essentiellement, l'estimateur GREG intègre les variables auxiliaires pertinentes par la voie de leurs totaux de population de contrôle connus, même si les variables auxiliaires sont connues pour chaque unité de la population (Cassel, Särndal et Wretman 1976; Särndal 1980; Deville et Särndal 1992; Särndal et coll. 1992; Jiang et Lahiri 2006). De nos jours, il est assez fréquent que l'on dispose d'information auxiliaire complète : les données de recensement, les registres administratifs, les données de télédétection et les enquêtes antérieures constituent une mine de renseignements précieux qui peuvent être utilisés pour accroître la précision de la procédure d'estimation (Montanari et Ranalli 2003). Par conséquent, des modèles complexes et des méthodes souples s'appuyant sur l'information auxiliaire complète ont été introduits dans les sondages ces dernières années. Les techniques des splines pénalisées ont été adaptées pour construire des estimateurs fondés sur un modèle (Zheng et Little 2004) et assistés par un modèle (Breidt, Claeskens et Opsomer 2005) pour un total de population finie en se basant sur des données d'enquête complexe. Breidt et Opsomer (2000) ont considéré un estimateur par la régression non paramétrique, assisté par modèle, recourant à la régression par polynômes locaux et ont montré que la régression non paramétrique peut accroître significativement l'efficacité des estimateurs quand les modèles paramétriques sont spécifiés incorrectement. Leurs travaux ont été étendus du modèle à une seule covariable au cas du modèle additif semi-paramétrique. Wu et Sitter (2001) ajustent un modèle de travail général, qui pourrait comprendre à la fois des composantes linéaires et non linéaires, puis le calent sur les valeurs prédites résultantes par régression linéaire simple. Montanari et Ranalli (2005) combinent l'estimation par calage fondé sur un modèle et des méthodes non paramétriques, et proposent des estimateurs non paramétriques assistés par modèle pour une moyenne de population finie.

Alors qu'en statistique classique, on procède souvent à une transformation appropriée de la variable dépendante dans le modèle hypothétique pour obtenir la normalité, la linéarité et l'homoscédasticité (Carroll et Ruppert 1988), la littérature sur les transformations dans le contexte de l'inférence en population finie n'est pas très riche. L'intérêt croît cependant pour l'élaboration de méthodes utilisant une transformation appropriée et des données de sondage. Dans

1. Yan Li, University of Maryland at College Park. Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS; Executive Plaza South; Room 8014, 6120 Executive Boulevard, MSC7244, Bethesda, MD 20892, États-Unis. Courriel : lisherry@mail.nih.gov.

certaines applications des sondages, particulièrement les enquêtes auprès des entreprises et des établissements, les variables étudiées sont souvent continues, positives et fortement asymétriques (par exemple, le revenu). Pour estimer le total de population finie, l'application d'un modèle linéaire à la variable étudiée pourrait ne pas convenir, mais être raisonnable pour une transformation strictement monotone de cette variable. Chen et Chen (1996) ont envisagé l'utilisation de données d'enquête transformées pour améliorer la précision de l'approximation normale. Korn et Graubard (1998) ont comparé divers intervalles de confiance, y compris des intervalles fondés sur une transformation logit, pour l'estimation de proportions dans le cas d'un petit nombre prévu de dénombrements positifs. Karlberg (2000) a proposé un estimateur basé sur un modèle de superpopulation lognormal-logistique pour prédire le total de population finie d'une variable étudiée à distribution fortement asymétrique. Les résultats de simulation ont indiqué que l'estimateur fondé sur le modèle lognormal-logistique offre une alternative sensée à d'autres estimateurs, particulièrement quand la taille de l'échantillon est faible. Chambers et Dorfman (2003) ont discuté de l'estimation d'une moyenne de population finie sous certaines transformations générales, mais connues, des données continues.

Les chercheurs jugent la technique de transformation fructueuse pour l'analyse des données d'enquête. L'étape clé consiste à trouver une transformation appropriée, bien adaptée à ces données. Dans de nombreuses applications, la forme de la transformation est déterminée subjectivement. Malheureusement, les renseignements *a priori* ou la théorie ne suggèrent pas nécessairement celle qu'il convient d'utiliser. Le cas échéant, il serait commode d'identifier la transformation de manière adaptative, en utilisant les données d'enquête.

Les travaux de Box et Cox (1964) ont mené à l'élaboration de méthodes du type « les données décident de la transformation » pour la construction de modèles avec erreurs indépendantes et identiquement distribuées. Leur article et d'autres traitant du sujet, dont Tukey (1957), John et Draper (1980), ainsi que Bickel et Doksum (1981), ont inspiré une foule d'études appliquées. Spitzer (1976) a examiné la relation entre la demande d'argent et la trappe à liquidité à l'aide d'un modèle de Box-Cox généralisé. Dans le contexte de la recherche sur la malaria, Newman (1977) a conclu que la spécification fonctionnelle de Box-Cox était supérieure aux spécifications antérieures. Miner (1982), ainsi que Davison, Arnade et Hallahan (1989) ont étudié la modélisation des fonctions du rendement du soja et de l'exportation de soja par les États-Unis, respectivement. Ils ont conclu que la transformation de Box-Cox produit des termes d'erreur qui suivent approximativement une loi normale. Une bibliographie sur la transformation de Box-Cox peut être consultée dans un article de synthèse publié par Sakia (1992). Pour une application de la méthodologie de Box-Cox à un modèle linéaire mixte, voir Gurka (2004, 2006).

Li et Lahiri (2007) appliquent la transformation de Box-Cox à la variable étudiée pour obtenir des prédicteurs fondés sur un modèle d'un total de population finie qui sont robustes. Ils mentionnent aussi les estimateurs assistés par modèle dans une sous-section (section 2.6), mais n'étudient pas leurs propriétés. Le présent article fournit cette analyse. Nous employons la technique de Box-Cox pour ajuster une régression de la variable étudiée sur un ensemble de variables auxiliaires. Puis, nous utilisons la régression ajustée pour prédire les valeurs de la variable étudiée pour les unités non observées de la population finie qui, à leur tour, fournissent un estimateur de type régression adaptative dans le cadre de travail assisté par un modèle.

La présentation de l'article est la suivante. À la section 2, nous proposons un nouvel estimateur et examinons les propriétés analytiques de ce dernier en fonction du plan de sondage. À la section 3, pour mieux évaluer la robustesse et l'efficacité de l'estimateur proposé, nous le comparons aux estimateurs GREG fondés sur les modèles de travail linéaire et loglinéaire sous-jacents par simulations de Monte Carlo. À la section 4, nous discutons des résultats. Enfin, à la section 5, nous tirons certaines conclusions au sujet des domaines dans lesquels les travaux de recherche pourraient se poursuivre.

2. Un nouvel estimateur du total de population finie

Supposons que la quantité d'intérêt est le total de population finie

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

où $U = \{1, ..., N\}$ dénote une population finie de taille connue N, et $y_i > 0$ est la valeur de la variable étudiée associée à l'unité i. Écrivons $\mathbf{Y} = (y_1, ..., y_N)'$. Pour estimer T, nous tirons un échantillon s de taille n de la population finie selon un plan d'échantillonnage probabiliste. Soit w_i le poids d'échantillonnage de l'unité i. Ce poids est simplement l'inverse de la probabilité d'inclusion de l'unité i, dénotée par $\pi_i = P(i \in s)$ $(i = 1, ..., N)$. Nous supposons que nous avons de l'information sur $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)'$, où $\mathbf{x}_i = (1, x_{i1}, ..., x_{ik})'$ est un vecteur colonne de k variables auxiliaires pour l'unité i. Pour tout échantillon s de taille n, nous redéfinissons $\mathbf{Y}$ et $\mathbf{X}$ de façon que les n premières lignes de $\mathbf{Y}$ et $\mathbf{X}$ correspondent à celles figurant dans l'échantillon. Écrivons

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{pmatrix},$$

où

- $\mathbf{y}_s$ est un vecteur colonne de dimensions $n \times 1$ de valeurs observées de la variable étudiée;
- $\mathbf{y}_r$ est un vecteur colonne de dimensions $(N - n) \times 1$ de valeurs non observées de la variable étudiée;
- $\mathbf{X}_s$ est une matrice de dimensions $n \times (k + 1)$ de variables auxiliaires connues dans l'échantillon;
- $\mathbf{X}_r$ est une matrice de dimensions $(N - n) \times (k + 1)$ de variables auxiliaires connues hors de l'échantillon.

Tout au long de l'article, nous utilisons E_d et V_d pour dénoter l'espérance, ou valeur prédite, et la variance sous le plan d'échantillonnage.

2.1 Estimateurs GREG de totaux de population finie

L'estimateur GREG est défini ici comme étant

$$\hat{T}_G = \sum_{i \in U} \hat{y}_{i,w} + \sum_{i \in s} (y_i - \hat{y}_{i,w})/\pi_i,$$

où $\hat{y}_{i,w}$ est le prédicteur de y_i fondé sur un modèle. Quel que soit le degré avec lequel le modèle sous-jacent décrit bien la population, les estimateurs GREG du total de population finie sont convergents par rapport au plan sous des contraintes faibles (Särndal et coll. 1992). Le modèle utilisé le plus fréquemment est le modèle de régression linéaire standard, donné par

$$\mathbf{M}_1 : \ \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

où $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I})$, une loi de probabilité à N variables de vecteur moyen $\mathbf{0}$ et de matrice de variance-covariance $\sigma^2\mathbf{I}$, et $\mathbf{I}$ est la matrice d'identité de dimensions $N \times N$ (rien ne serait perdu dans ce contexte en remplaçant $\sigma^2\mathbf{I}$ par une matrice définie positive plus générale). Dans cette équation, $\boldsymbol{\beta}$ est un vecteur colonne de dimensions $(k + 1) \times 1$ de coefficients de régression. Les paramètres σ^2 et $\boldsymbol{\beta}$ sont l'un et l'autre des paramètres de superpopulation inconnus. Un prédicteur sans biais pour la i^{e} unité est donné par

$$\hat{y}_{i,w} = \mathbf{x}_i' \hat{\boldsymbol{\beta}}_w, \qquad (1)$$

où $\hat{\boldsymbol{\beta}}_w$ est l'estimateur des moindres carrés pondérés (MCP) de $\boldsymbol{\beta}$ sous $\mathbf{M}_1$ et

$$\hat{\boldsymbol{\beta}}_w = \left(\sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i \in s} w_i \mathbf{x}_i y_i \right).$$

Dans certaines applications, surtout dans les enquêtes auprès des entreprises et les enquêtes agricoles, un modèle linéaire n'est pas nécessairement approprié pour y, mais peut être raisonnable pour une transformation strictement monotone de y. Pour l'ensemble de données utilisé dans Royall et Cumberland (1981), Chen et Chen (1996) ont observé que la distribution de la population finie était fortement asymétrique et que la transformation logarithmique aidait à la rendre symétrique. Le besoin d'effectuer la transformation logarithmique et les avantages de celle-ci étaient évidents. Par conséquent, nous considérons le modèle de régression loglinéaire, où la transformation logarithmique est appliquée à la variable dépendante

$$\mathbf{M}_2 : \ \log\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

où $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I})$. Un prédicteur évident pour la i^{e} unité est donné par

$$\hat{y}_{i,w} = e^{\mathbf{x}_i' \hat{\boldsymbol{\beta}}_{lw}}, \qquad (2)$$

où

$$\hat{\boldsymbol{\beta}}_{lw} = \left(\sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i \in s} w_i \mathbf{x}_i \log y_i \right).$$

Le modèle $\mathbf{M}_2$ requiert une spécification subjective de la transformation appliquée à la variable étudiée, ce qui peut être raisonnable dans des situations où nous savons quelle est la transformation appropriée grâce à des données empiriques *a priori* ou à la théorie. En l'absence de toute connaissance *a priori* au sujet de la transformation, il est prudent de choisir cette dernière parmi une famille flexible de transformations en utilisant les données.

Tukey (1957) a examiné la famille de transformations puissance :

$$y^{(\lambda)} = \begin{cases} y^{\lambda} & \lambda \neq 0, \\ \log(y) & \lambda = 0, \end{cases}$$

où $y > 0$. Afin d'éliminer la discontinuité à $\lambda = 0$, Box et Cox (1964) ont proposé la famille de transformations suivante :

$$y^{(\lambda)} = \begin{cases} (y^{\lambda} - 1)/\lambda & \lambda \neq 0, \\ \log(y) & \lambda = 0, \end{cases}$$

où $y > 0$. Le paramètre λ détermine la nature de la transformation. Par exemple, les valeurs $\lambda = 1,\ 0,\ 0{,}5,\ -1$ correspondent à l'absence de transformation, une transformation logarithmique, une transformation racine carrée et une transformation inverse, respectivement. Le paramètre de transformation λ est estimé d'après les données. L'analyse de Box-Cox peut mener à une transformation logarithmique, mais tout aussi bien aboutir à une autre transformation de la famille susmentionnée, selon les données réelles observées.

Nous considérons le modèle de superpopulation suivant pour la variable étudiée transformée :

$$\mathbf{M}_3 : \mathbf{Y}^{(\lambda)} = \mathbf{X}\beta + \boldsymbol{\varepsilon},$$

où les résidus $\boldsymbol{\varepsilon}$ suivent une loi approximativement normale de moyenne $\mathbf{0}$ et de matrice de variance $\sigma^2\mathbf{I}$.

Schlesselman (1971) a montré que les estimateurs du maximum de vraisemblance des paramètres du modèle de Box-Cox sont indépendants de l'échelle, si bien que le rééchelonnement de la variable y originale aboutit à la même fonction de log-vraisemblance à condition que le modèle de régression contienne un terme constant. Inspirés par cette étude et par la plupart des articles sur les modèles de Box-Cox, nous incluons un terme constant dans le modèle.

Sous $\mathbf{M_3}$, le prédicteur pour la i^{e} unité non observée s'obtient par une simple rétrotransformation à partir de la transformée de Box-Cox :

$$\hat{y}_{i,w} = g_i(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) = (\hat{\lambda}_w \mathbf{x}_i' \hat{\boldsymbol{\beta}}_w + 1)^{1/\hat{\lambda}_w}, \qquad (3)$$

où $\hat{\boldsymbol{\beta}}_w$ et $\hat{\lambda}_w$ sont les estimateurs des paramètres du modèle. La méthode d'estimation est expliquée à la section suivante. Les équations (2) et (3) ne fournissent pas de prédicteur sans biais de y_i pour les modèles sous-jacents respectifs. Li et Lahiri (2007) ont montré que, si la variance des résidus est faible, c'est-à-dire si le modèle est très bien ajusté aux données, les deuxièmes membres des équations (2) et (3) sont de bonnes solutions de rechange pour les prédicteurs sans biais. Pour simplifier et pour réduire le fardeau des calculs, nous traiterons les deuxièmes membres des équations (2) et (3) comme des alternatives appropriées pour les prédicteurs sans biais. Rappelons que notre but est de décrire des estimateurs assistés par modèle pour des totaux de population finie. Le modèle sous-jacent est utilisé uniquement pour suggérer un estimateur, qui sera évalué dans le cadre de randomisation. Même si les prédicteurs des y_i individuels sont biaisés, nous pouvons encore construire des estimateurs convergents par rapport au plan pour des totaux de population finie, contrairement aux estimateurs strictement fondés sur un modèle proposé par Li et Lahiri (2007).

Nous dénotons les estimateurs GREG sous les trois modèles $\mathbf{M_1}$ à $\mathbf{M_3}$ par $\hat{T}_{G_L}$, $\hat{T}_{G_LOGL}$, et $\hat{T}_{G_BC}$, respectivement. L'estimateur $\hat{T}_{G_BC}$ diffère des estimateurs $\hat{T}_{G_L}$ et $\hat{T}_{G_LOGL}$, parce que les données dictent la transformation qu'il faut utiliser.

Il est possible d'intégrer des transformations de Box-Cox sur la variable y ainsi que les variables x. Par le passé, diverses formes fonctionnelles du modèle de Box-Cox ont été étudiées. Khan et Ross (1977), Spitzer (1976), Zarembka (1968), Boylan, Cuddy et O'Muircheartaigh (1980), et d'autres ont considéré un cas particulier du modèle général de Box-Cox quand on émet l'hypothèse d'un paramètre de transformation commun pour la variable y et les variables x. Gemmill (1980), Boylan, Cuddy et O'Muircheartaigh (1982), et d'autres ont appliqué le modèle général de Box-Cox avec divers paramètres de transformation à la variable y et aux variables x. Li et Lahiri (2007) ont également utilisé le modèle général de Box-Cox pour prédire le total de population finie sous un cadre fondé sur un modèle. Dans l'avenir, cette méthode pourrait être étendue à diverses formes fonctionnelles de la transformation de Box-Cox.

Ici, nous discutons uniquement d'une transformation Box-Cox de la variable y, ce qui permet une comparaison plus équitable des trois estimateurs GREG.

2.2 Estimation des paramètres du modèle et de la transformation $\boldsymbol{\varphi} = (\boldsymbol{\beta}, \lambda, \sigma^2)'$ par la méthode du pseudo maximum de vraisemblance (PMV)

Afin de faciliter l'estimation de λ en utilisant les procédures de calcul existantes, il faut remplacer $\mathbf{Y}^{(\lambda)}$ dans le modèle $\mathbf{M_3}$ par une transformation à l'échelle $\mathbf{Y}^{*(\lambda)}$. Pour la i^{e} unité,

$$y_i^{*(\lambda)} = \begin{cases} (y_i^{\lambda} - 1)/\lambda \tilde{y}^{\lambda-1} & \lambda \neq 0, \\ \tilde{y}\log(y_i) & \lambda = 0, \end{cases}$$

où $\tilde{y}$ est la moyenne géométrique des y. Les calculs qui suivent seront fondés sur le nouveau modèle à l'échelle :

$$\mathbf{M}_4 : \mathbf{Y}^{*(\lambda)} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*,$$

où les $\boldsymbol{\varepsilon}^*$ suivent une loi approximativement normale de moyenne $\mathbf{0}$ et de matrice de variance $\sigma_e^{*2}\mathbf{I}$. Soit $\boldsymbol{\varphi}^* = (\boldsymbol{\beta}^*, \lambda, \sigma_e^{*2})'$.

L'estimateur du maximum de vraisemblance (EMV) de $\boldsymbol{\varphi}^*$ maximise la log-vraisemblance

$$l(\boldsymbol{\varphi}^*) = \sum_i \log f(y_i; \boldsymbol{\varphi}^*, \tilde{y}),$$

où

$$f(y_i; \boldsymbol{\varphi}^*, \tilde{y}) = (2\pi\sigma_e^{*2})^{-1/2} \exp\left\{-(2\sigma_e^{*2})^{-1}(y_i^{*(\lambda)} - \mathbf{x}_i'\boldsymbol{\beta}^*)^2\right\} \cdot (y_i/\tilde{y})^{\lambda-1}.$$

Skinner, Holt et Smith (1989) redéfinissent $\boldsymbol{\varphi}^*$ comme la valeur de $\tilde{\boldsymbol{\varphi}}^*$ qui maximise

$$l(\tilde{\boldsymbol{\varphi}}^*) = \sum_{i \in U} \log f(y_i, \tilde{\boldsymbol{\varphi}}^*),$$

la somme étant calculée sur l'ensemble des unités de la population finie. Donc, parmi tous les modèles possibles $f(y_i, \tilde{\boldsymbol{\varphi}}^*)$, on choisit celui qui « est le mieux ajusté » à la population finie. Si nous choisissons mal la famille $f(y_i, \tilde{\boldsymbol{\varphi}}^*)$, cet ajustement le meilleur restera médiocre, mais dans notre inférence, nous le traitons comme la cible que nous essayons d'atteindre avec nos données d'échantillon. Par conséquent, il est important de choisir les options appropriées pour $f(y_i, \tilde{\boldsymbol{\varphi}}^*)$.

Pour la population finie, $\boldsymbol{\varphi}^*$ satisfait

$$\dot{l}_U(\boldsymbol{\varphi}^*) = \sum_{i\in U}\left[\partial \log f(y_i;\ \boldsymbol{\varphi}^*,\ \tilde{y})/\partial\boldsymbol{\varphi}^*\right] = 0,$$

où

$$\tilde{y} = \prod_{i=1}^{N} y_i^{1/N}.$$

Pour $\boldsymbol{\varphi}^*$ donné, posons que $\dot{l}_U(\boldsymbol{\varphi}^*)$, sommation de la dérivée première de la log-vraisemblance par rapport à $\boldsymbol{\varphi}^*$, est un paramètre de population finie. Nous tirons un échantillon et, en approximant $\log f(y_i; \boldsymbol{\varphi}^*, \tilde{y})$ pour chaque unité i de l'échantillon par $\log f(y_i; \boldsymbol{\varphi}^*, \tilde{y}_w)$, nous estimons le total de population, $\dot{l}_U(\boldsymbol{\varphi}^*)$, par $\dot{l}_s(\hat{\boldsymbol{\varphi}}^*_{\mathrm{PMV}})$:

$$\dot{l}_s(\hat{\boldsymbol{\varphi}}^*_{\mathrm{PMV}}) = \sum_{i\in s} w_i\left[\partial \log f(y_i;\ \boldsymbol{\varphi}^*,\ \tilde{y}_w)/\partial\boldsymbol{\varphi}^*\right]_{\boldsymbol{\varphi}^*=\hat{\boldsymbol{\varphi}}^*_{\mathrm{PMV}}},$$

où

$$\tilde{y}_w = \prod_{i\in s} y_i^{w_i/\sum_{i\in s} w_i},$$

la moyenne géométrique pondérée des y dans l'échantillon et $\hat{\boldsymbol{\varphi}}^*_{\mathrm{PMV}}$ est l'estimateur du pseudo maximum de vraisemblance de $\boldsymbol{\varphi}^*$, qui satisfait $\dot{l}_s(\hat{\boldsymbol{\varphi}}^*_{\mathrm{PMV}}) = 0$ (Wu et Sitter 2001). L'estimateur PMV, $\hat{\boldsymbol{\varphi}}^*_{\mathrm{PMV}} = (\hat{\boldsymbol{\beta}}^*_w,\ \hat{\lambda}_w,\ \hat{\sigma}^{*2}_{e,w})'$, peut être obtenu par une méthode de balayage de l'espace des solutions. Autrement dit, calculer et tracer les valeurs de log-vraisemblance pondérées,

$$\log L(\boldsymbol{\varphi}^*) = \sum_{i\in s} w_i \log f(y_i;\ \boldsymbol{\varphi}^*,\ \tilde{y}_w) \qquad (4)$$

en fonction de l'ensemble de valeurs de λ nous permettra de localiser l'estimation PMV, $\hat{\lambda}_w$, du paramètre de transformation. Lorsque nous évaluons la fonction de log-vraisemblance à chaque valeur fixe de λ dans le contexte de l'échantillonnage, $\boldsymbol{\beta}^*$ et σ_e^{2*} sont estimés en intégrant les poids d'échantillonnage comme il suit :

$$\hat{\boldsymbol{\beta}}_w^{\ *} = \left(\sum_{i\in s} w_i\,\mathbf{x_i}\,\mathbf{x_i}'\right)^{-1}\left(\sum_{i\in s} w_i\,\mathbf{x}_i\,y_i^{*(\lambda)}\right),$$

$$\hat{\sigma}^{*2}_{e,w} = \sum_{i\in s} w_i\,(y_i^{*(\lambda)} - \mathbf{x_i}'\hat{\boldsymbol{\beta}}^*_w)^2 \Big/ \sum_{i\in s} w_i\,.$$

Puisque c'est le modèle $\mathbf{M_3}$ qui nous intéresse, la reconversion de $\hat{\boldsymbol{\varphi}}^*_{\mathrm{PMV}}$ qui maximise (4) à $\hat{\boldsymbol{\varphi}}_{\mathrm{PMV}}$ dans le modèle $\mathbf{M_3}$ est nécessaire et $\hat{\boldsymbol{\beta}}_w = \tilde{y}_w^{\hat{\lambda}_w-1}\hat{\boldsymbol{\beta}}^*_w$, $\hat{\sigma}^2_{e,w} = \tilde{y}_w^{2(\hat{\lambda}_w-1)}\hat{\sigma}^{*2}_{e,w}$.

Puisque nous prévoyons utiliser un estimateur convergent par rapport au plan pour le total de population, il est raisonnable de nous demander pourquoi nous avons besoin des poids d'échantillonnage pour estimer les paramètres du modèle. Sverchkov et Pfeffermann (2004) soutiennent que l'intégration des poids peut produire de meilleures estimations des paramètres du modèle quand la spécification de ce dernier est correcte dans la population, mais incorrecte dans l'échantillon. Cette situation peut se produire quand les probabilités ou les sélections sont corrélées à l'intérieur des ε_i.

Soit B^* et Λ les deux premières composantes de la solution en échantillon complet de la maximisation de $\sum_U \log f(y_i, \tilde{\boldsymbol{\varphi}}^*)$, et $\mathbf{B} = \tilde{y}^{\Lambda-1}\mathbf{B}^*$. Cela nous permet de définir $\boldsymbol{\theta}_N = (\mathbf{B}, \Lambda)$ comme étant les paramètres de population finie, quelle que soit la validité du modèle.

Nous pouvons montrer que $\hat{\boldsymbol{\theta}}_w = (\hat{\beta}_w, \hat{\lambda}_w)$ est un estimateur convergent par rapport au plan de $\boldsymbol{\theta}_N$, c'est-à-dire que

$$\hat{\boldsymbol{\beta}}_{\mathbf{w}} \to \mathbf{B} \text{ en probabilité et } \hat{\lambda}_w \to \Lambda \text{ en probabilité}$$

sous certaines conditions de régularité en utilisant des arguments semblables à ceux de Binder (1983), Wu (1999), ainsi que Wu et Sitter (2001). Ici, la convergence probabiliste est celle par rapport au plan d'échantillonnage.

2.3 Propriété de convergence de l'estimateur $\hat{T}_{G_BC}$

Il est bien connu que l'estimateur $\hat{T}_{G_L}$ a la propriété désirable de convergence par rapport au plan sous des contraintes faibles (Särndal et coll. 1992). Autrement dit, l'écart relatif entre l'estimateur et la grandeur qu'il estime converge en probabilité vers 0 à mesure que l'échantillon devient arbitrairement grand, que tienne ou non le modèle de travail sur lequel il est fondé. Cette propriété peut être conservée pour l'estimateur $\hat{T}_{G_BC}$.

Définissons

$$\hat{T}_{D_G_BC} = \sum_{i\in U} g_i(\boldsymbol{\theta}_N) + \sum_{i\in s}(y_i - g_i(\boldsymbol{\theta}_N))/\pi_i,$$

et

$$\hat{T}_{G_BC} = \sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) + \sum_{i\in s}(y_i - g_i(\hat{\boldsymbol{\theta}}_w))/\pi_i.$$

Théorème : Sous les hypothèses suivantes, l'estimateur GREG fondé sur la transformation de Box-Cox $\hat{T}_{G_BC}$ est convergent par rapport au plan pour T, en ce sens que $N^{-1}(\hat{T}_{G_BC} - T) = O_p(1/\sqrt{n})$. De surcroît, la variance asymptotique de $\hat{T}_{G_BC}$ est donnée par

$$AV_d(\hat{T}_{G_BC}) =$$
$$\sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)(y_i - g_i(\boldsymbol{\theta}_N))(y_j - g_j(\boldsymbol{\theta}_N))/(\pi_i\,\pi_j),$$

qui peut être estimée par

$$\hat{V}_d(\hat{T}_{G_BC}) =$$
$$\sum_{i\in s}\sum_{j\in s}((\pi_{ij} - \pi_i\pi_j)/\pi_{ij}\ \ )(y_i - g_i(\hat{\boldsymbol{\theta}}_w))(y_j - g_j(\hat{\boldsymbol{\theta}}_w))\Big/(\pi_i\pi_j). \quad (5)$$

Hypothèse 1 : $\hat{\boldsymbol{\theta}}_w = \boldsymbol{\theta}_N + O_p(1/\sqrt{n})$.

Hypothèse 2 : Pour chaque $\mathbf{x}_i$, $\partial g_i(\mathbf{t})/\partial \mathbf{t}$, est continu en $\mathbf{t}$ et $|\partial g_i(\mathbf{t})/\partial \mathbf{t}| \le h(\boldsymbol{\theta})$ pour $\mathbf{t}$ dans le voisinage de $\boldsymbol{\theta}$.

Hypothèse 3 : Les estimateurs d'Horvitz-Thompson calculés avec les poids de sondage de base pour certains totaux de population suivent une loi asymptotiquement normale.

Hypothèse 4 : Pour chaque $\mathbf{x}_i$, la dérivée seconde de $g_i(\mathbf{t})$ par rapport à $\mathbf{t}$ est continue et bornée dans le voisinage de $\boldsymbol{\theta}$.

Preuve : (voir l'annexe).

L'estimateur proposé de la variance représenté par l'équation (5) est basé sur des approximations en grand échantillon. Pour un niveau nominal donné $1 - \alpha$, l'intervalle de confiance habituel fondé sur l'approximation normale pour l'estimateur de la variance donne approximativement un taux de couverture de $100(1 - \alpha)$ % dans les grands échantillons répétés. Malheureusement, dans certains cas, on a constaté que les propriétés de couverture de ce type d'estimateur de la variance peuvent être médiocres pour certains choix du modèle auxiliaire (*assisting model*) pour les estimateurs $\hat{T}_{G_L}$ (Särndal 1982; Särndal, Swensson et Wretman 1989; *etc.*). Les études théoriques et empiriques de la propriété de couverture de l'estimateur proposé de la variance devront être approfondies.

L'estimateur $\hat{T}_{G_BC}$ est convergent par rapport au plan pour le total de population finie T sous l'approche de randomisation, et la technique de Box-Cox permet qu'une transformation raisonnable de la variable dépendante soit déterminée automatiquement par les données à partir d'une grande famille de fonctions, donc permet d'accroître l'efficacité.

3. Une étude par simulation

L'objectif de cette étude par simulation est d'évaluer les propriétés de divers estimateurs GREG pour un total de population finie. Dans cet exercice de simulation, nous produisons une population finie à partir de l'Australian Agricultural and Grazing Industries Survey (AAGIS). Ces données d'enquête contiennent des renseignements sur le nombre de têtes de bétail (y) et la superficie agricole (x) pour chacune des 431 fermes.

Nous considérons une population finie de taille N = 4 000, générée à partir du modèle suivant :

$$\mathbf{M_5}: y_i^{(\lambda)} = (y_i^{\lambda} - 1)/\lambda = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

où les ε_i sont indépendants et de loi approximativement normale $N(0, \sigma^2)$, et x_i est le logarithme d'une valeur produite à partir d'une distribution exponentielle de moyenne μ_x et d'erreur-type σ_x. Afin d'imiter une situation réelle, nous choisissons $\lambda = 0{,}1$, $\beta_0 = 4{,}20$ et $\beta_1 = 2{,}66$ qui sont les estimations obtenues en ajustant les données réelles d'enquête au modèle $\mathbf{M_3}$. Nous fixons $\mu_x = 1\,040$ et $\sigma_x = 1\,000$ pour être certain que $y_i > 0$ pour presque toutes les unités i. Strictement parlant, nous avons une distribution normale tronquée de y, puisque nous écartons toutes les valeurs négatives. L'effet en est négligeable, car moins de 0,1 % des valeurs produites de y doivent être supprimées. Le même phénomène a été observé par Taylor (1986).

La simulation est fondée sur un échantillonnage répété à partir de la population finie générale. Nous étudions deux plans d'échantillonnage, à savoir l'échantillonnage aléatoire simple (EAS) et l'échantillonnage aléatoire simple stratifié (EASS). Lorsqu'un échantillon est sélectionné par EASS, nous appliquons des probabilités de sélection inégales entre les diverses strates. Nous définissons deux strates en utilisant comme valeur limite la médiane des valeurs de y dans la population finie. Pour la strate h de taille N_h, nous sélectionnons un échantillon aléatoire simple de taille n_h. Définissons les probabilités de sélection p_1 et p_2 pour la strate 1 et la strate 2, respectivement. Nous spécifions $p_1 = 2 \times p_2$. Pour une taille d'échantillon fixe n, $n_1 = N_1 \times p_1$, et $n_2 = N_2 \times p_2$.

Nous souhaitons estimer le total de population finie

$$T = \sum_{i \in U} y_i.$$

Dans cette étude par simulation, nous étudions les propriétés des estimateurs $\hat{T}_{G_BC}$, $\hat{T}_{G_L}$, et $\hat{T}_{G_LOGL}$, ainsi que l'estimateur d'Horvitz-Thompson fondé sur le plan $(\hat{T}_D)$, où les indices inférieurs « -L », « -LOGL » et « -BC » dénotent les modèles linéaire, loglinéaire et Box-Cox sous-jacents, respectivement.

Un millier d'échantillons sont tirés de la population finie simulée pour chaque taille d'échantillon $n \in (30, 80, 150)$. Quatre estimateurs sont produits pour chaque échantillon sélectionné. L'estimateur du paramètre de transformation en population finie Λ est également produit pour chaque échantillon. Pour les besoins de la comparaison, nous utilisons deux méthodes pour estimer Λ. Soit $\hat{\lambda}(\hat{\lambda}_w)$ les estimateurs MCO/MV (PMV) de Λ. Sur l'ensemble des 1 000 échantillons, nous calculons les biais relatifs empiriques en pourcentage (BiaisRel) et les racines carrées des erreurs quadratiques moyennes (REQM) pour évaluer ces estimateurs en utilisant les formules suivantes :

$$\text{BiaisRel} = B^{-1} \sum_{b=1}^{B} (\hat{\varpi}_b - \varpi)/\varpi,$$

et

$$\text{REQM} = \sqrt{B^{-1}\sum_{b=1}^{B}(\hat{\varpi}_b - \varpi)^2},$$

où B est le nombre de répétitions dans la simulation de Monte Carlo et $\hat{\varpi}$ représente une estimation arbitraire du paramètre de population finie ϖ.

4. Résultats

Au tableau 1, nous présentons le biais relatif et la racine carrée de l'erreur quadratique moyenne de quatre estimateurs pour différents plans d'échantillonnage et diverses tailles d'échantillon quand $\sigma = 0{,}5$. Les quatre estimateurs donnent tous un biais relatif proche de 0 [le maximum des valeurs absolues du biais relatif (|BiaisRel|) est inférieur à 0,01 dans le tableau 1]. Parmi ces estimateurs, $\hat{T}_{G_BC}$ est celui dont le biais relatif absolu et la REQM sont les plus faibles sur diverses tailles d'échantillon et plans d'échantillonnage. Par conséquent, la technique de Box-Cox protège $\hat{T}_{G_BC}$, qui voit son efficacité accrue comparativement à d'autres estimateurs GREG.

Tableau 1
Biais relatif et racine carrée de l'erreur quadratique moyenne de quatre estimateurs pour différents plans d'échantillonnage et diverses tailles d'échantillon ($\lambda = 0{,}1$)

		$\hat{T}_D^1$	$\hat{T}_{G_L}^2$	$\hat{T}_{G_\text{LOG}L}^3$	$\hat{T}_{G_BC}^4$
Échantillonnage aléatoire simple					
BiaisRel($\times 10^{-3}$)	$n = 30$	4,37	-9,82	5,46	3,62
	$n = 80$	1,43	-3,67	1,24	0,65
	$n = 150$	-2,60	-1,24	1,22	0,53
REQM($\times 10^{7}$)	$n = 30$	7,17	3,54	2,02	1,94
	$n = 80$	4,26	2,09	1,18	1,09
	$n = 150$	3,20	1,58	0,88	0,79
Échantillonnage aléatoire simple stratifié					
BiaisRel($\times 10^{-3}$)	$n = 30$	6,01	-5,82	3,84	1,92
	$n = 80$	9,93	-1,01	3,04	0,98
	$n = 150$	1,75	-1,85	1,06	0,42
REQM($\times 10^{7}$)	$n = 30$	5,63	3,67	2,29	2,11
	$n = 80$	3,51	2,31	1,43	1,28
	$n = 150$	2,49	1,59	1,01	0,90

[1] $\hat{T}_D$: estimateur d'Horvitz-Thompson fondé sur le plan;
[2] $\hat{T}_{G_L}$: estimateur GREG avec modèle linéaire sous-jacent;
[3] $\hat{T}_{G_\text{LOG}L}$: estimateur GREG avec modèle loglinéaire sous-jacent;
[4] $\hat{T}_{G_BC}$: estimateur GREG avec modèle de Box-Cox sous-jacent.

Nous étudions aussi le biais relatif et la REQM sous les mêmes conditions quand $\sigma = 1$ et $\sigma = 2$. La figure 1 donne la REQM pour les trois estimateurs GREG ($\hat{T}_{G_L}$, $\hat{T}_{G_\text{LOG}L}$ et $\hat{T}_{G_BC}$) lorsqu'on utilise différents plans d'échantillonnage. Nous notons que $\hat{T}_{G_BC}$ possède systématiquement la plus petite REQM quand $\sigma = 0{,}5$ et 1. Donc, un modèle robuste choisi par la méthode de Box-Cox réduit la REQM, surtout quand le modèle est approprié pour un petit σ.

Nous pourrions aussi nous intéresser aux propriétés de $\hat{T}_{G_L}$ comparativement à $\hat{T}_{G_BC}$ quand $\lambda = 1$, c'est-à-dire la situation qui favorise $\hat{T}_{G_L}$. Si nous examinons le tableau 2, nous voyons que $\hat{T}_{G_BC}$ est relativement comparable à $\hat{T}_{G_L}$, surtout quand la taille d'échantillon est grande. Ce résultat implique que le fait d'utiliser $\hat{T}_{G_BC}$ alors que nous devrions utiliser $\hat{T}_{G_L}$ n'entraîne pas une grande perte en ce qui concerne le biais relatif et la REQM.

Afin de mieux évaluer la robustesse et l'accroissement de l'efficacité de $\hat{T}_{G_BC}$ comparativement à $\hat{T}_{G_L}$ et à $\hat{T}_{G_\text{LOG}L}$, nous produisons une population finie de taille $N = 4\,000$, non plus conformément au modèle de Box-Cox $\mathbf{M_5}$, mais à partir de

$$\mathbf{M_6}: y_i^{(\lambda)} = \beta_0 + \beta_1 x_i + z_i + \varepsilon_i,$$

où $z_i = x_i^2$. Nous utilisons les mêmes x_i et les mêmes valeurs de paramètres que celles spécifiées à la section 3 pour générer les y_i. Nous étudions les quatre mêmes estimateurs basés sur la nouvelle population finie $\{(x_i, y_i)$ $i = 1, \ldots, 4{,}000\}$. Il s'agit de la situation non idéale pour n'importe lequel des estimateurs GREG. Les résultats sont présentés au tableau 3. Les avantages de l'utilisation de $\hat{T}_{G_BC}$ sont manifestes en ce qui concerne le biais relatif et la REQM.

Le tableau 4 donne le biais relatif et la REQM pour $\hat{\lambda}$ et $\hat{\lambda}_w$ sous échantillonnage EASS avec diverses tailles d'échantillon et les σ fondés sur les valeurs de population produites à partir de $\mathbf{M_5}$. Puisque nous avons stratifié sur la variable y, l'utilisation des pondérations devrait avoir un impact, du moins sur le biais de l'estimation des paramètres. Quand σ est petit, c'est-à-dire quand les données simulées sont bien ajustées au modèle hypothétique, $\hat{\lambda}_w$ donne un biais relatif plus proche de 0, mais $\hat{\lambda}_w$ et $\hat{\lambda}$ donnent l'un et l'autre d'aussi bons résultats en ce qui concerne la REQM. Par contre, quand σ est grand, $\hat{\lambda}_w$ donne systématiquement des valeurs absolues plus petites du biais relatif et de la REQM que $\hat{\lambda}$, quoique les REQM restent proches. En effet, en ce qui concerne l'estimation de T, ni l'une ni l'autre approche ne l'emporte pour ce qui est du biais empirique ou de la racine carrée de l'erreur quadratique moyenne, quelle que soit la taille de l'échantillon ou la valeur choisie de σ (données non présentées). Cela pourrait tenir au fait que l'estimateur de T est biaisé sous le modèle quelle que soit la qualité de l'estimation de λ.

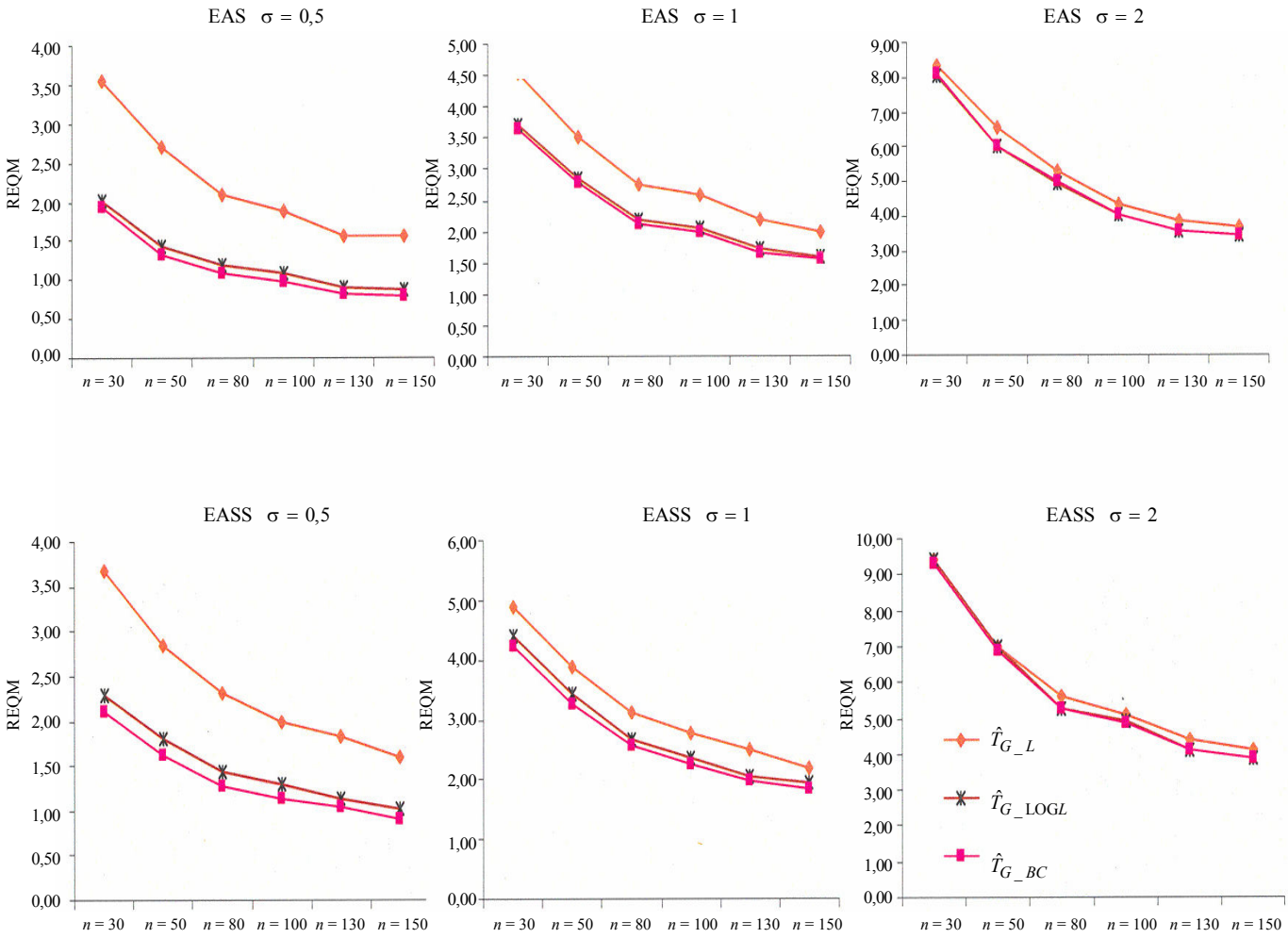

Figure 1 Comparaison de la racine carrée de l'erreur quadratique moyenne de $\hat{T}_{G_L}$, $\hat{T}_{G_LOGL}$ et $\hat{T}_{G_BC}$ pour divers plans d'échantillonnage, tailles d'échantillon et écarts-types

Tableau 2
Biais relatif et racine carrée de l'erreur quadratique moyenne de quatre estimateurs pour différents plans d'échantillonnage et diverses tailles d'échantillon ($\lambda = 1$)

		$\hat{T}_D^1$	$\hat{T}_{G_L}^2$	$\hat{T}_{G_LOGL}^3$	$\hat{T}_{G_BC}^4$
Échantillonnage aléatoire simple					
BiaisRel($\times 10^{-3}$)	$n = 30$	0,31	0,16	0,63	0,24
	$n = 80$	-0,37	-0,09	0,03	-0,09
	$n = 150$	-0,25	0,03	0,10	0,02
REQM($\times 10^{7}$)	$n = 30$	20,91	3,51	3,98	3,63
	$n = 80$	12,22	2,12	2,36	2,13
	$n = 150$	8,98	1,57	1,75	1,57
Échantillonnage aléatoire simple stratifié					
BiaisRel($\times 10^{-3}$)	$n = 30$	0,51	0,27	0,56	0,22
	$n = 80$	3,97	-0,11	-0,11	-0,13
	$n = 150$	-0,23	0,04	0,06	0,04
REQM($\times 10^{7}$)	$n = 30$	12,54	3,79	4,20	3,91
	$n = 80$	8,39	2,27	2,61	2,29
	$n = 150$	5,48	1,67	1,90	1,67

[1] $\hat{T}_D$: estimateur d'Horvitz-Thompson fondé sur le plan;
[2] $\hat{T}_{G_L}$: estimateur GREG avec modèle linéaire sous-jacent;
[3] $\hat{T}_{G_LOGL}$: estimateur GREG avec modèle loglinéaire sous-jacent;
[4] $\hat{T}_{G_BC}$: estimateur GREG avec modèle de Box-Cox sous-jacent.

Tableau 3
Biais relatif et racine carrée de l'erreur quadratique moyenne de quatre estimateurs pour différents plans d'échantillonnage et diverses tailles d'échantillon (valeurs de y produites à partir d'un modèle M_6)

		$\hat{T}_D^1$	$\hat{T}_{G_L}^2$	$\hat{T}_{G_LOGL}^3$	$\hat{T}_{G_BC}^4$
Échantillonnage aléatoire simple					
BiaisRel($\times 10^{-3}$)	$n = 30$	-16,89	-54,56	28,81	-2,17
	$n = 80$	-5,65	-23,15	11,76	-1,43
	$n = 150$	-13,78	-13,78	10,73	-0,76
REQM ($\times 10^{11}$)	$n = 30$	30,08	24,68	12,11	2,87
	$n = 80$	17,95	13,85	7,25	1,78
	$n = 150$	13,60	10,29	5,50	1,33
Échantillonnage aléatoire simple stratifié					
BiaisRel($\times 10^{-3}$)	$n = 30$	1,11	-36,13	31,95	-7,87
	$n = 80$	5,59	-18,13	13,11	-3,26
	$n = 150$	-2,79	-7,30	7,33	-1,43
REQM($\times 10^{11}$)	$n = 30$	34,10	27,15	14,19	4,37
	$n = 80$	19,61	15,93	8,49	2,61
	$n = 150$	14,60	12,19	6,62	2,02

[1] $\hat{T}_D$: estimateur d'Horvitz-Thompson fondé sur le plan;
[2] $\hat{T}_{G_L}$: estimateur GREG avec modèle linéaire sous-jacent;
[3] $\hat{T}_{G_LOGL}$: estimateur GREG avec modèle loglinéaire sous-jacent;
[4] $\hat{T}_{G_BC}$: estimateur GREG avec modèle de Box-Cox sous-jacent.

Tableau 4
Biais relatif et racine carrée de l'erreur quadratique moyenne de $\hat{\lambda}^1$ et $\hat{\lambda}_w^2$ sous échantillonnage EASS pour diverses tailles d'échantillon et divers écarts-types

	$\sigma = 2$		$\sigma = 1$		$\sigma = 0{,}5$	
	$\hat{\lambda}$	$\hat{\lambda}_w$	$\hat{\lambda}$	$\hat{\lambda}_w$	$\hat{\lambda}$	$\hat{\lambda}_w$
Biais relatif						
$n = 30$	-0,58	0,13	-0,28	0,10	-0,16	-0,01
$n = 80$	-0,42	0,14	-0,19	0,11	-0,13	-0,02
$n = 150$	-0,39	0,10	-0,16	0,09	-0,10	-0,01
Racine de l'erreur quadratique moyenne						
$n = 30$	0,14	0,12	0,11	0,11	0,07	0,07
$n = 80$	0,08	0,07	0,06	0,06	0,04	0,04
$n = 150$	0,06	0,05	0,05	0,04	0,03	0,03

[1] L'estimateur $\hat{\lambda}$ est obtenu par la méthode des moindres carrés ordinaires/méthode du maximum de vraisemblance.
[2] L'estimateur $\hat{\lambda}_w$ est obtenu par la méthode du pseudo maximum de vraisemblance.

5. Conclusion

Nous avons proposé un estimateur par la régression généralisée d'un total de population finie fondé sur une technique de transformation de Box-Cox sous un plan d'échantillonnage à probabilités inégales. Étant convergent par rapport au plan, l'estimateur proposé retient la propriété de robustesse de l'estimateur GREG, même si le modèle sous-jacent est défaillant. Dans de nombreuses situations, une version du modèle $\mathbf{M_3}$ fournira au moins une approximation utile du comportement de la variable dépendante. La technique de Box-Cox permet qu'une transformation raisonnablement ajustée de la variable dépendante soit automatiquement déterminée par les données. Nous avons évalué analytiquement et par simulation de Monte Carlo la robustesse et l'efficacité de l'estimateur proposé.

Lorsqu'on compare un estimateur GREG basé sur un modèle linéaire sous-jacent ($\hat{T}_{G_L}$) à un autre fondé sur un modèle de Box-Cox ($\hat{T}_{G_BC}$), il ne faut pas perdre de vue que $\hat{T}_{G_L}$ ne requiert pas d'information auxiliaire complète. En outre, il peut produire un jeu unique de poids applicable à toutes les variables d'intérêt, contrairement à l'estimateur $\hat{T}_{G_BC}$. Toutefois, pour accroître l'efficacité de l'un et l'autre estimateur, il faut habituellement utiliser des poids distincts pour des variables d'intérêt différentes, parce les meilleures prédictions sont obtenues en ajustant un modèle de travail particulier pour chaque variable étudiée. L'estimateur $\hat{T}_{G_BC}$ peut même être plus efficace que l'estimateur $\hat{T}_{G_L}$, mais au prix de nécessiter des données complètes sur les variables x. Bien que ce genre d'information soit rarement disponible dans les sondages nord-américains auprès des ménages, elle l'est souvent dans les sondages auprès des entreprises.

Les sondages sont peu souvent réalisés pour mesurer une seule variable d'intérêt. La question est de savoir comment estimer les totaux finis pour des sous-population mutuellement exclusives et exhaustives de telle façon que la somme de ces estimations concorde avec l'estimation du total fini pour l'ensemble de la population. Nous devons accorder une attention particulière à ce problème, car les estimateurs de Box-Cox ne sont pas linéaires en y. Ces estimations peuvent être obtenues en utilisant un outil d'étalonnage standard décrit dans Li et Lahiri (2007). Une autre approche consiste à utiliser le calage fondé sur un modèle (Wu et Sitter 2001). Le fait de traiter la prédiction $\hat{y}_{i,w}$ tirée de l'équation (3) comme la variable auxiliaire dans un estimateur par la régression linéaire convergent par rapport au plan, de sorte que

$$\hat{T}_{G_BC} = \sum_{i\in s} y_i/\pi_i + \left(\sum_{i\in U}\mathbf{z}_i' - \sum_{i\in s}\mathbf{z}_i'/\pi_i\right)\mathbf{d},$$

où

$$\mathbf{d} = \left(\sum_{i\in s} w_i\,\mathbf{z}_i\,\mathbf{z}_i'\right)^{-1}\left(\sum_{i\in s} w_i\,\mathbf{z}_i\,y_i\right), \text{ et } \mathbf{z}_i = (1\ \ \hat{y}_{i,w})',$$

produira un ensemble de poids de calage

$$\left(w_{i,c} = \left[1 + \left(\sum_U \mathbf{z}_i' - \sum_S \mathbf{z}_i'/\pi_i\right)\left(\sum_S w_j\,\mathbf{z}_j\,\mathbf{z}_j'\right)^{-1}\mathbf{z}_i\right]w_i\right)$$

utilisables de façon générale. En outre, nous pouvons, en principe, intégrer plus d'un ensemble de prédicteurs comme variables auxiliaires, pour diverses variables d'intérêt ou pour la même variable d'intérêt décomposée en plusieurs sous-populations. Il s'agit là d'un domaine prometteur pour de futurs travaux de recherche.

Plusieurs autres extensions de notre méthode courante méritent d'être explorées plus en détail. Nous n'avons pas considéré ici la possibilité que les erreurs de modélisation unitaires présentent une structure de corrélation complexe. Quand une telle structure existe mais est ignorée, notre méthode produit un estimateur qui est convergent par rapport au plan, mais dont l'efficacité souffre vraisemblablement. Il sera intéressant de voir si la procédure d'estimation peut être rendue plus efficace en permettant certaines structures de corrélation dans les données.

Dans le présent article, nous n'avons transformé que la variable y par la technique de Box-Cox. Dans l'avenir, notre méthode pourrait être étendue à diverses formes fonctionnelles de la transformation de Box-Cox.

L'estimateur de variance de l'estimateur $\hat{T}_{G_BC}$ proposé ici est fondé sur une approximation en grand échantillon. Certaines études ont révélé des propriétés médiocres pour ce type d'estimateur de variance sous certains choix des modèles auxiliaires pour l'estimateur $\hat{T}_{G_L}$. Les études théoriques et empiriques de la propriété de couverture de l'estimateur de variance proposé doivent se poursuivre.

Remerciements

Je remercie le rédacteur en chef et trois examinateurs dont plusieurs suggestions constructives m'ont permis d'améliorer considérablement le manuscrit original. Je remercie M. Phil Kott de sa révision copieuse du texte et M. Alan Dorfman d'avoir fourni les données d'enquête utilisées dans l'étude. Je tiens aussi à exprimer ma reconnaissance au professeur Partha Lahiri, mon directeur de thèse de doctorat, et aux professeurs Katherine Abraham, Wolfgang Jank, Fritz Scheuren et Paul Smith, et à M. Paul D. Williams, membres du comité de supervision, pour leur encadrement généreux tout au long de mes travaux de recherche de thèse.

Annexe

Soit

$$\hat{T}_{G_BC} = \sum_{i\in s} y_i/\pi_i + \left(\sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) - \sum_{i\in s} g_i(\hat{\boldsymbol{\theta}}_w)/\pi_i\right).$$

Par développement en série de Taylor, sous l'hypothèse (2), nous avons

$$g_i(\hat{\boldsymbol{\theta}}_w) = g_i(\boldsymbol{\theta}_N) + \left(\partial g_i(\mathbf{t})/\partial \mathbf{t}\right)|_{t=\boldsymbol{\theta}^*}\ (\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N),$$

où $\boldsymbol{\theta}^* \in (\boldsymbol{\theta}_N, \hat{\boldsymbol{\theta}}_w)$ ou $(\hat{\boldsymbol{\theta}}_w, \boldsymbol{\theta}_N)$ et $\left(\partial g_i(\mathbf{t})/\partial \mathbf{t}\right)$ est un vecteur ligne.

En vertu des hypothèses (1) et (2),

$$N^{-1}\sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) = N^{-1}\sum_{i\in U} g_i(\boldsymbol{\theta}_N) + O_p(1/\sqrt{n})$$

$$N^{-1}\sum_{i\in s} \pi_i^{-1} g_i(\hat{\boldsymbol{\theta}}_w) = N^{-1}\sum_{i\in s} \pi_i^{-1} g_i(\boldsymbol{\theta}_N) + O_p(1/\sqrt{n}).$$

Notons aussi qu'en vertu de l'hypothèse (3),

$$N^{-1}\sum_{i\in U} g_i(\boldsymbol{\theta}_N) = N^{-1}\sum_{i\in s} \pi_i^{-1} g_i(\boldsymbol{\theta}_N) + O_p(1/\sqrt{n}).$$

Donc,

$$N^{-1}\left(\sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) - \sum_{i\in s} \pi_i^{-1} g_i(\hat{\boldsymbol{\theta}}_w)\right) = O_p(1/\sqrt{n}).$$

En outre, en vertu de l'hypothèse (3),

$$N^{-1}\left(\sum_{i\in U} y_i - \sum_{i\in s} \pi_i^{-1} y_i\right) = O_p(1/\sqrt{n})\ .$$

Par conséquent, $N^{-1}(\hat{T}_{G_BC} - T) = O_p(1/\sqrt{n})$, *i.e.*, $\hat{T}_{G_BC}$ converge en probabilité vers T à l'ordre $O_p(1/\sqrt{n})$.

En plus, sous l'hypothèse (4), une approximation par série de Taylor de deuxième ordre de $g_i(\hat{\boldsymbol{\theta}}_w)$ peut être développée comme il suit :

$$g_i(\hat{\boldsymbol{\theta}}_w) = g_i(\boldsymbol{\theta}_N) + (\partial g_i(\mathbf{t})/\partial \mathbf{t})'\,|_{t=\boldsymbol{\theta}_N}$$
$$(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N) + (\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N)'\left(\partial^2 g_i(\mathbf{t})/\left(\partial \mathbf{t}\partial \mathbf{t}'\right)\right)|_{t=\boldsymbol{\theta}^*}\ (\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N),$$

où $\boldsymbol{\theta}^* \in (\boldsymbol{\theta}_N, \hat{\boldsymbol{\theta}}_w)$ ou $(\hat{\boldsymbol{\theta}}_w, \boldsymbol{\theta}_N)$. Il découle des hypothèses (1) et (4) que

$$N^{-1}\sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) = N^{-1}\sum_{i\in U} g_i(\boldsymbol{\theta}_N)$$
$$+ N^{-1}\sum_{i\in U}\left(\partial g_i(\mathbf{t})/\partial \mathbf{t}'\right)|_{t=\boldsymbol{\theta}_N}\ (\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N) + O_p(n^{-1}),$$

$$N^{-1}\sum_{i\in s} \pi_i^{-1} g_i(\hat{\boldsymbol{\theta}}_w) = N^{-1}\sum_{i\in s} \pi_i^{-1} g_i(\boldsymbol{\theta}_N)$$
$$+ N^{-1}\sum_{i\in s} \pi_i^{-1}\left(\partial g_i(\mathbf{t})/\partial \mathbf{t}'\right)|_{t=\boldsymbol{\theta}_N}\ (\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_N) + O_p(n^{-1}).$$

En vertu des hypothèses (1) et (3), $\hat{\boldsymbol{\theta}}_w = \boldsymbol{\theta}_N + O_p(1/\sqrt{n})$ et

$$N^{-1}\sum_{i\in s} \pi_i^{-1}\partial g_i(\mathbf{t})/\partial \mathbf{t}\,|_{t=\boldsymbol{\theta}_N} = N^{-1}\sum_{i\in U} \partial g_i(\mathbf{t})/\partial \mathbf{t}\,|_{t=\boldsymbol{\theta}_N} + O_p(1/\sqrt{n}).$$

Donc,

$$N^{-1}\left(\sum_{i\in U} g_i(\hat{\boldsymbol{\theta}}_w) - \sum_{i\in s} \pi_i^{-1} g_i(\hat{\boldsymbol{\theta}}_w)\right) =$$
$$N^{-1}\left(\sum_{i\in U} g_i(\boldsymbol{\theta}_N) - \sum_{i\in s} \pi_i^{-1} g_i(\boldsymbol{\theta}_N)\right) + O_p(n^{-1}).$$

Par conséquent,

$$\hat{T}_{G_BC} = \sum_{i\in s} y_i/\pi_i$$
$$+ \left(\sum_{i\in U} g_i(\boldsymbol{\theta}_N) - \sum_{i\in s} \pi_i^{-1} g_i(\boldsymbol{\theta}_N)\right) + O_p(N/n).$$

La variance asymptotique de $\hat{T}_{AG}$ sous le plan est :

$$AV_d(\hat{T}_{G_BC})$$
$$\approx \sum_{i\in U}\sum_{j\in U} (\pi_{ij} - \pi_i\pi_j)(y_i - g_i(\boldsymbol{\theta}_N))(y_j - g_j(\boldsymbol{\theta}_N))/(\pi_i\pi_j),$$

qui peut être estimée par

$$\hat{V}_d(\hat{T}_{G_BC})$$
$$\approx \sum_{i\in s}\sum_{j\in s} (\pi_{ij} - \pi_i\pi_j)/\pi_{ij}(y_i - g_i(\hat{\boldsymbol{\theta}}_w))(y_j - g_j(\hat{\boldsymbol{\theta}}_w))/(\pi_i\pi_j).$$

Bibliographie

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex survey. *Revue Internationale de Statistique*, 51, 279-92.

Bickel, P.J., et Doksum, K.A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76, 296-311.

Box, G.E., et Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Séries B, 26, 211-252.

Boylan, T.A., Cuddy, M.P. et O'Muircheartaigh, I.G. (1980). The functional form of the aggregate import demand equation: A comparison of three european economies. *Journal of international economics*, 10, 561-566.

Boylan, T.A., Cuddy, M.P. et O'Muircheartaigh, I.G. (1982). Import demand equations: An application of a generalized Box-Cox methodology. *Revue Internationale de Statistique*, 50, 103-112.

Breidt, F.J., Claeskens, G. et Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92, 831-846.

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

Carrol, R.J., et Ruppert, D. (1988). *Transformations and weighting in Regression.* Londres : Chapman and Hall.

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

Chambers, R.L., et Dorfman, A.H. (2003). Transformed variables in survey sampling. Rapport technique.

Chen, G., et Chen, J. (1996). Une méthode de transformation applicable à l'échantillonnage de populations finies calée par une méthode de vraisemblance empirique. *Techniques d'enquête*, 22, 139-147.

Davison, C.W., Arnade, C.A. et Hallahan, C.B. (1989). Box-Cox estimation of U.S. soyabean exports. *Journal of Agricultural Economics Research*, 41, 8-15.

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

Estevao, V., Hidiroglou, M. et Särndal, C.-E. (1995). Methodological principles for a generalized estimation system at statistics Canada. *Journal of Official Statistics*, 11, 181-204.

Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5–25.

Fuller, W.A., Loughin, M. et Baker, H. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.

Gemmil, G. (1980). Using the Box-Cox Form for Estimating Demand: A Comment. *The review of economics and statistics*, 62, 147-148.

Gurka, M. (2004). The Box-Cox transformation in the general linear mixed model for longitudinal data. Ph.D. Dissertation.

Gurka, M. (2006). Expanding the Box-Cox transformation to the linear mixed model. *Journal of the Royal Statistical Society*, Séries A, 169, 273-288.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite population. *Journal of American Statistical Association*, 47, 663-685.

Jayasuriya, B., et Valliant, R. (1996). Application de l'estimation par régression restreinte dans une enquête-ménage. *Techniques d'enquête*, 22, 127-138.

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

John, N.R., et Draper, J.A. (1980). An alternative family of transformations. *Applied Statistics*, 29, 190–197.

Karlberg, F. (2000). Survey estimation for highly skewed population in the presence of zeroes. *Journal of Official Statistics*, 16, 229-241.

Khan, M.S., et Ross, K.Z. (1977). The functional form of aggregate import demand. *Journal of International Economics*, 7, 149-160.

Korn, E., et Graubard, B. (1998). Intervalles de confiance pour les proportions à petit nombre d'événements positifs prévus estimées au moyen des données d'enquête. *Techniques d'enquête*, 24, 209-218.

Li, Y., et Lahiri, P. (2007). Robust model-based and model-assisted predictors of the finite population total. *Journal of the American Statistical Association*, 102, 664-673.

Miner, A.G. (1982). The contribution of weather and technology to U.S. soybean yield. *Unpublished Dissertation*. University of Minnesota.

Montanari, G.E., et Ranalli, M.G. (2003). Nonparametric model calibration estimation in survey sampling. Manuscript.

Montanari, G.E., et Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, 1429–1442.

Newman, P. (1977). Malaria and Mortality. *Journal of the American Statistical Association*, 72, 257-263.

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

Sakia, R.M. (1992). The Box-Cox transformation technique: A review. *The Statistician*, 41, 169-178.

Särndal, C.-E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

Särndal, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Plan Inference*, 7, 155-170.

Särndal, C.-E., Swensson, B. et Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

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

Schlesselman, J. (1971). Power families: A note on the Box and Cox transformation. *Journal of the Royal Statistical Society*, Séries B, 33, 307-311.

Skinner, C.J., Holt, D. et Smith, T.M.F. (Éds.) (1989). *Analysis of complex surveys.* Chichester : Wiley.

Spitzer, J.J. (1976). The demand for money, the liquidity trap and functional forms. *The International Economics Review*, 17, 220-227.

Sverchkov, M., et Pfeffermann, D. (2004). Prévision des totaux de population finie basée sur la distribution échantillonnale. *Techniques d'enquête*, 30, 87-101.

Taylor, M.J.G. (1986). The retransformed mean after a fitted power transformation. *Journal of the American Statistical Association*, 81, 114-118.

Tukey, J.W. (1957). The comparative anatomy of transformation. *Annals of Mathematical Statistics*, 28, 601-632.

Wu, C. (1999). The effective use of complete auxiliary information from survey data. Ph.D. dissertation, Simon Fraser University, Canada.

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

Zarembka, P. (1968). Functional form in the demand for money. *Journal of the American Statistical Association*, 63, 502-511.

Zheng, H., et Little, R.J.A. (2004). Modèles non paramétriques mixtes à fonction spline pénalisée pour l'inférence au sujet d'une moyenne de population finie d'après des échantillons à deux degrés. *Techniques d'enquête*, 30, 233-243.

Techniques d'enquête, juin 2008
Vol. 34, N° 1, pp. 99-113
Statistique Canada, N° 12-001-X au catalogue

L'algorithme BACON-EEM pour la détection d'observations aberrantes multivariées dans des données d'enquête incomplètes

Cédric Béguin et Beat Hulliger [1]

Résumé

Dans le cas de données multivariées complètes, l'algorithme BACON (Billor, Hadi et Vellemann 2000) donne une estimation robuste de la matrice de covariance. La distance de Mahalanobis correspondante peut être utilisée pour la détection des observations aberrantes multivariées. Quand des items manquent, l'algorithme EM est un moyen commode d'estimer la matrice de covariance à chaque étape d'itération de l'algorithme BACON. Dans l'échantillonnage en population finie, l'algorithme EM doit être amélioré pour estimer la matrice de covariance de la population plutôt que de l'échantillon. Une version de l'algorithme EM pour données d'enquête suivant un modèle normal multivarié, appelée algorithme EEM (espérance estimée/maximisation), est proposée. La combinaison des deux algorithmes, dénommée algorithme BACON-EEM, est appliquée à deux ensembles de données et comparée à d'autres méthodes.

Mots clés : Méthode de recherche avant (forward search); détection des observations aberrantes; données multivariées; valeurs manquantes; échantillonnage; robustesse; algorithme EM.

1. Introduction

Le problème qui sous-tend les méthodes présentées ici est celui d'une enquête par sondage portant sur des données quantitatives, comme les ventes de divers produits, pour lesquelles des valeurs manquent et des observations aberrantes existent. Souvent, à la phase de vérification du sondage, les observations aberrantes sont décelées par inspection des questionnaires individuels ou par des méthodes de détection des observations aberrantes univariées. Toutefois, il existe peu de méthodes systématiques permettant de déceler des observations aberrantes multivariées dans des données d'enquête incomplètes.

La détection des observations aberrantes est un aspect important de la vérification des données statistiques. Celles qui ne sont pas décelées peuvent avoir un effet important et indésirable sur les résultats de l'enquête. La plupart des méthodes existantes de détection des observations aberrantes sont conçues pour être appliquées à des données complètes univariées ou bivariées. Toutefois, les observations aberrantes réelles dans les données d'enquête sont souvent de nature multivariée. Le problème qu'elles posent devient nettement plus difficile à résoudre en trois dimensions ou plus qu'en une ou en deux dimensions. Dans le contexte unidimensionnel, les observations aberrantes ne peuvent être qu'extrêmement petites ou extrêmement grandes (du moins pour les distributions unimodales), mais, dans le cas d'une dimensionnalité plus élevée, la question de la « direction » de l'observation aberrante devient de plus en plus importante. Les observations aberrantes peuvent être relativement proches de la majorité des données ou d'un modèle si la distance est mesurée dans une métrique euclidienne, parce que cette dernière ne vérifie que les directions des axes. Par contre, si l'on utilise une métrique appropriée à la structure de corrélation de la majorité des données, l'observation aberrante peut être éloignée. Donc, pour les dimensionnalités plus élevées, la forme du nuage de points de la majorité des données doit être bien reflétée par la métrique utilisée pour pouvoir déceler les observations aberrantes.

La détection des observations aberrantes nécessite un modèle pour la majorité des données afin de pouvoir distinguer les observations auxquelles le modèle n'est pas bien ajusté. Donc, la détection des observations aberrantes est intrinsèquement liée aux modèles et à leur estimation robuste. Dans un contexte d'échantillonnage, le modèle devrait être approprié à la majorité de la population plutôt que seulement à la majorité de l'échantillon. Par conséquent, le plan d'échantillonnage devrait être pris en compte pour la détection des observations aberrantes dans des données d'enquête par sondage. Nous revenons à la discussion du rôle des poids d'échantillonnage aux sections 1 et 5.

Les données de sondage contiennent souvent des observations manquantes. La détection des observations aberrantes quand les réponses à certains items manquent doit comporter l'estimation du modèle pour la majorité des données en tenant compte de ces données manquantes. L'estimation sous valeurs manquantes sera fondée sur la relation entre les variables observées et les variables manquantes. Cette relation doit être modélisée robustement afin de la protéger contre les observations aberrantes. Si une observation est classée comme étant aberrante en se basant sur l'information complète, mais que les valeurs causant l'aberrance manquent, l'observation aberrante n'apparaîtra

1. Cédric Béguin, Université de Neuchâtel, 2010 Neuchâtel, Suisse; Beat Hulliger, Université des sciences appliquées de la Suisse du Nord-Ouest, 4600 Olten, Suisse. Courriel : beat.hulliger@fhnw.ch.

pas comparativement à un modèle robuste. Donc, il sera difficile de déceler une valeur aberrante qui ne l'est que dans ses valeurs manquantes. Cette situation est analogue au concept de valeur manquant au hasard (MAR, pour *missingness at random*) (Little et Rubin 1987). Nous avons besoin d'information dans les valeurs observées pour inférer qu'une observation est aberrante. Nous pourrions appeler cette situation « aberrance au hasard ». Nous pouvons l'exprimer formellement en disant que le mécanisme de création d'observations aberrantes ne dépend pas des données non observées, qui comprennent les valeurs réelles non observées de l'observation aberrante au cas où l'observation aberrante serait une erreur. Cependant, pour la détection des observations aberrantes, cette condition est trop stricte, parce que nous pourrions arriver à détecter des observations aberrantes dans des valeurs observées, même si le mécanisme dépend des valeurs non observées. Cela est possible, parce que le modèle doit tenir pour la majorité des données seulement et non pour les observations aberrantes. Si les valeurs observées d'une observation aberrante s'écartent suffisamment du modèle, cette dernière est détectée. Cependant, en ce qui concerne l'imputation des valeurs réelles pour les observations désignées aberrantes, nous sommes dans la même situation que pour les valeurs manquantes. Si le mécanisme de création des observations aberrantes, sachant les valeurs observées, dépend encore des valeurs réelles non observées de l'observation aberrante, nous ne pouvons pas estimer un modèle pour les valeurs non observées. Dans le présent article, nous utilisons l'imputation uniquement comme un outil ponctuel pour améliorer la détection des observations aberrantes. Néanmoins, nous supposons que, sachant les données observées, le mécanisme de non-réponse et le mécanisme de création d'observations aberrantes sont indépendants et que ni l'un ni l'autre ne dépend des données inobservées.

Dans un progiciel de travail du projet EUREDIT sur le développement et l'évaluation de nouvelles méthodes de vérification et d'imputation (EUREDIT 2003), les auteurs ont développé une méthode de détection des observations aberrantes pour traiter la situation difficile que représentent les données de sondage incomplètes multivariées. Deux de ces méthodes, appelées Transformed Rank Correlations et Epidemic Algorithm, sont présentées dans Béguin et Hulliger (2004). La troisième méthode, BACON-EEM, est présentée ici.

Dans le présent article, nous nous concentrons sur la détection des observations aberrantes. Selon notre scénario, une fois qu'une observation aberrante est décelée, elle peut être vérifiée et traitée manuellement ou elle peut être traitée par imputation. L'estimation robuste remplacerait à la fois la détection et l'imputation, mais elle est moins bien adaptée à la pratique des statistiques officielles. Nous ne faisons pas la distinction entre les cas aberrants représentatifs et non représentatifs (Chambers 1986), puisque les uns et les autres doivent être décelés, bien qu'ils doivent éventuellement être traités différemment.

Dans le cas des données complètes, les méthodes multivariées existantes peuvent être classées en deux grandes familles. Les auteurs de nombreuses méthodes supposent que les données suivent une certaine loi elliptique et essayent d'estimer robustement le centre et la matrice de covariance. Puis, ils utilisent une distance de Mahalanobis correspondante pour détecter les observations aberrantes. La deuxième classe de méthodes ne s'appuie pas sur une hypothèse distributionnelle, mais utilise une mesure de profondeur des données (voir Liu, Parelius et Singh 1999 pour une revue) qui sert de mesure d'aberrance. La deuxième famille est, à première vue, plus séduisante, mais malheureusement, elle produit souvent des méthodes dont les calculs ne sont pas faisables avec de grands ensembles de données.

De nombreux estimateurs robustes de la matrice de covariance ont été publiés dans la littérature. Les estimations M (Huber 1981; Maronna 1976) ont l'avantage d'être relativement faciles à calculer par une itération simple à partir d'un bon point de départ (Rocke et Woodruff 1993). Toutefois, leur point de défaillance, c'est-à-dire la plus petite fraction des données dont la modification arbitraire peut amener un estimateur au-delà de toutes les bornes, est, au mieux $1/(p+1)$, où p est la dimension des données (Donoho 1982; Maronna 1976; Stahel 1981). Ce handicap est important dans le cas des statistiques officielles, qui sont souvent de dimensionnalité élevée. De nombreux autres estimateurs robustes équivariants affines, c'est-à-dire des estimateurs qui se transforment de manière cohérente quand les données sont transformées linéairement, ont été étudiés par Donoho (1982), mais tous ont un point de défaillance correspondant, au mieux, à $1/(p+1)$. D'autres approches ont abouti à des estimateurs équivariants affines à point de défaillance élevé, par exemple l'estimateur de Stahel-Donoho (SD) (Stahel 1981; Donoho 1982) ou les estimateurs par le déterminant minimal de covariance (MCD, pour *Minimum Covariance Determinant*) (Rousseeuw 1985; Rousseeuw et Leroy 1987), mais elles ont l'inconvénient d'être informatiquement coûteuses. Une approche de Gnanadesikan et Kettenring s'appuyant sur une construction de la matrice de covariance sacrifie l'équivariance affine, mais gagne en simplicité et en vitesse. Cette approche a été réactualisée dans Maronna et Zamar (2002) et dans l'une des méthodes présentées dans Béguin et Hulliger (2004), appelée corrélations de rang transformées (TRC pour *Transformed Rank Correlations*). La méthode TRC consiste à calculer une matrice initiale de corrélations de rang de Spearman bivariées. Pour s'assurer que la

matrice de covariance est définie positive, les données sont transformées dans l'espace des vecteurs propres de la matrice initiale. Les médianes des coordonnées et les écarts absolus par rapport à la médiane dans ce nouvel espace sont alors rétrotransformés dans l'espace original pour obtenir une estimation du centre et une matrice de covariance définie positive.

Une autre idée de Gnanadesikan et Kettenring (1972) est reliée aux méthodes dites de recherche avant (*forward search*), qui sont étroitement apparentées à la méthode proposée ici. Ces méthodes sont fondées sur le concept consistant à « faire pousser » un bon sous-ensemble d'observations. Par « bon sous-ensemble », il est entendu un sous-ensemble exempt ou presque exempt d'observations aberrantes. L'idée est de partir d'un petit sous-ensemble de données, puis d'ajouter des observations non aberrantes jusqu'à ce qu'il n'y en ait plus aucune disponible.

La notion d'un algorithme de recherche avant a été proposée pour la première fois dans Wilks et Gnanadesikan (1964) et décrite en détail dans Gnanadesikan et Kettenring (1972). Les articles de Hadi (1992) et d'Atkinson (1993) ont démontré l'efficacité de ce genre de méthodes. Dans les deux articles, le « bon sous-ensemble » est agrandi un point à la fois en utilisant les distances de Mahalanobis pour classer les observations. Par la suite, les travaux de recherche ont porté surtout sur l'élaboration de méthodes plus rapides et plus perfectionnées fondées sur la même idée. Les deux dernières, qui sont les plus efficaces, ont été élaborées par Billor, Hadi et Vellemann (2000) et Kosinski (1999). Ces algorithmes ont été comparés par Béguin (2002) et l'algorithme BACON (Billor et coll. 2000) s'est avéré être la méthode de recherche avant la plus robuste et la plus rapide dans le cas de données normales multivariées complètes. En particulier, le point de défaillance s'est révélé très élevé dans les applications pratiques. En outre, comparativement à d'autres méthodes de type Mahalanobis, la performance de l'algorithme BACON sur des données complètes est très bonne (Béguin et Hulliger 2003).

Aucune des méthodes susmentionnées n'est conçue pour traiter des données incomplètes issues de sondages, c'est-à-dire avec valeurs manquantes et poids de sondage. Le premier article dans lequel a été abordé le problème de la détection d'observations aberrantes multivariées dans des données incomplètes est celui de Little et Smith (1987). Ces auteurs proposent des distances de Mahalanobis pour détecter les observations aberrantes, et obtiennent des estimations du centre et de la dispersion par l'algorithme ER, dans lequel l'estimateur du maximum de vraisemblance de l'étape de maximisation de l'algorithme EM (Dempster, Laird et Rubin 1977) est remplacé par un M-estimateur robuste en une étape. Cependant, l'algorithme ER a pour point de départ la moyenne et la covariance non robustes classiques et, par conséquent, son point de défaillance est 0. Autrement dit, même une seule observation aberrante peut porter l'estimateur au-delà de toute limite. Pour corriger le faible point de défaillance de cet algorithme, Cheng et Victoria-Feser (2000) ont utilisé un algorithme MCD pour l'étape de maximisation de l'algorithme EM. Cependant, la combinaison des procédures itératives du MCD et de l'EM rend les calculs trop lents pour les applications pratiques quand les ensembles de données sont grands. En outre, l'introduction des poids d'échantillonnage n'est pas simple.

L'algorithme TRC de Béguin et Hulliger (2004) utilise des imputations par régression linéaire robuste par le meilleur prédicteur univarié pour traiter les valeurs manquantes. Les corrélations de rang de Spearman sont exprimées comme des fonctionnelles de la fonction de répartition empirique de l'échantillon pour obtenir des estimations des corrélations de rang de Spearman dans la population.

Comme l'algorithme BACON est fondé sur la loi normale multivariée, nous avons choisi l'algorithme EM pour données normales multivariées pour imputer les valeurs manquantes dans les itérations BACON. Afin de tenir compte de l'échantillonnage, nous avons remplacé les estimations de l'algorithme BACON par des estimateurs de type Horvitz-Thompson et nous avons élaboré une version spéciale de l'algorithme EM où les espérances au niveau de la population sont estimées d'après l'échantillon. À la section 2, nous décrivons la notation, rappelons rapidement en quoi consiste l'algorithme BACON et présentons son adaptation aux poids d'échantillonnage. À la section 3, nous introduisons l'algorithme d'espérance estimée/maximisation (EEM) et à la section 1, nous discutons de l'adaptation de la distance de Mahalanobis aux valeurs manquantes. À la section 4, nous expliquons comment les algorithmes BACON et EEM sont fusionnés de manière efficace pour devenir l'algorithme BACON-EEM. À la section 5, nous montrons l'application de l'algorithme BACON-EEM à deux ensembles de données. Nous comparons les résultats aux méthodes concurrentes, c'est-à-dire la méthode des corrélations de rang transformées (TRC) élaborées dans Béguin et Hulliger (2004), l'algorithme ER et un algorithme de référence utilisant le MCD après imputation non robuste fondée sur l'algorithme EM.

2. L'algorithme BACON

L'algorithme BACON est présenté dans Billor et coll. (2000). Deux versions sont décrites : l'une pour les données multivariées en général et l'autre pour les données de régression. Nous ne considérerons ici que le premier cas.

Les données sont stockées dans une matrice X de n lignes (observations $x_1, ..., x_n$) et de p colonnes (variables

$x^1, ..., x^p$). Nous supposons que la majorité d'entre-elles sont unimodales et approximativement elliptiques symétriques. La moyenne des coordonnées (matrice de covariance, resp.) calculée sur X est dénotée par m_X (C_X resp.). Le carré de la distance de Mahalanobis d'un point y basé sur m_X et C_X est $\mathrm{MD}_X^2(y) = (y - m_X)^\top \; C_X^{-1}(y - m_X)$. Si la moyenne et la covariance sont calculées uniquement sur un sous-ensemble G des données, nous les dénotons m_G et C_G avec la distance de Mahalanobis correspondante MD_G.

La première étape de l'algorithme est le choix d'un sous-ensemble initial G de « bonnes données ». Deux versions sont proposées dans la littérature. La première consiste simplement à sélectionner les cp points ayant les plus petites distances de Mahalanobis $\mathrm{MD}_X(x_i)$, $i \in \{1, ..., n\}$, où c est un entier choisi par l'analyste des données. Sa valeur peut être fixée à $c = 3$ par défaut. La deuxième version consiste à choisir les cp points ayant les plus petites distances euclidiennes par rapport à la médiane des coordonnées, avec c défini comme auparavant. La deuxième version est plus robuste, mais elle perd l'équivariance affine. D'autres points de départ que la médiane des coordonnées pourraient être envisagés, comme une médiane spatiale. Dans cet article, nous nous concentrons sur la deuxième version du bon sous-ensemble de base. Dans les deux versions, si C_G est singulière, le sous-ensemble de base est étendu en ajoutant les observations ayant les plus petites distances jusqu'à ce que C_G soit de plein rang. Après cela, le processus itératif commence.

Soit $\chi^2_{p,\beta}$ le $1-\beta$ centile de la loi χ^2 avec p degrés de liberté et $|G|$, le nombre d'éléments dans l'ensemble G. Les étapes de l'algorithme BACON sont :

1. Calculer les carrés des distances de Mahalanobis $\mathrm{MD}_G^2(x_i)$ pour $i \in \{1, ..., n\}$.
2. Définir un sous-ensemble G' incluant tous les points pour lesquels $\mathrm{MD}_G^2(x_i) < c_{npr}\, \chi^2_{p,\alpha/n}$, où $c_{npr} = c_{np} + c_{hr}$ est un facteur de correction avec $c_{np} = 1 + (p+1)/(n-p) + 1/(n-h-p)$, $c_{hr} = \max\{0, (h-r)/(h+r)\}$, $h = \lceil (n+p+1)/2 \rceil$ et $r = |G|$.
3. Si $G' = G$, arrêter; sinon, fixer G à G' et aller à l'étape 1.

Soulignons que le facteur de correction c_{npr} est proche de 1 si n est grand. Les observations qui ne sont pas contenues dans l'ensemble de données G final sont déclarées aberrantes. Une autre approche consiste à choisir pour la distance de Mahalanobis un seuil au-delà duquel les observations sont déclarées aberrantes en inspectant la distribution de la distance de Mahalanobis.

La complexité des calculs que requiert l'algorithme BACON dépend de la configuration des données. Comparativement à d'autres algorithmes, elle est faible et, en particulier, augmente lentement avec la taille de l'échantillon (voir aussi la section 5). La méthode BACON convient donc particulièrement bien pour les grands ensembles de données.

Notons que le critère de sélection original de l'étape 2 est conçu pour une loi normale multivariée, qui implique que les carrés des distances de Mahalanobis suivent asymptotiquement une loi χ^2 avec p degrés de liberté. Supposons que tous les points suivent une loi normale multivariée et que la distance de Mahalanobis est calculée en utilisant la moyenne d'échantillon et la matrice de covariance. Le test $\mathrm{MD}_X^2(x_i) > \chi^2_{p,\alpha}$ déclare qu'environ 100α pour cent des points sont des observations aberrantes. Au lieu d' α, nous utilisons souvent α/n. À l'aide des inégalités de Bonferroni, on peut montrer que, sous la loi normale, le test avec le niveau α/n n'aboutit à la déclaration d'aucune observation aberrante ayant une probabilité supérieure à $1-\alpha$ (c'est-à-dire $P(\mathrm{MD}_X^2(x_i) < \chi^2_{p,\alpha/n},\ \forall i \in \{1, ..., n\}) \geq 1-\alpha$). L'utilisation de α/n pour le test mène très rarement à la détection de points qui ne sont pas des observations aberrantes, mais elle réduit aussi la sensibilité aux observations aberrantes proches quand n devient grand. Il pourrait être utile d'exécuter la méthode avec les deux types de niveau de test et de comparer les résultats.

2.1 Adaptation aux poids d'échantillonnage

Dans le contexte de l'échantillonnage, nous utilisons la notation qui suit. Les données proviennent d'un échantillon aléatoire s de la population finie U contenant les N éléments. L'échantillon de taille n est tiré selon le plan d'échantillonnage $p(s)$ et les probabilités d'inclusion de premier ordre sont dénotées $\pi_i = \sum_{s|i \in s} p(s)$. Les poids seront les inverses des probabilités d'inclusion des observations $w_i = 1/\pi_i$, tels que l'estimateur d'Horvitz-Thompson du total de population, $\sum_{i \in U} x_i$, soit $\sum_s w_i x_i = \sum_{i=1}^n w_i x_i$. En outre, nous supposons que $\sum_s w_i \approx N$. La moyenne m_X et la matrice de covariance C_X peuvent être estimées par les estimateurs de Hájek

$$\hat{m}_X = \frac{\sum_s w_i x_i}{\sum_s w_i}$$

et (1)

$$\hat{C}_X = \frac{\sum_s w_i (x_i - \hat{m}_X)(x_i - \hat{m}_X)^\top}{\sum_s w_i}.$$

L'estimation d'échantillon de la médiane est définie comme dans Béguin et Hulliger (2004) : soit x_u^k la valeur la plus petite telle que $\sum_s w_i 1_{x \leq x_u^k}(x_i^k) \geq 0{,}5 \sum_s w_i$ et x_v^k, la valeur la plus petite telle que $\sum_s w_i 1_{x \leq x_v^k}(x_i^k) > 0{,}5 \sum_s w_i$; alors, l'estimation est donnée par

$$\widehat{\mathrm{med}}_X = (w_u x_u^k + w_v x_v^k)/(w_u + w_v). \quad (2)$$

Afin d'adapter l'algorithme BACON à l'échantillonnage, nous sélectionnons le sous-ensemble initial en utilisant les estimateurs de Hájek $\hat{m}_X$ et $\hat{C}_X$ ou la médiane $\widehat{\text{med}}_X$. Pour le processus itératif, nous dénotons par s_G les « bonnes observations » sélectionnées dans l'échantillon. Ces observations sont représentatives d'un « bon sous-ensemble virtuel » G de l'ensemble de la population dont la taille estimée est $\hat{r} = \sum_{s_G} w_i$. La moyenne et la matrice de covariance de ce sous-ensemble sont estimées par les estimateurs de Hájek

$$\hat{m}_G = \frac{\sum_{s_G} w_i x_i}{\sum_{s_G} w_i}$$

et (3)

$$\hat{C}_G = \frac{\sum_{s_G} w_i (x_i - \hat{m}_G)(x_i - \hat{m}_G)^\top}{\sum_{s_G} w_i}.$$

Nous utilisons ces estimations pour calculer les estimations des distances de Mahalanobis $\widehat{\text{MD}}_G(x_i), x_i \in s$. Enfin, nous calculons le facteur de correction $c_{Npr} = c_{Np} + c_{hr}$ du critère de sélection en utilisant les estimations $\hat{N} = \sum_s w_i$ et $\hat{r} = \sum_{s_G} w_i$. Si N est connu, nous utilisons sa valeur réelle.

S'il n'y a pas de valeurs manquantes dans les données, l'algorithme BACON peut être utilisé pour estimer la moyenne et la covariance de population. L'hypothèse fondamentale pour l'algorithme BACON est encore que la majorité des observations de la population suivent une loi elliptique. Nous pouvons utiliser l'algorithme BACON sans pondération et comparer les résultats à la version pondérée. L'obtention de résultats différents indique que les variables du plan ou le modèle utilisé pour la pondération de la non-réponse ne sont pas bien reflétés dans le modèle. Nous recommandons d'utiliser les poids, en particulier dans les applications de routine, afin d'offrir une certaine protection contre l'erreur de spécification du modèle. Quoiqu'il en soit, les paramètres estimés devraient être la moyenne et la covariance de la majorité de la population.

Soulignons que la distance de Mahalanobis ne fait pas intervenir directement les poids d'échantillonnage. Le poids d'une observation éventuellement aberrante n'influe sur la distance de Mahalanobis que par la voie du modèle, c'est-à-dire la moyenne et la covariance.

3. L'algorithme EEM

Les problèmes de non-réponse sont importants en statistique officielle et de nombreux sondages ne peuvent pas fournir un ensemble de données complètes. Le problème de la non-réponse totale, c'est-à-dire des observations dont toutes les valeurs manquent, est habituellement résolu par utilisation de pondérations appropriées et il n'est pas traité ici. La non-réponse partielle, c'est-à-dire les observations pour lesquelles une partie seulement de l'information est disponible, ne peut pas être traitée en écartant toutes les observations incomplètes, parce que trop d'information serait perdue. L'approche suivie ici demeure hautement efficace dans le cas de données normales multivariées. À chaque étape d'itération BACON, la moyenne et la matrice de covariance du bon sous-ensemble d'observations seront calculées en utilisant une version modifiée de l'algorithme EM pour données normales multivariées. Les espérances calculées à l'étape E seront remplacées par des estimations d'échantillon, d'où le nom EEM (espérance estimée/maximisation) de l'algorithme. Notons que nous présentons ici cette adaptation pour des données normales multivariées, mais que les résultats peuvent être généralisés à d'autres distributions de la famille exponentielle ordinaire.

Nous réutilisons dans ce paragraphe la description et la notation de l'algorithme EM donnée dans Schafer (2000). Tous les détails au sujet de l'algorithme non présentés ici peuvent être consultés dans les trois premiers chapitres et à la section 5.3 de ce livre. Nous recourons également ici à l'abus de notation suivant : le X dénotera simultanément une variable aléatoire de dimension p et la matrice de dimension $N \times p$ contenant les valeurs réalisées de la variable X de la population U. Si l'on procédait à un recensement de l'ensemble de la population pour mesurer la variable X, on obtiendrait un certain ensemble de valeurs observées et de valeurs manquantes $X = X_o \cup X_m$. L'algorithme EM repose sur l'hypothèse que le mécanisme de création des données manquantes est ignorable (Schafer 2000 section 2.2). Ici, nous supposons en outre que la présence de données manquantes est indépendante de l'échantillonnage. Les observations des données peuvent être modélisées comme étant des tirages indépendants et identiquement distribués (iid) à partir d'une loi de probabilité normale multivariée de densité $f(x, \theta)$. En utilisant les hypothèses et la factorisation $P(X \mid \theta) = P(X_o \mid \theta) P(X_m \mid X_o, \theta)$ nous pouvons écrire la log-vraisemblance sous données complètes sous la forme $l(\theta \mid X) = l(\theta \mid X_o) + \log(P(X_m \mid X_o, \theta)) + c$, où $l(\theta \mid X_o)$ est la log-vraisemblance sous les données observées et où c est une constante choisie arbitrairement. Le terme $P(X_m \mid X_o, \theta)$ reflète l'interdépendance de X_m et de θ dont tire parti l'algorithme EM. Comme $P(X_m \mid X_o, \theta)$ est inconnue, la moyenne de $l(\theta \mid X)$ sur $P(X_m \mid X_o, \theta^{(t)})$ est prise à chaque étape E, où $\theta^{(t)}$ est une estimation provisoire du paramètre inconnu. L'estimation suivante $\theta^{(t+1)}$ est trouvée en maximisant le résultat de l'étape de l'espérance (étape M). La séquence d'étapes E et M est itérée jusqu'à la convergence. Les conditions sous lesquelles cette séquence $\theta^{(t)}$ converge vers un point stationnaire de la vraisemblance

sous les données observées sont décrites dans Dempster et coll. (1977). Dans le cas de problèmes « dociles », ce point stationnaire est un maximum global.

Pour une loi de probabilité appartenant à la famille exponentielle ordinaire, la log-vraisemblance sous données complètes peut s'écrire

$$l(\theta \mid X) = \eta(\theta)^{\top} \cdot T(X) + Ng(\theta) + c, \qquad (4)$$

où $\eta(\theta) = (\eta_1(\theta), \eta_2(\theta), ..., \eta_k(\theta))^{\top}$ est la forme canonique du paramètre θ et $T(X) = (T_1(X),\ T_2(X), ..., T_k(X))^{\top}$ est le vecteur des statistiques suffisantes sous données complètes. En outre, chacune des statistiques suffisantes possède une forme additive $T_j(X) = \sum_{i=1}^{N} h_j(x_i)$ pour une certaine fonction h_j. Comme $l(\theta \mid X)$ est une fonction linéaire des statistiques suffisantes, dans l'étape E, $T_j(X)$ est remplacé par $E(T_j(X) \mid X_o, \theta^{(t)})$. Autrement dit, l'étape E comble les parties manquantes des statistiques suffisantes sous données complètes. Pour une loi normale multivariée $X = (X^1, \ldots, X^p)$, les statistiques suffisantes sont composées de deux types d'éléments : les sommes $\sum_{i=1}^{N} x_i^k$ et les sommes des produits $\sum_{i=1}^{N} x_i^k x_i^l$, $1 \le k, l \le p$. L'étape E se réduit au calcul des espérances conditionnelles de ces sommes sachant les données observées X_o et le paramètre provisoire $\theta^{(t)}$.

Pour un seul indice de sommation *i*, on peut montrer (Schafer 2000, section 5.3) que ces espérances dépendent uniquement des composantes observées de la même observation, c'est-à-dire de x_i^{obs}. Cela nous mène à

$$E\left(\sum_{i=1}^{N} x_i^k \,\middle|\, X_o, \theta^{(t)}\right) = \sum_{i=1}^{N} E(x_i^k \mid X_o, \theta^{(t)}) = \sum_{i=1}^{N} E(x_i^k \mid x_i^{\text{obs}}, \theta^{(t)}), 1 \le k \le p \qquad (5)$$

et à la forme analogue de la somme des produits. Naturellement, $E(x_i^k \mid x_i^{\text{obs}}, \theta^{(t)}) = x_i^k$ si $x_i^k \in x_i^{\text{obs}}$. Si x_i^k manque, alors cette espérance est la valeur ajustée d'une régression de x^k sachant le paramètre $\theta^{(t)}$ sur les variables qui sont observées pour l'observation *i*. Donc, les statistiques suffisantes sont composées des sommes de population des valeurs observées (T_o) et des sommes des valeurs ajustées (T_m).

Dans la situation où nos données proviennent d'un échantillon d'une population finie, nous considérons la population finie comme une réalisation de la loi normale multivariée, et les sommes (5) et les sommes des produits doivent être estimées d'après l'échantillon. La forme de (5) nous permet d'utiliser les simples estimateurs d'Horvitz-Thompson. L'estimation de (5) est

$$T^{k0} = \sum_{s} w_i E(x_i^k \mid x_i^{\text{obs}}, \theta^{(t)}),\ 1 \le k \le p, \qquad (6)$$

et $E(\sum_{i=1}^{N} x_i^k x_{i_s}^l \mid X_o, \theta^{(t)})$ est estimée par

$$T^{kl} = \sum_{s} w_i E(x_i^k x_i^l \mid x_i^{\text{obs}}, \theta^{(t)}),\ 1 \le k, l \le p. \qquad (7)$$

Brièvement, nous remplaçons les sommes de population de T_o et T_m par les estimateurs d'Horvitz-Thompson $\hat{T}_o$ et $\hat{T}_m$. Nous appelons le calcul de T^{k0} et de T^{kl} l'étape de l'espérance estimée (étape EE). En introduisant ces estimateurs dans l'expression (4), nous obtenons un estimateur de la fonction de vraisemblance de population moyenne.

Pour l'étape M, c'est-à-dire la maximisation de l'estimation de la vraisemblance de population moyenne, nous devons résoudre les équations normales pondérées. Nous obtenons la solution par une simple opération matricielle en appliquant l'opérateur SWEEP (Schafer 2000, section 5.3) à la matrice symétrique $(T^{kl})_{0 \le k, l \le p}$ de dimensions $(p+1) \times (p+1)$ des espérances estimées des statistiques suffisantes (avec T^{00} fixé à 1) divisée par *N*, qui est estimé par la somme des poids s'il est inconnu :

$$\theta^{(t+1)} = \text{SWP}[0]\left(\frac{(T^{kl})_{0 \le k, l \le p}}{\sum_s w_i}\right), \qquad (8)$$

où SWP[0] est l'opérateur SWEEP sur la première ligne/colonne de la matrice.

L'algorithme EEM itère les étapes EE et M. Du point de vue des calculs, la différence entre l'étape EE et l'étape E de l'algorithme EM original tient à l'utilisation de sommes pondérées au lieu de sommes non pondérées en ne faisant pas varier les poids au cours des itérations. Par conséquent, nous nous attendons à ce que la convergence de l'algorithme EEM reste semblable à celle de l'algorithme EM. Pour l'algorithme BACON-EEM, nous avons seulement besoin d'une approximation grossière de la solution à chaque étape BACON. Donc, nous n'utilisons qu'un petit nombre d'itérations de l'algorithme EEM.

3.1 Distance de Mahalanobis avec valeurs manquantes

La distance de Mahalanobis est établie pour des observations complètes et doit être adaptée aux observations incomplètes. Une option consiste à utiliser l'estimation EEM pour imputer la moyenne conditionnelle pour les valeurs manquantes, sachant les valeurs observées, puis à calculer la distance de Mahalanobis en utilisant les valeurs imputées. Sous une hypothèse MAR (données manquant au hasard), il existe un modèle valide fondé sur la partie observée des données pour imputer les valeurs manquantes. Dans le cas de la détection des observations aberrantes, nous supposons que le modèle d'imputation peut tenir pour la majeure partie des données seulement et qu'il est estimé de manière robuste. Mais alors, nous ne pouvons pas nous attendre à ce qu'une valeur aberrante soit prédite par le modèle, sauf si la partie observée d'une observation est déjà

aberrante. Par conséquent, l'imputation avant la détection des observations aberrantes n'offre aucun avantage et nous préférons adapter directement la distance de Mahalanobis à la situation où des valeurs manquent. Deux versions distinctes de cette distance sont possibles dans cette situation.

Nous appelons la première version distance de Mahalanobis *marginale*. Elle consiste à utiliser la distance de Mahalanobis dans l'espace des variables observées et à la rééchelonner à l'aide d'un facteur p/q, où $q = \sum_k r_{ik}$ est le nombre de variables non manquantes et p est le nombre total de variables. Plus précisément, nous supposons qu'une observation x est partitionnée en $x = (x_o^\top, x_m^\top)^\top$ (éventuellement après un réarrangement), où x_o dénote la partie observée et x_m la partie non observée de l'observation. Alors, la distance de Mahalanobis marginale est

$$\mathrm{MD}^2_{\mathrm{marg}} = \frac{p}{q}(x_o - m_o)^\top (S_{oo})^{-1}(x_o - m_o), \qquad (9)$$

où S_{oo} est la partie de la matrice de covariance qui correspond à x_o. Cette version est également utilisée dans Little et Smith (1987).

La deuxième version de la distance de Mahalanobis en cas de valeurs manquantes s'obtient en réduisant à zéro la contribution des valeurs manquantes à cette distance. Cela reviendrait à remplacer toutes les valeurs manquantes par leur moyenne, c'est-à-dire $x_m = m_m$. Autrement dit, nous imputerions une moyenne sans tenir compte de la matrice de covariance et les arguments à l'encontre de la détection des observations aberrantes avec valeurs imputées s'appliqueraient également ici. Néanmoins, nous avons testé cette deuxième version de la distance de Mahalanobis. Elle produit des distances erratiques (Béguin 2002 ainsi que Béguin et Hulliger 2003) et nous n'avons pas poursuivi son utilisation.

4. L'algorithme BACON-EEM

Les algorithmes BACON et EEM demandent tous deux une grande puissance de calcul et en les fusionnant d'une manière commode, nous augmentons la performance. La structure de « croissance » de l'algorithme BACON implique des redondances qui peuvent être utilisées pour éviter des calculs supplémentaires à chaque étape de l'algorithme EEM. À chaque étape BACON, le point crucial est que les estimations de la moyenne et de la matrice de covariance à partir de l'algorithme EEM permettent d'exclure les points aberrants du bon sous-ensemble, ce qui ne requiert pas d'estimations extrêmement précises. Donc, il n'est pas nécessaire d'itérer EEM jusqu'à la convergence chaque fois que la moyenne et la covariance sont requises. Nous utilisons seulement cinq itérations par défaut. En outre, nous prenons le résultat de la dernière itération EEM de la dernière étape BACON comme valeur de départ pour l'algorithme EEM.

Autant d'information que possible provenant des itérations antérieures devrait être réutilisée. En fait, la statistique suffisante T^G calculée sur un bon sous-ensemble G possède une partie observée de la somme T_o^G et une partie manquante de la somme T_m^G. L'espérance calculée par l'étape E peut par conséquent s'écrire

$$E(T^G \mid X_o^G, \theta) = T_o^G + E(T_m^G \mid X_o^G, \theta). \qquad (10)$$

Comme les sous-ensembles G grandissent habituellement, $\hat{T}_o^G$ n'est pas recalculé à chaque étape de la boucle BACON, mais une variable globale pour $\hat{T}_o^G$ est mise à jour chaque fois que G change (habituellement l'ajout de points seulement, parfois l'élimination de quelques-uns).

À chaque itération de l'algorithme BACON-EEM, quand l'algorithme EEM a obtenu les estimations du centre et de la dispersion du bon sous-ensemble, les distances de Mahalanobis marginales de toutes les observations sont utilisées à l'étape 2 de l'algorithme BACON.

Mentionnons le point crucial en ce qui concerne la robustesse de l'algorithme : EEM n'est pas robuste, mais à chaque étape BACON, il est exécuté uniquement sur les points dont la distance de Mahalanobis marginale était la plus petite et, par conséquent, non aberrante à l'étape précédente. Autrement dit, l'observation x sera utilisée par EEM si, et uniquement si, x_o est suffisamment petit pour la mesure donnée par $(S_{oo})^{-1}$ à l'étape précédente. Par conséquent, si le premier sous-ensemble de bons points est exempt ou presque exempt d'observations aberrantes, le processus d'imputation dans EEM ne créera jamais d'observations aberrantes au cours de l'exécution complète de l'algorithme BACON-EEM. Cela signifie que l'algorithme EEM non robuste est protégé par l'approche de recherche avant générale de l'algorithme BACON, de la même façon que sont protégées la moyenne et la covariance non robustes de l'algorithme BACON original.

En résumé, les étapes de l'algorithme BACON-EEM sont les suivantes :

1. Calculer la médiane des coordonnées pondérée $\widehat{\mathrm{med}}(x)$ en ignorant les valeurs manquantes dans chaque variable séparément. Déterminer la distance euclidienne par rapport à la médiane de chaque observation en omettant les valeurs manquantes, mais en normant pour le nombre de valeurs présentes : $a_i = \|x_i - \widehat{\mathrm{med}}(x)\|\sqrt{p/q}$. Sélectionner les $m = cp$ observations ayant les a_i les plus faibles pour constituer le sous-ensemble G initial.
2. Calculer un centre $\hat{m}_G$ et une dispersion $\hat{C}_G$ en utilisant l'algorithme EEM et mettre à jour l'estimation de la statistique suffisante de la partie observée $\hat{T}_o^G$.

3. Calculer les carrés des distances de Mahalanobis marginales $\mathrm{MD}_G^2(x_i)$ pour $i = 1, ..., n$. Le nouvel ensemble G' contient les observations pour lesquelles $\mathrm{MD}_G^2(x_i) < c_{\hat{N}p\hat{r}}\, \chi^2_{p,\alpha}$.
4. Si $G' = G$, arrêter; sinon, fixer G à G' et passer à l'étape 2.

Si les objectifs principaux de l'algorithme BACON-EEM sont les estimations de la moyenne et de la covariance plutôt que la détection des observations aberrantes, l'algorithme EEM peut être itéré plus loin sans changer G. À l'étape 3, on pourrait aussi utiliser α/n au lieu de α (voir la section 2).

5. Applications

À la présente section, nous comparons l'algorithme BACON-EEM (BEM) à la méthode des corrélations de rang transformées (TRC) de Béguin et Hulliger (2004) et à l'algorithme ER de Little et Smith (1987). En guise de valeurs repères supplémentaires, nous utilisons une application sous le modèle normal multivarié avec estimation de la moyenne et de la covariance par l'algorithme EM. Autrement dit, nous créons une imputation non robuste. Puis, nous obtenons des estimations robustes de la position multivariée et de la matrice de covariance à l'aide de l'estimateur par déterminant minimal de covariance (MCD) calculé sur les données imputées et, enfin, nous détectons les observations aberrantes en utilisant les distances de Mahalanobis correspondantes. La méthode de référence est appelée GIMCD pour *Gauss Imputation followed by MCD detection* (imputation gaussienne suivie d'une détection MCD). Les algorithmes sont implémentés en `R` (R Development Core Team 2006) avec l'aide des progiciels `R norm` (Novo et Schafer 2002) et `MASS` (Venables et Ripley 2002).

5.1 Données sur les feux de brousse

La réaction de l'algorithme BACON-EEM à l'introduction de valeurs manquantes est illustrée à l'aide d'un ensemble de données réelles de 38 observations et 5 variables. Il a été utilisé par Maronna et Zamar (2002) pour localiser les empreintes de feux de brousse. Cet exemple bien connu est également étudié dans Maronna et Yohai (1995) et dans Maronna et Zamar (2002). Il permet de produire un graphique bidimensionnel (dans les variables 2 et 3) qui révèle la plupart des observations aberrantes (voir la figure 1). Les données contiennent une grappe d'observations aberrantes 33 à 38, une deuxième grappe d'observations aberrantes 7 à 11, et quelques autres observations aberrantes isolées, à savoir les observations 12, 13, 31 et 32. Nous avons ajouté l'observation 31 à la liste des observations aberrantes éventuelles, parce que les méthodes MCD, BACON et d'autres méthodes étudiées dans Maronna et Zamar (2002) indiquent qu'il s'agit d'un cas limite. Les valeurs manquantes sont créées selon un mécanisme MCAR (valeur manquant entièrement au hasard). Deux ensembles de données sont créés en utilisant respectivement 20 % et 40 % d'items manquants. L'ensemble de données pour lequel 40 % des valeurs manquent contient des observations pour lesquelles jusqu'à quatre valeurs sur cinq manquent et, par conséquent, présentent un défi quelle que soit la méthode. Comme la taille n de l'ensemble de données est faible, nous exécutons l'algorithme BACON-EEM avec le test $\chi^2_{p,\alpha/n}$. Les résultats sont présentés au tableau 1. Pour les observations 7 à 13 et 31 à 38, nous indiquons individuellement si elles sont détectées ou non, tandis que pour les 23 autres bons points, nous indiquons le nombre d'observations désignées comme étant aberrantes. La limite au dessous de laquelle une distance de Mahalanobis indique qu'une valeur est aberrante a été déterminée pour chaque passage machine en examinant le graphique des quantiles de la distance de Mahalanobis.

Tableau 1 Observations aberrantes détectées pour trois taux de valeurs manquantes

(1 - *q*)%	Méthode	7-11	12,13,21	32-38	N^bre non ab.
0	MD	11100	000	0000000	0
0	MCD	11111	001	1111111	0
0	ER	11111	100	0000000	0
0	BEM	11111	001	1111111	0
0	TRC	01100	000	1111111	0
20	GIMCD	11111	001	1111111	0
20	ER	10000	000	0000000	0
20	BEM	11111	000	1111111	1
20	TRC	01111	111	1111111	1
40	GIMCD(1)	11100	000	1111111	0
40	GIMCD(2)	11100	000	0100000	5
40	ER	10000	010	0000000	2
40	BEM	11111	000	1111111	1
40	TRC	11111	010	1111111	1

MD : distance de Mahalanobis classique, MCD : déterminant minimal de covariance, GIMCD : imputation non robuste sous modèle gaussien suivi de MCD (GIMCD(1) et GIMCD(2) sont deux réalisations de l'algorithme GIMCD), ER : espérance-maximisation avec une étape M à la maximisation, BEM : BACON-EEM, TRC : corrélations de rang transformées. La première colonne indique la proportion de valeurs manquantes, la dernière colonne donne le nombre d'autres points (non aberrants) déclarés comme des observations aberrantes, les colonnes intermédiaires sont des indicateurs de détection des observations dans la première ligne.

Quand les observations sont complètes, aucune observation n'est désignées comme étant aberrante par aucune des méthodes. La distance de Mahalanobis avec moyenne et covariance non robustes détecte trois observations aberrantes, mais manquent les autres. Les méthodes MCD et BACON-EEM aboutissent à la sélection du même sous-ensemble de données comme sous-ensemble de bons points et, par conséquent, donnent le même résultat, c'est-à-dire exactement la même distance de Mahalanobis pour chacune

des observations. Dans les deux cas, les observations 12 et 13 ne sont pas considérées comme étant aberrantes, mais toutes les autres le sont. La méthode ER détecte le groupe d'observations aberrantes 7,...,12, mais aucune des autres observations aberrantes. La méthode TRC détecte le groupe 32,...,38 et deux observations aberrantes supplémentaires.

Dans le cas où 20 % des valeurs manquent, les méthodes BEM et TRC désignent un bon point de donnée comme étant une observation aberrante. La méthode TRC détecte 14 des 15 observations éventuellement aberrantes. La méthode ER manque toutes les observations aberrantes, sauf l'observation 7. La méthode GIMCD détecte les 13 mêmes observations aberrantes que dans la situation où aucune valeur ne manque. Notons toutefois qu'il existe une certaine variabilité dans les résultats de la GIMCD à cause de l'imputation aléatoire. La méthode BEM détecte les mêmes observations aberrantes que sous les conditions de données complètes, sauf l'observation 11.

Dans le cas où 40 % de valeurs manquent, l'algorithme ER désigne les observations 7 et 13 et deux bonnes observations comme étant aberrantes. Puisque l'imputation est aléatoire, le résultat de GIMCD présente une certaine variabilité. Deux réalisations, GIMCD(1) et GIMCD(2) sont présentées au tableau 1. Dans un bon cas, GIMCD détecte 10 des 15 observations aberrantes et ne désigne aucune bonne observation comme étant aberrante. Dans un mauvais cas, GIMCD ne détecte que quatre observations aberrantes, mais indique que cinq bonnes observations sont aberrantes. La méthode BEM détecte 12 des observations aberrantes et déclare qu'une bonne observation est aberrante. La méthode TRC détecte 13 observations aberrantes et déclare qu'une bonne observation est aberrante.

5.2 Données MU281

L'ensemble de données MU284 tiré de Särndal, Swensson et Wretman (1992) contient des données sur les municipalités suédoises. Nous utilisons les variables *population en 1975* et *population en 1985* (pop75 et pop85), *recettes des impôts municipaux 1985* (RMT85), *nombre d'employés municipaux 1984* (ME84) et *valeur immobilière 1984* (REV84). Les trois plus grandes villes selon pop75 sont écartées, parce qu'elles représentent d'énormes observations aberrantes qui seraient traitées séparément en pratique. Nous supposons que les municipalités restantes constituent un échantillon stratifié d'une population plus grande. Les strates sont définies conformément à $0 < \text{pop75} < 20, 20 \leq \text{pop75} < 100, 100 \leq \text{pop75}$. Le tableau 2 donne les tailles de population supposées et les poids correspondants. Ce plan d'échantillonnage reflète une stratification typique des enquêtes auprès des établissements avec une strate à tirage complet pour les établissements les plus grands, où, en bout de ligne, huit des dix établissements répondent à l'enquête.

Tableau 2 Taille de population et d'échantillon pour MU281

	strate		
	1	2	3
pop75	0 à 19	20 à 99	100+
N	1 600	250	10
n	171	102	8
w	9,36	2,45	1,25

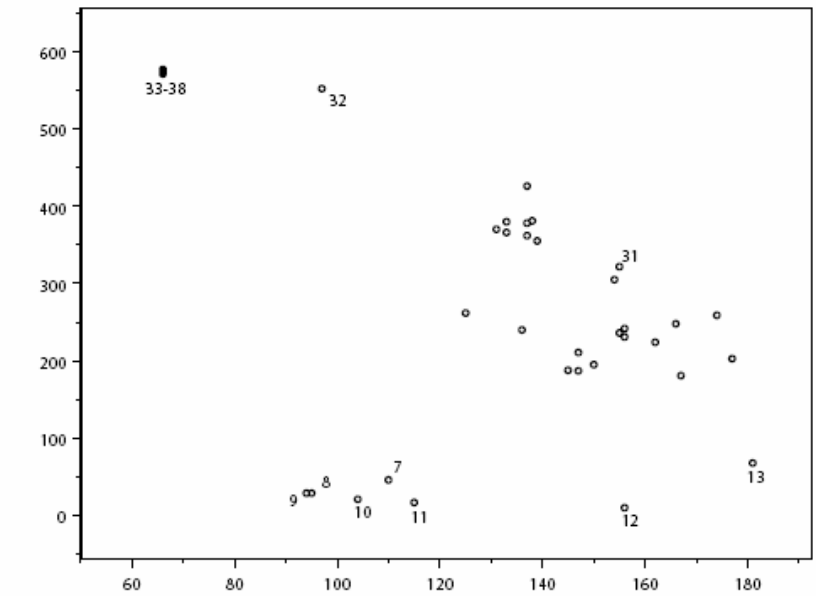

Figure 1 Données sur les feux de brousse

Les trois variables RMT85, ME84 et REV84 sont divisées par pop85 pour obtenir les chiffres par habitant. Les variables dont la valeur est exprimée par habitant sont dénotées par des noms en lettres minuscules (rmt85, me84 et rev84). La figure 2 montre la distribution de ces trois variables et de la variable auxiliaire pop75. Les chiffres par habitant suivent une distribution approximativement elliptique. Il y a un lien linéaire entre rmt84 et me84 et un lien légèrement non linéaire entre ces variables et pop75. Il ne semble exister aucune relation entre rev84 et rmt85, ni entre rev84 et me84, mais la variabilité de rev84 est clairement plus importante quand pop75 est faible. Les distributions de variables pop75 et rev84 sont asymétriques. Il existe une grande observation aberrante dans rmt85 et dans me84, et au moins deux dans rev84. Nous incluons pop75 dans tous nos calculs. En pratique, on inclurait la variable auxiliaire qui définit le plan d'échantillonnage dans un modèle. Notons que pop75 ne présente pas de valeurs manquantes.

Le diagramme qq des distances de Mahalanobis basé sur MCD montre uniquement les deux observations aberrantes manifestes dans rev84. La grande observation aberrante relevée conjointement dans rmt85 et dans me84 possède la 25^{e} plus grande distance de Mahalanobis. Nous donnons aux 25 observations les plus grandes le nom d'observations aberrantes de base non pondérées. Dans l'ensemble de données MU284 original, ces observations ont les étiquettes 3, 4, 29, 31, 46, 47, 56, 79, 83, 117, 126, 131, 140, 158, 199, 211, 222, 246, 248, 252, 254, 260, 262, 272 et 273. Dans le cas des distances de Mahalanobis non robustes classiques, seulement 12 des observations aberrantes de base sont désignées comme étant aberrantes, c'est-à-dire sont parmi les 25 observations ayant la plus grande distance (classique) de Mahalanobis. Des méthodes robustes sont nécessaires pour détecter les observations aberrantes dans l'ensemble de données MU281. Afin de pouvoir comparer les méthodes, nous fixons le nombre d'observations qui doivent être considérées comme aberrantes à 25 pour le moment. Donc, nous considérons pour chaque méthode les 25 observations pour lesquelles la distance de Mahalanobis est la plus grande comme étant des observations aberrantes. Soulignons qu'il ne s'agit pas forcément du seuil que l'on choisirait après inspection du graphique qq des distances de Mahalanobis.

Le tableau 3 montre le nombre d'observations aberrantes de base détectées par les méthodes exécutées sur l'ensemble de données complètes. Évidemment, MCD détecte ses propres 25 observations aberrantes. Les algorithmes ER, BEM et TRC détectent 25 ou 24 de ces observations aberrantes si aucune pondération n'est appliquée, mais uniquement 11 ou 15, s'ils sont pondérés. Un suffixe « w » derrière l'acronyme de la méthode indique que les poids d'échantillonnage ont été utilisés. Puisque les petites municipalités ont plus de poids, les estimations sont attirées vers elles et d'autres observations aberrantes apparaîtrons parmi les 25 distances de Mahalanobis les plus grandes pour ERw, BEMw et TRCw (voir aussi le tableau 4). Une inspection plus approfondie révèle que nombre des observations aberrantes non pondérées sont situées dans la queue de la distribution de rev84, tandis que la plupart des observations aberrantes pondérées sont situées dans la queue de la distribution de pop75. Les résultats des méthodes pondérées coïncident pour 20 observations considérées comme aberrantes. Nous les appellerons observations aberrantes de base pondérées. Ces dernières ont pour étiquette dans l'ensemble de données MU284 original 16, 28, 36, 45, 46, 55, 97, 113, 115, 121, 155, 185, 196, 208, 233, 241, 245, 265, 267 et 270. Seulement dix de ces observations figurent parmi les observations aberrantes de base non pondérées. Le tableau 4 donne le nombre d'observations aberrantes de base pondérées et non pondérées dans les strates. Dans la strate 1, il y a 12 observations aberrantes non pondérées, mais seulement deux pondérées. Donc, la pondération influe clairement sur la détection des observations aberrantes. L'influence s'exerce principalement sur le modèle qui est attiré vers les petites observations dont les poids sont grands. Naturellement, cette situation peut être considérée comme une sorte de masquage des observations aberrantes, mais dans le contexte de la modélisation, une meilleure explication est que le modèle n'est pas entièrement adéquat sur toutes les strates et que le modèle pondéré est mieux ajusté à la population que le modèle non pondéré.

La deuxième ligne du tableau 3 donne le temps de calcul des algorithmes. L'algorithme ER est nettement plus lent que ses concurrents, ce qui pourrait toutefois tenir à une implémentation inefficace. L'algorithme le plus rapide est BEM, suivi de TRC et, à une certaine distance, de MCD. L'algorithme TRC peut toutefois devenir lent si le taux de valeurs manquantes est élevé.

Tableau 3 MU281 complet, nombre d'observations aberrantes de base non pondérées détectées

Méthode	MCD	ER	ER_W	BEM	BEM_W	TRC	TRC_W
Nombre détectées	25	25	11	24	15	24	15
Temps de calcul	0,81	3,17	2,52	0,07	0,04	0,14	0,14

Le suffixe w indique que l'algorithme est exécuté avec les poids d'échantillonnage.

Tableau 4 Nombre d'observations aberrantes de base par strate

Strate	1	2	3	Total
Non pondérées	12	5	8	25
Pondérées	2	10	8	20

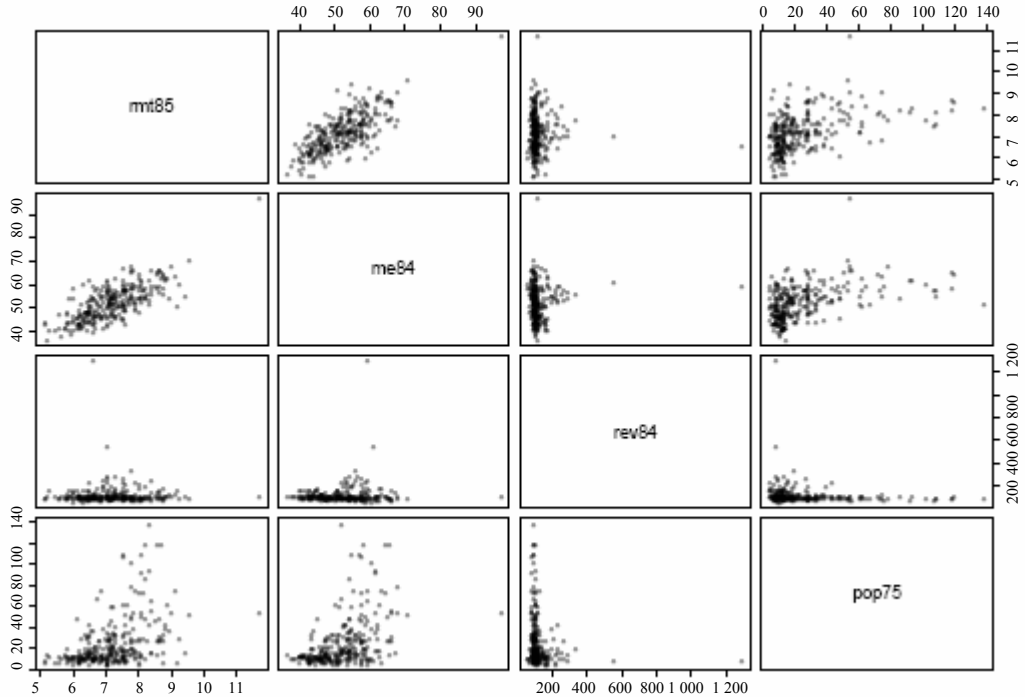

Figure 2 MU281 – Chiffres par habitant et pop75

5.2.1 Valeurs manquantes

Nous introduisons maintenant des valeurs manquantes dans les variables rmt85, me84 et rev84 selon un mécanisme MAR léger. Nous attribuons un profil de valeurs manquantes aux observations en fonction du modulo de leur étiquette dans les données MU284 originales. Le mécanisme MAR est reflété par des taux plus élevés de valeurs manquantes pour les strates 1 et 2 (voir l'annexe). Les profils de réponse et les taux de valeurs manquantes par strate sont présentés au tableau 5. Par exemple, rev84 est la seule valeur manquante dans 15 observations de la strate 1 et dans deux observations de la strate 2. Dans l'ensemble, 187 observations demeurent complètes et la proportion d'observations pour lesquelles des valeurs manquent (taux de valeurs manquantes) est de 33 %.

Tableau 5 Fréquence des profils de réponse par strate pour rmt85, me84 et rev84

Indicateur de réponse			Strate		
rmt85	me84	rev84	1	2	3
0	1	1	11	4	0
1	0	1	13	2	0
1	1	0	15	2	0
1	0	0	13	2	1
0	1	0	14	2	0
0	0	1	11	4	0
1	1	1	94	86	7
Taux de valeurs manquantes			0,450	0,157	0,125

Parmi les 35 observations aberrantes de base pondérées ou non pondérées, il existe 17 observations incomplètes. Le tableau 6 indique combien de ces observations aberrantes de base ont été détectées après l'introduction des valeurs manquantes. Les 20 observations aberrantes de base pondérées sont bien détectées par les algorithmes pondérés ERw, BEMw et TRCw. GIMCD détecte 4 de ces observations et 14 des observations aberrantes de base non pondérées. Donc, la présence de valeurs manquantes affecte les capacités de l'algorithme MCD qui avait, en fait, été utilisé pour définir les observations aberrantes de base non pondérées. Une mise en garde est toutefois de rigueur. L'exécution répétée de l'imputation gaussienne a révélé une certaine variabilité des résultats de la méthode de GIMCD. Cependant, même avec un résultat d'imputation favorable, GIMCD ne bat pas les méthodes ER, BEM ou TRC dans la détection des observations aberrantes de base non pondérées. Les versions pondérées ERw, BEMw et TRCw détectent également les observations aberrantes de base pondérées. Le nombre d'observations complètes parmi les observations aberrantes désignées par les diverses méthodes est indiqué à la dernière ligne du tableau 6. Toutes les méthodes désignent aussi comme étant aberrantes des observations comportant des valeurs manquantes. Puisque le taux de valeurs manquantes est élevé dans la strate des petites observations et que les versions pondérées des méthodes désignent moins d'observations aberrantes dans cette strate, le nombre d'observations aberrantes complètes

est habituellement plus grand pour les algorithmes pondérés. Dans l'ensemble, l'introduction de valeurs manquantes n'altère pas beaucoup les capacités des algorithmes ER, BEM et TRC, tandis qu'elle affecte moyennement celles de la méthode GIMCD.

Tableau 6 Ensemble de données MU281 avec valeurs manquantes, nombre d'observations aberrantes de base détectées

Méthode	GIMD	GIMCD	ER	ERw	BEM	BEMw	TRC	TRCw
Pondérées	14	4	10	20	9	19	12	17
Non pondérées	12	14	23	11	22	15	22	17
Complètes	16	8	17	20	16	18	19	18

GIMD : Imputation non robuste sous un modèle gaussien suivie d'une détection par distance classique de Mahalanobis.

5.2.2 Observations aberrantes supplémentaires

En plus des observations aberrantes dans les données originales, nous en introduisons maintenant de nouvelles. Les observations qui devraient devenir des observations aberrantes supplémentaires sont déterminées par le modulo de l'étiquette originale. Si (LABEL mod $8 = 1$ et pop75 ≥ 10) ou (LABEL mod $16 = 1$ et pop75 < 10) alors l'observation est une observation aberrante supplémentaire. Donc, le taux d'observations aberrantes est plus élevé pour les grandes municipalités. Cependant, il n'est pas influencé par les valeurs des autres variables. Nous pouvons dire que l'aberrance est aléatoire. Notons que nous aurions pu tirer un échantillon aléatoire au lieu de l'échantillon systématique susmentionné. Nous avons opté pour l'échantillon systématique afin de simplifier la répétition des résultats et d'éviter un caractère aléatoire supplémentaire.

Il se fait que deux des observations aberrantes de base pondérées et l'une de celles non pondérées sont aussi des observations aberrantes supplémentaires. Nous continuons de les traiter comme des observations aberrantes de base. Compte tenu de cette situation, il y a 32 observations aberrantes supplémentaires dans l'échantillon. Combinées aux 25 observations aberrantes de base non pondérées et aux 25 pondérées définies plus haut, elles donnent 57 ou 52 observations aberrantes à déceler (20,3 % ou 18,5 % d'observations aberrantes). À partir d'ici, nous fixons le seuil pour les distances de Mahalanobis à la 57^{e} distance la plus grande pour simplifier la comparaison des méthodes.

Les valeurs des observations aberrantes supplémentaires sont créées comme il suit : $\text{rmt} = 0,2 * \text{rmt85} + 8$, $\text{me} = 0,1 * \text{me84} + 50$, $\text{rev} = 0,4 * \text{rev84} + 300$. Notons que nous omettons le suffixe indiquant l'année pour les variables contaminées. La dépendance à l'égard des anciennes valeurs est négligeable. Il est utilisé uniquement pour éviter un modèle explicite pour l'erreur autour du point $(\text{rmt}, \text{me}, \text{rev}) = (8, 50, 300)$. Il s'agit du type de contamination qu'il est difficile de déceler pour les estimateurs de covariance robustes (Rocke et Woodruff 1996), c'est-à-dire une contamination concentrée et proche du nuage de points des bonnes observations.

La figure 3 montre les trois variables contaminées et la position des observations aberrantes supplémentaires.

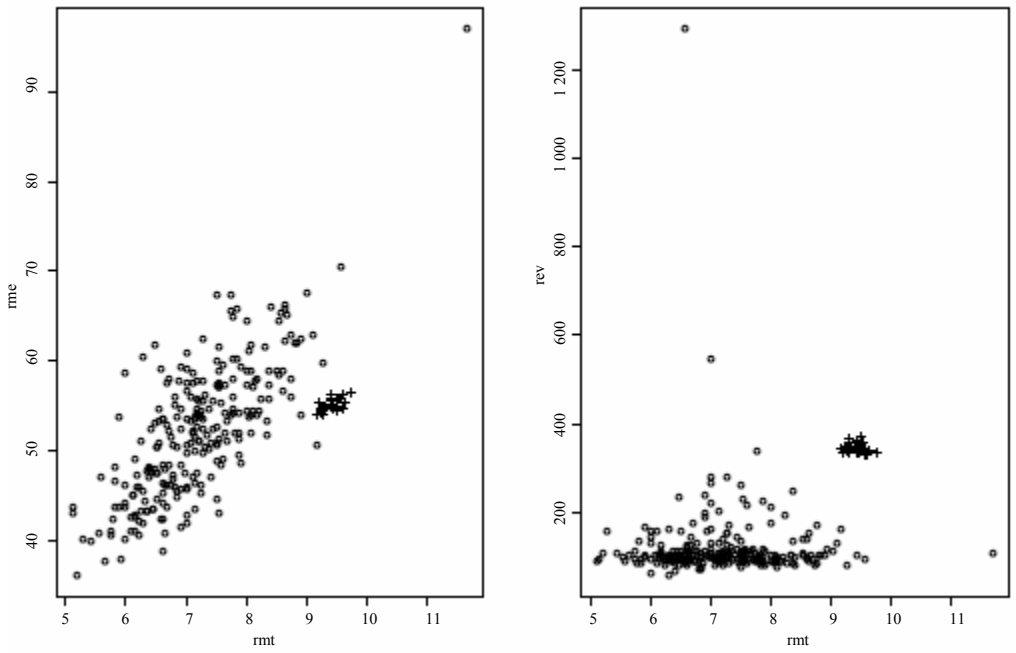

Figure 3 Variables me et rev c. rmt. Les observations aberrantes supplémentaires sont marquées d'une croix

Le tableau 7 montre le nombre d'observations aberrantes décelées. GIMCD détecte 31 des 32 observations aberrantes supplémentaires, tandis que BEM, BEMw, TRC et TRCw en détectent un grand nombre, mais pas toutes. ER et ERw détecte un moins grand nombre d'observations aberrantes supplémentaires. Les observations aberrantes de base pondérées sont toutes décelées par ER et ERw, BEMw et TRCw. Les versions non pondérées de BEM et de TRC détectent un moins grand nombre d'observations aberrantes de base pondérées et GIMCD n'en détecte que quatre. Pondérés ou non, BEM et TRC sont les algorithmes qui détectent le mieux les observations aberrantes de base non pondérées.

La dernière ligne du tableau 7 montre que l'imputation non robuste de GIMCD permet de désigner plus d'observations aberrantes avec valeurs manquantes que les autres méthodes dont les imputations sont déjà robustifiées avant la phase de détection.

Tableau 7 MU281 avec valeurs manquantes et contamination supplémentaire modérée

Méthode	N^{bre} OA	GIMCD	ER	ERw	BEM	BEMw	TRC	TRCw
Additionnelles	32	31	19	6	27	27	28	27
De base pondérées	20	4	20	20	11	20	12	20
De base non pondérées	25	13	15	12	23	19	23	18
Complètes		24	34	43	38	40	40	40

N^{bre} OA : nombre d'observations aberrantes, n^{bre} complètes : nombre d'observations complètes parmi les observations aberrantes détectées.

Afin de déterminer la défaillance des méthodes quand un nombre élevé d'observations aberrantes supplémentaires contaminent les données, nous établissons que des observations sont aberrantes si (LABEL mod $2 = 0$ et pop75 ≥ 20) ou (LABEL mod $3 = 1$ et pop75 < 20). À part les observations qui sont déjà des observations aberrantes de base, cela fait 98 observations supplémentaires. Donc, regroupées avec les 25 observations de base non pondérées, elles donnent 43,8 % d'observations aberrantes. Le seuil pour les méthodes est par conséquent fixé à la 123^{e} distance de Mahalanobis la plus grande. Le tableau 8 montre qu'à cause de ce seuil élevé, toutes les observations aberrantes de base pondérées sont détectées par toutes les méthodes. Les méthodes GIMCD, ER et ERw ne peuvent faire face au taux élevé d'observations aberrantes. La méthode BEM détecte la plupart des observations aberrantes, suivie de près par BEMw et TRC. La méthode TRCw détecte un peu moins d'observations aberrantes de base non pondérées et d'observations aberrantes supplémentaires.

Tableau 8 MU281 avec valeurs manquantes et forte contamination supplémentaire

Méthode	N^{bre} OA	GIMCD	ER	ERw	BEM	BEMw	TRC	TRCw
Additionnelles	98	20	19	37	85	85	85	80
De base pondérées	20	20	20	20	20	20	20	20
De base non pondérées	25	21	19	17	23	18	18	13

6. Conclusion

L'algorithme EM pour données normales multivariées peut être adapté à un contexte d'échantillonnage. L'algorithme BACON protège l'algorithme EEM non robuste contre les observations aberrantes quand ce dernier est appliqué dans une itération de l'algorithme BACON. L'algorithme ER utilise une robustification à l'intérieur de l'algorithme EM. Les applications révèlent que cette robustification pourrait ne pas suffire. Toutefois, une explication possible pourrait être le point de départ non robuste de l'étape M dans l'algorithme ER.

La méthode de GIMCD, qui est un algorithme EM non robuste suivi d'une imputation et d'une détection avec l'estimateur de covariance MCD, donne de remarquablement bons résultats dans des conditions modérées de valeurs manquantes et de contamination. Sa variabilité quand le taux de valeurs manquantes est élevé est un inconvénient. Des solutions plus stables tenant également compte du plan d'échantillonnage devraient être étudiées. L'algorithme BACON-EEM manifeste de très bonnes capacités de détection, en particulier quand le taux de valeurs manquantes et le taux de contamination sont élevés.

Malgré sa simplicité, l'algorithme TRC est une bonne méthode dans de nombreuses circonstances. Son principal inconvénient semble être l'imputation ponctuelle avec une seule covariable, qui peut poser un problème quand le taux de valeurs manquantes est élevé.

Afin de trouver un bon modèle pour la population, il est important d'utiliser les poids d'échantillonnage. Néanmoins, il est conseillé d'utiliser aussi une version non pondérée et d'examiner les différences. Il se pourrait que les observations aberrantes soient masquées par les poids d'échantillonnage élevés, parce qu'elles peuvent alors dominer dans l'estimation du modèle.

Remerciements

Le projet de recherche EUREDIT faisait partie du Programme de la technologie de la société d'information (TSI) du cinquième programme cadre de l'Union européenne. La participation de la Suisse à EUREDIT a été

financée par le Bureau fédéral pour l'Éducation et la Science de la Suisse. Une partie importante de ces travaux de recherche a été effectuée pendant que les deux auteurs travaillaient au Bureau fédéral de la statistique de la Suisse. Les auteurs remercient les examinateurs et les rédacteurs de leurs commentaires constructifs.

Annexe

Valeurs manquantes dans MU281

Le profil de réponse par défaut est 111, qui indique que les valeurs des trois variables rmt85, me84 et rev84 sont toutes présentes. Un 0 dans la chaîne de caractères indique une valeur manquante pour la variable correspondante. Pour commencer, nous changeons le profil de réponse dans toutes les strates conformément au scénario suivant avec les paramètres $(a, b, c) = (1, 2, 3)$:

$$\text{profil de réponse} = \begin{cases} 011, & \text{si LABEL mod } 20 = \text{a;} \\ 101, & \text{si LABEL mod } 20 = \text{b;} \\ 110, & \text{si LABEL mod } 20 = \text{c;} \\ 100, & \text{si LABEL mod } 30 = \text{a;} \\ 010, & \text{si LABEL mod } 30 = \text{b;} \\ 001, & \text{si LABEL mod } 30 = \text{c .} \end{cases}$$

En outre, nous appliquons de nouveau le scénario susmentionné avec les paramètres $(a, b, c) = (5, 6, 7)$ pour la strate 1.

Bibliographie

Atkinson, A. (1993). Stalactite plots and robust estimation for the detection of multivariate outliers. Dans *Data Analysis and Robustness*. (Éds., S. Morgenthaler, E. Ronchetti et W. Stahel), Birkäuser.

Béguin, C. (2002). Outlier detection in multivariate data. Master's thesis, Université de Neuchâtel.

Béguin, C., et Hulliger, B. (2003). Robust multivariate outlier detection and imputation with incomplete survey data. Deliverable D4/5.2.1/2 Part C, EUREDIT.

Béguin, C., et Hulliger, B. (2004). Multivariate oulier detection in incomplete survey data: The epidemic algorithm and transformed rank correlations. *Journal of the Royal Statistical Society*, A 167(Part 2.), 275-294.

Billor, N., Hadi, A.S. et Vellemann, P.F. (2000). BACON: Blocked Adaptative Computationally-efficient Outlier Nominators. *Computational Statistics and Data Analysis*, 34(3), 279-298.

Campbell, N. (1989). Bushfire mapping using noaa avhrr data. Rapport technique, CSIRO.

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81(396), 1063-1069.

Cheng, T.-C., et Victoria-Feser, M.-P. (2000). Robust correlation estimation with missing data. Rapport technique 2000.05, Université de Genève.

Dempster, A.P., Laird, N.M. et Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (c/r: P22-37). *Journal of the Royal Statistical Society*, Séries B, *Methodological*, 39, 1-22.

Donoho, D. (1982). *Breakdown Properties of Multivariate Location Estimators*. Ph.d. qualifying paper, Department of Statistics, Harvard University.

EUREDIT (2003). *Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit project*, Volume 1 and 2. EUREDIT consortium. http://www.cs.york.ac.uk/euredit/results/results.html.

Gnanadesikan, R., et Kettenring, J.R. (1972, March). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28, 81-124.

Hadi, A.S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society*, B, 54(3), 761-771.

Huber, P.J. (1981). *Robust Statistics*. New York : John Wiley & Sons, Inc.

Kosinski, A.S. (1999). A procedure for the detection of multivariate outliers. *Computational Statistics & Data Analysis*, 29, 145-161.

Little, R., et Smith, P. (1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association*, 82, 58-68.

Little, R.J.A., et Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York : John Wiley & Sons, Inc.

Liu, R.Y., Parelius, J.M. et Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, 27(3), 783-858.

Maronna, R., et Zamar, R. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4), 307-317.

Maronna, R.A. (1976). Robust *M*-estimators of multivariate location and scatter. *The Annals of Statistics*, 4, 51-67.

Maronna, R.A., et Yohai, V.J. (1995). The behaviour of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429), 330-341.

Novo, A.A., et Schafer, J.L. (2002). *norm*: *Analysis of multivariate normal datasets with missing values*. R package version 1.0-9.

R Development Core Team (2006). *R*: *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rocke, D., et Woodruff, D. (1993). Computation of robust estimates of multivariate location and shape. *Statistica Neerlandica*, 47, 27-42.

Rocke, D., et Woodruff, D. (1996). Identification of outlier in multivariate data. *Journal of the American Statistical Association*, 91(435), 1047-1061.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. Dans *Mathematical Statistics and Applications*, Volume B, 283-297. Elsevier.

Rousseeuw, P.J., et Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York : John Wiley & Sons, Inc.

Schafer, J. (2000). *Analysis of Incomplete Multivariate Data*, Volume 72 of *Monographs on Statistics and Applied Probability*. Chapman & Hall.

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

Stahel, W. (1981). *Robuste Schätzungen*: *Infinitesimale optimalitätund Schätzungen von Kovarianzmatrizen*. Thèse de Doctorat, Swiss Federal Institute of Technology.

Venables,W.N., et Ripley, B.D. (2002). *Modern Applied Statistics with S* (quatrième éd.). New York : Springer. ISBN 0-387-954570.

Wilks, S.S., et Gnanadesikan, R. (1964). Graphical methods for internal comparisons in multiresponse experiments. *Annals of Mathematical Statistics*, 35, 623-631.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2008
Vol. 34, N° 1, pp. 115-130
Statistique Canada, N° 12-001-X au catalogue

Offre de primes d'encouragement aux répondants dans une enquête par panel multimodes : effets cumulatifs sur la non-réponse et le biais

Annette Jäckle et Peter Lynn [1]

Résumé

L'offre de primes en vue d'encourager la participation aux enquêtes est un moyen utilisé de plus en plus fréquemment pour combattre la chute des taux de réponse et les risques résultant de biais dû à la non-réponse. Cette dernière pose tout spécialement un problème dans les enquêtes par panel, car même un faible taux de non-réponse d'une vague à l'autre peut entraîner des pertes cumulées importantes et, si la non-réponse est différentielle, cette situation peut aboutir à l'accroissement du biais d'une vague à l'autre. Les effets des primes d'encouragement ont été bien étudiés dans le contexte d'enquêtes transversales, mais l'on en sait fort peu quant à leurs effets cumulés sur l'ensemble des vagues d'un panel. Nous fournissons de nouvelles preuves des effets des paiements incitatifs continus sur l'attrition, ou érosion de l'échantillon, sur le biais et sur la non-réponse partielle en nous appuyant sur les données d'une expérience d'offre de primes d'encouragement à grande échelle, multivagues, à mode d'interview mixte, intégrée dans une enquête par panel réalisée auprès des jeunes par le gouvernement britannique. Dans cette étude, les primes d'encouragement ont donné lieu à une réduction significative de l'attrition, qui compensait de loin les effets négatifs sur les taux de réponse partiels en ce qui concerne la quantité d'information recueillie dans le cadre de l'enquête par cas émis. Les primes d'encouragement ont eu des effets proportionnels sur les taux de rétention pour une gamme de caractéristiques des répondants et, par conséquent, n'ont pas réduit le biais d'attrition relatif à ces caractéristiques. Les effets des primes d'encouragement sur le taux de rétention étaient plus importants pour les primes inconditionnelles que pour les primes conditionnelles, et plus importants également pour le mode de collecte par la poste que pour celui par téléphone. D'une vague à l'autre, les effets sur l'attrition ont diminué quelque peu, mais les effets sur la non-réponse partielle et l'absence d'effet sur le biais sont demeurés constants. Les effets des primes d'encouragement offertes lors de vagues plus tardives semblent avoir été indépendants des traitements relatifs aux primes et des modes de collecte des données appliqués lors de vagues antérieures.

Mots clés : Attrition; non-réponse partielle; enquête par la poste; ITAO; prime d'encouragement inconditionnelle; prime d'encouragement conditionnelle.

1. Introduction

Partout dans le monde, les taux de réponse aux enquêtes sont à la baisse, ce qui accroît les risques de biais de non-réponse si les caractéristiques des non-répondants diffèrent systématiquement de celles des répondants. Dans le cas des enquêtes par panel, cette situation pose un problème particulier, car même de faibles taux de non-réponse lors de chaque vague peuvent donner lieu à des pertes cumulées importantes. Si la non-réponse est différentielle, le biais pourrait augmenter avec la durée du panel. Afin de relever les taux de participation, les organismes d'enquête offrent de plus en plus fréquemment des primes d'encouragement aux répondants. Le présent article fournit de nouvelles preuves des effets cumulatifs des primes d'encouragement sur l'attrition, le biais d'attrition et le biais de non-réponse partielle, en s'appuyant sur des données provenant d'une expérience d'offre de primes d'encouragement à grande échelle, multivagues, à mode de collecte mixte, intégrée dans une enquête par panel menée auprès des jeunes par le gouvernement britannique.

Les effets des primes d'encouragement ont été étudiés dans de nombreuses circonstances. Ainsi, les incitations monétaires accroissent plus la réponse que les cadeaux et les loteries (Church 1993; Singer, Hoewyk, Gebler, Raghunathan et McGonagle 1999), les primes d'encouragement inconditionnelles (c'est-à-dire, celles qui sont données au moment de la demande de participation à l'enquête) accroissent plus la réponse que les primes conditionnelles (celles qui sont promises en échange de la participation) (Church 1993; Goyder 1994; Hopkins et Gullickson 1992; Singer et coll. 1999), les taux de réponse augmentent avec la valeur de la prime (Armstrong 1975; Church 1993; Fox, Crask et Kim 1988; Hopkins et Gullickson 1992; Rodgers 2002; Yu et Cooper 1983), les primes ont des effets plus importants dans les études où les taux de réponse sont faibles, ainsi que dans les enquêtes avec envoi du questionnaire par la poste que dans celles où le questionnaire est administré par un intervieweur (Singer et coll. 1999). Toutefois, la plupart des données révélant des différences d'effet des primes d'encouragement selon le mode de collecte des données émanent de comparaisons entre études distinctes et ne tiennent pas compte de l'effet des différences entre d'autres mesures qui ont une incidence sur la réponse. Par conséquent, les écarts constatés entre les effets des primes d'encouragement ne sont pas nécessairement de

1. Annette Jäckle, Institute for Social and Economic Research, University of Essex, Colchester CO4 3SQ, UK. Courriel : aejack@essex.ac.uk; Peter Lynn, Institute for Social and Economic Research, University of Essex, Colchester CO4 3SQ, UK. Courriel : plynn@essex.ac.uk.

vrais effets de mode. L'étude réalisée par Ryu, Couper et Marans (2006) fait exception. Les auteurs comparent les effets de primes monétaires et de cadeaux dans une enquête à mode de collecte mixte par la poste et sur place. Néanmoins, leur étude ne comporte pas de scénario sans prime et ne permet donc pas d'évaluer l'ordre de grandeur des effets des primes d'encouragement entre les modes. Nous avons comparé les effets de primes d'encouragement dans une enquête à mode de collecte mixte par la poste et par interview téléphonique assistée par ordinateur (ITAO) et, dans le mode de collecte par la poste, nous avons également examiné les effets de primes conditionnelles et inconditionnelles.

La recherche sur les effets des primes d'encouragement s'est concentrée sur les taux de réponse et l'on en sait fort peu au sujet des effets sur le biais, lequel est la raison pour laquelle les faibles de taux de réponse sont préoccupants. Les études sur les primes d'encouragement sont limitées principalement à l'examen des effets sur le biais de composition de l'échantillon et certaines ont révélé que ces primes accroissent de manière disproportionnée la participation des répondants habituellement sous-représentés, comme les personnes peu instruites (Singer, Van Hoewyk et Maher 2000), les pauvres (James 1997), les Noirs ou les pauvres (Mack, Huggins, Keathley et Sundukchi 1998), les membres de groupes ethniques minoritaires noirs ou indiens, vivant dans de grands ménages ou dans des ménages avec enfants à charge, âgés de 0 à 20 ans, ou les personnes seules (Stratford, Simmonds et Nicolaas 2003). Cependant, les biais de composition de l'échantillon ne sont pas nécessairement corrélés aux biais qui entachent d'importantes estimations par sondage, notamment parce qu'en général, quelques caractéristiques compositionnelles seulement sont étudiées. En dernière analyse, les études du biais de non-réponse sont restreintes par le manque d'information au sujet des non-répondants, limite que permettent de surmonter en partie les études par panel, où l'information sur les non-répondants peut être extraite des vagues qui ont précédé le décrochage de l'enquête. Nous estimons l'importance du biais dû à l'attrition en fonction de la composition de l'échantillon et des variables étudiées. Puis, nous évaluons l'efficacité avec laquelle les primes d'encouragement réduisent le biais, en exploitant l'information disponible dans le panel sur les personnes qui cessent de participer à l'enquête.

En outre, les renseignements sont peu abondants sur les effets des primes d'encouragement d'une vague à l'autre d'une enquête par panel, que le même traitement soit répété ou qu'il change d'une vague à l'autre. Dans une étude de l'utilisation de primes d'encouragement dans les enquêtes longitudinales, Laurie et Lynn (sous presse) concluent que, compte tenu des répercussions de la modification des conditions d'encouragement sur les coûts, il existe étonnamment peu de données sur les effets de long terme de tels changements en vue d'orienter les praticiens des sondages. Citant un mémoire interne du US Census Bureau, Ward, Boggess, Selvavel et McMahon (2001) écrivent (voir la page 2) qu'un examen des études longitudinales bien connues (Downs 1999) a révélé que toutes les études non réalisées par le Census Bureau comportaient l'offre d'une prime monétaire à chaque vague, mais qu'aucun test scientifique n'avait été réalisé en vue de déterminer l'efficacité de ces primes. Si l'attrition correspond au décrochage des participants les moins coopérants, l'échantillon pourrait compter de plus en plus de répondants déterminés à participer, donc moins sensibles aux primes d'encouragement, parce qu'ils sont suffisamment motivés pour participer même sans les primes (Laurie et Lynn, sous presse). Dans ces conditions, l'effet marginal des primes sur le taux de réponse pourrait décroître au cours de la vie du panel. De même, l'effet sur le biais d'attrition pourrait croître, si les primes ont un effet disproportionné sur les membres de l'échantillon qui, en leur absence, seraient plus susceptibles de décrocher. Bien que quelques études sur les primes d'encouragement aient été réalisées dans le contexte de panels, elles ne portaient pour la plupart que sur une seule vague ou examinaient des changements de traitement incitatif d'une vague à la suivante. Martin, Abreu et Winters (2001) et Ward et coll. (2001), par exemple, ont étudié les effets des primes d'encouragement sur les taux de conversion des non-répondants aux vagues antérieures; Rodgers (2002) et Laurie (2007) ont examiné les effets des changements de valeur des primes dans un panel. Les seules études s'intéressant aux effets des primes d'encouragement sur plus de deux vagues semblent être celles de James (1997), de Mack, Huggins, Keathley et Sundukchi (1998) et de Laurie et Lynn (sous presse), qui ont signalé que l'effet positif d'un paiement incitatif au début d'un panel persiste pendant plusieurs vagues, même si le paiement n'est pas répété. Ces études ne portent toutefois que sur l'effet d'un paiement incitatif lors d'une seule vague et n'examinent pas les effets cumulatifs de primes d'encouragement offertes au cours de plusieurs vagues successives. Nous examinons ici les effets cumulatifs de paiements incitatifs continus au cours de trois vagues s'étendant sur une période de trois ans, ainsi que les effets du passage d'un mode de collecte par téléphone à un mode de collecte par la poste et du passage d'un traitement incitatif conditionnel à un traitement incitatif inconditionnel.

Enfin, on trouve dans la littérature des données contradictoires quant aux effets des primes d'encouragement sur la qualité des données. Bien que soient fréquemment exprimées des craintes que les primes d'encouragement aboutissent à la collecte de données de moins bonne qualité

parce qu'elles accroissent marginalement la motivation de répondants qui, autrement, auraient cessé de participer à l'étude et qui ne sont pas suffisamment aptes ou motivés pour répondre diligemment, les auteurs des études existantes concluent soit que les primes suscitent un plus grand effort de la part du répondant et produisent une non-réponse partielle plus faible (James et Bolstein 1990; Mack et coll. 1998; Singer et coll. 2000), soit qu'il n'existe aucune association (Berk, Mathiowetz, Ward et White 1987; Davern, Rockwood, Sherrod et Campbell 2003; Goyder 1994; Shettle et Mooney 1999; Singer et coll. 1999; Teisl, Roe et Vayda 2005; Tzamourani et Lynn 1999; Willimack, Schuman, Pennell et Lepkowski 1995). La non-réponse partielle est un facteur qui peut être crucial, parce que les analystes n'utilisent habituellement que les cas pour lesquels les données sont complètes, ce qui entraîne une perte d'efficacité due à la réduction de la taille des échantillons et, comme dans le cas de la non-réponse totale, peut biaiser les estimations et invalider l'inférence si les personnes qui ne répondent pas à certaines questions ne représentent pas un sous-ensemble aléatoire de l'échantillon (Mason, Lesser et Traugott 2002). Les problèmes que pose la non-réponse partielle deviennent plus épineux dans le cas de l'analyse multivariée si les profils de données manquantes varient d'une question à l'autre, ainsi que dans le cas de l'analyse des changements, qui, en plus, dépend de renseignements complets à divers points dans le temps. Puisque les primes d'encouragement peuvent avoir une incidence à la fois sur la non-réponse totale et la non-réponse partielle, leur effet éventuel sur des mesures répétées dérivées d'une étude par panel n'est pas clair. Nous avons examiné l'effet des primes d'encouragement sur les taux de non-réponse partielle et calculé leur effet net sur l'attrition et la réponse partielle.

2. Hypothèses testées

Les résultats mesurés dans le cadre de la présente analyse sont le taux d'attrition, le taux de non-réponse partielle et le biais d'attrition. Ici, l'attrition est un état sans retour, car l'enquête ne comportait pas la reprise des non-répondants dans l'échantillon à l'occasion de vagues ultérieures. La non-réponse partielle est mesurée par le nombre de questions non filtrées pour lesquelles la réponse manquait, à cause d'un refus ou d'une réponse « Ne sais pas ». (Les questions non filtrées sont celles qui s'appliquent à tous les membres de l'échantillon; les questions pour lesquelles l'admissibilité est déterminée par la réponse à une question antérieure sont exclues de notre mesure de la non-réponse partielle.) Le biais d'attrition est mesuré en fonction des caractéristiques sociodémographiques et des mesures des variables étudiées de la première vague. Ces trois mesures des résultats sont utilisées pour tester les hypothèses suivantes :

*H*1 : *Effets des primes d'encouragement sur l'attrition, la non-réponse partielle et le biais d'attrition.*

Dans le cadre d'études antérieures, les primes d'encouragement faisaient généralement croître les taux de réponse, que ce soit parce que les normes d'échanges sociaux obligent les répondants à rendre la pareille (norme de réciprocité, Gouldner 1960) ou parce que la prime d'encouragement se substitue au manque de motivation à participer pour d'autres raisons, comme le devoir civique ou l'intérêt pour le sujet [théorie du levier et de la saillance (*leverage-salience theory*), Groves, Singer et Corning 2000]. En plus de motiver les répondants à fournir des réponses de meilleure qualité, les primes d'encouragement peuvent réduire la non-réponse partielle. Parallèlement, elles peuvent modifier la composition de l'échantillon en y intégrant un plus grand nombre de répondants ne manifestant pas de diligence à répondre aux questions de l'enquête et, par conséquent, accroître la non-réponse partielle. Enfin, les primes d'encouragement peuvent avoir des effets différents sur l'attrition selon les membres de l'échantillon. En principe, ceux qui ont une forte propension à participer à l'enquête sans la prime devraient être moins affectés par l'offre de primes d'encouragement que ceux qui sont plus susceptibles de décrocher de l'enquête. Par conséquent, les primes d'encouragement pourraient réduire le biais d'attrition.

*Hypothèse nulle H*1 : *Les primes d'encouragement n'ont pas d'effet sur l'attrition, la non-réponse partielle ni le biais d'attrition.*

*H*2 : *Effets des primes d'encouragement d'une vague à l'autre.*

L'effet d'accroissement des taux de réponse totale et de réponse partielle des primes d'encouragement pourrait s'affaiblir d'une vague à l'autre, si l'attrition se traduit par l'abandon des membres les moins motivés à participer et que les membres restants sont suffisamment motivés pour d'autres raisons et donc moins sensibles aux primes d'encouragement (Laurie et Lynn, sous presse). Cependant, la portée de la réduction du biais de non-réponse pourrait croître au fil des vagues de l'enquête si les primes d'encouragement retiennent de manière disproportionnée les membres de l'échantillon les plus susceptibles d'abandonner.

*Hypothèse nulle H*2 : *Les effets des primes d'encouragement ne varient pas d'une vague à l'autre.*

*H*3 : *Effets des primes d'encouragement inconditionnelles et conditionnelles dans le contexte d'un panel.*
Selon des études antérieures portant sur des enquêtes transversales, les primes d'encouragement inconditionnelles ont des effets plus importants sur la non-réponse totale, peut-être parce que le paiement préalable signale que l'organisme d'enquête pense qu'il peut compter sur la participation des membres de l'échantillon, ce qui renforce la norme de réciprocité. Il reste à déterminer si les diverses conditions incitatives ont des effets différents sur la non-réponse partielle.
*Hypothèse nulle H*3 : *Les primes d'encouragement inconditionnelles et conditionnelles ont les mêmes effets dans un contexte de panel.*

*H*4 : *Effets des primes d'encouragement dans les modes de collecte par la poste et par téléphone.*
Les comparaisons d'études antérieures donnent à penser que les primes d'encouragement ont un effet plus important lorsque la collecte de données se fait par la poste, peut-être parce que dans le mode de collecte par téléphone, l'intervieweur agit déjà comme un motivateur externe accroissant la réponse totale ainsi que partielle (Singer et coll. 1999) et que l'importance des améliorations supplémentaires est plus faible. Il pourrait ne pas en être de même dans le contexte d'un panel, où l'effet du mode de collecte sur la réponse peut être modifié par l'expérience du répondant lors des vagues antérieures.
*Hypothèse nulle H*4 : *Les primes d'encouragement ont les mêmes effets en mode de collecte par la poste et par téléphone.*

*H*5 : *Effets des changements de mode de collecte ou de traitement incitatif d'une vague de l'enquête à l'autre.*
Comparativement aux membres de l'échantillon affectés au même mode de collecte et au même traitement à toutes les vagues, ceux qui sont affectés à des traitements différents ou à des modes de collecte différents ont peut-être vécu des expériences différentes durant les vagues précédentes de l'enquête et ont peut-être des attentes différentes au sujet des vagues futures. Par conséquent, l'effet des primes d'encouragement pourrait non seulement être conditionnel au mode de collecte durant la vague courante, mais être aussi influencé par le traitement incitatif et le mode de collecte des vagues antérieures.
*Hypothèse nulle H*5 : *Les changements de mode de collecte ou de traitement incitatif entre les vagues n'ont pas d'effets persistants.*

*H*6 : *Effets des primes d'encouragement selon le niveau d'aptitude.*
Les membres de l'échantillon dont le niveau d'études est faible sont habituellement plus susceptibles de décrocher des enquêtes. Si les primes d'encouragement réduisent le biais d'attrition, elles devraient par conséquent réduire de manière disproportionnée l'attrition chez les personnes moins performantes. Les répondants dont le niveau d'aptitude est faible pourraient aussi être plus susceptibles de fournir des réponses incomplètes, s'ils trouvent plus difficile que les autres la tâche de répondre au questionnaire envoyé par la poste. Donc, les primes d'encouragement *peuvent* accroître les niveaux moyens de non-réponse partielle.
*Hypothèse nulle H*6 : *Les primes d'encouragement ont les mêmes effets quel que soit le niveau d'aptitude.*

3. Plan de l'étude

La Youth Cohort Study of England and Wales (YCS) a pour but d'étudier les transitions de la scolarité obligatoire à la poursuite d'études de niveau plus élevé ou au marché du travail et, en général, comprend l'échantillonnage tous les deux ans de cohortes de jeunes de 16 et 17 ans qui sont étudiées à plusieurs occasions à intervalle d'un an. L'expérience sur les primes d'encouragement a été intégrée dans les vagues 2, 3 et 4 de la cohorte 10. L'enquête est gérée et financée par le Department for Children, Schools and Families, qui a conçu l'expérience sur les primes d'encouragement en collaboration avec le National Centre for Social Research, qui a réalisé en sous-traitance les vagues 2 et 3 de la cohorte 10 de la YCS.

3.1 L'enquête

La population étudiée dans la cohorte 10 de l'YCS comprenait les élèves d'Angleterre et du pays de Galles ayant atteint l'âge minimal pour quitter l'école (16 ans) durant l'année scolaire 1998-1999 (Russell et Phelps 2001), c'est-à-dire une cohorte d'une année d'âge d'élèves nés entre le 1er septembre 1982 et le 31 août 1983. Un échantillon aléatoire au 1/10 a été tiré des registres des écoles (à l'exclusion des écoles spéciales et des écoles comptant moins de 20 élèves de cet âge) en 1999, en demandant aux écoles de fournir les noms et adresses des élèves nés le 5, le 15 et le 25 de chaque mois. Puis, un échantillon aléatoire systématique de 25 000 élèves a été tiré du fichier résultant contenant 31 424 noms et adresses. La première vague de l'enquête a eu lieu un an plus tard, au printemps 2000, la deuxième à la fin de 2000, la troisième au printemps 2002 et la quatrième au printemps 2003. Les non-répondants n'ont

pas été repris lors des vagues subséquentes et, par conséquent, l'attrition était monotone.

La vague 1 était une enquête par la poste avec suivi par téléphone des non-répondants après quatre envois par la poste (envoi du questionnaire initial et de trois rappels). D'après les résultats d'examens qu'ils ont déclarés, les répondants de la vague 1 ont été classés dans la catégorie des « élèves plus performants » s'ils avaient été reçus à cinq examens pour l'obtention du certificat général d'études secondaires (GCSE) avec une note allant de A* à C, ou dans la catégorie des « élèves moins performants », autrement. Environ le tiers des répondants de la vague 1 ont été classés ainsi dans la catégorie des moins performants. À l'occasion de la vague 2, environ le tiers des membres de l'échantillon utilisé sur le terrain ont été sélectionnés au hasard pour répondre à des questions supplémentaires sur des sujets particuliers et assignés à des interviews téléphoniques assistées par ordinateur (ITAO). En plus du questionnaire de base, les jeunes interviewés par téléphone ont répondu à un module sur les décisions relatives à la poursuite d'études supérieures (pour les élèves plus performants) ou sur les aspirations en matière d'études et d'emploi (pour les élèves moins performants). Les autres membres de l'échantillon ont reçu le questionnaire de base par la poste. À la vague 3, tous les élèves moins performants ont reçu le questionnaire de base envoyé par la poste, mais le module téléphonique a continué d'être administré à un tiers des élèves plus performants. À la quatrième vague, tous les répondants ont été affectés à l'enquête de base effectuée par la poste. La figure 1 illustre l'affectation aux divers modes de collecte et aux diverses primes d'encouragement.

Le questionnaire de base est demeuré en grande partie le même pour les trois vagues expérimentales. Les jeunes interrogés par téléphone ont d'abord répondu aux questions de base, puis aux modules supplémentaires. Le questionnaire de base était le même que le questionnaire envoyé par la poste, quoique certaines questions ont été adaptées en vue de leur administration par téléphone. L'interview téléphonique durait, en moyenne, 20 minutes. (Les questionnaires et les rapports techniques peuvent être consultés par la voie de UK Data Archive dans les annexes du Guide de l'utilisateur de l'YCS à http://www.data-archive.ac.uk/findingdata/snDescription.asp?sn=4571&key=YCS.)

3.2 L'expérience sur les primes d'encouragement

En réponse aux préoccupations croissantes suscitées par la baisse des taux de réponse, une expérience a été introduite dans la deuxième vague de la cohorte 10 afin d'étudier l'effet de paiements d'encouragement sur les taux de réponse et sur le biais de non-réponse. Un bon de 5 livres sterling (environ 10 $US ou 7 euros) a été envoyé à une partie des répondants de la vague 1 de l'enquête réalisée par la poste ainsi que de celle réalisée par téléphone, tandis que les groupes témoins n'ont pas reçu de prime d'encouragement. En outre, dans le cas de l'enquête par la poste, les primes d'encouragement étaient inconditionnelles (la prime a été envoyée avec l'envoi par la poste initial) ou conditionnelles (le bon a été promis dans l'envoi par la poste original, mais envoyé uniquement après réception du questionnaire dûment rempli). À l'occasion des vagues 3 et 4, toutes les primes d'encouragement ont été payées inconditionnellement.

3.3 Affectation des répondants aux modes de collecte et aux traitements incitatifs

À l'occasion de la vague 2, les répondants de la vague 1 ont été affectés aléatoirement au mode de collecte par téléphone ou par la poste. Par contre, l'affectation des traitements incitatifs a eu lieu au niveau de l'école (affectation randomisée par grappe selon le mode). Chaque école représentée dans l'échantillon a été affectée à un traitement par téléphone (témoin ou prime d'encouragement) et affectée indépendamment à un traitement, éventuellement différent, par la poste (témoin, prime d'encouragement inconditionnelle ou prime d'encouragement conditionnelle), de sorte que tous les membres de l'échantillon appartenant à une même école et approchés selon le même mode de collecte ont reçu le même traitement incitatif.

Vague 1	Questionnaire de base envoyé par la poste. Résultats d'examens déclarés utilisés pour classer les répondants pour l'affectation à la vague 2.									
	Élèves plus performants					Élèves moins performants				
Vague 2	T_x	T_u	P_x	P_u	P_c	T_x	T_u	P_x	P_u	P_c
Vague 3	T_x	T_u	P_x	P_u	--	--	--	P_x	P_u	--
Vague 4	--	--	P_x	P_u	--	--	--	P_x	P_u	--
Questionnaire	De base + études supérieures		De base			De base + études et emplois		De base		

Nota : T = téléphone, P = poste, x = témoin, u = prime inconditionnelle, c = prime conditionnelle. Les flèches indiquent les changements de traitement incitatif ou de mode de collecte d'une vague à l'autre.

Figure 1 Plan expérimental

Les 4 712 répondants moins performants de la vague 1 ont été stratifiés selon le numéro d'identification dans l'école, dans la région (*Government Office Region*) et affectés alternativement à un traitement par téléphone et à un traitement par la poste. Nous avons exclu de l'analyse 627 cas pour lesquels il n'existait aucun numéro de téléphone valide dans le dossier, car ceux de ces sujets auxquels on avait affecté un mode de collecte par téléphone ont été approchés par la poste. Par conséquent, l'analyse des élèves moins performants est limitée à 2 097 cas approchés par téléphone et 1 988 cas approchés par la poste.

Une procédure semblable a été suivie pour les élèves plus performants, excepté qu'une plus forte proportion d'entre eux a été affectée au traitement par la poste. La vague 1 comptait 8 909 répondants plus performants dont 751 pour lesquels on ne disposait pas de numéro de téléphone valide et qui ont donc été exclus de l'analyse. Après ces exclusions, 2 922 élèves plus performants ont été affectés au mode de collecte par téléphone et 5 236, au mode de collecte par la poste.

Pour l'affectation des écoles aux groupes de traitement incitatif, les écoles possédant des membres de l'échantillon interviewé par téléphone (c'est-à-dire toutes les écoles, sauf quelques-unes très petites dont moins de cinq élèves faisaient partie de l'échantillon) ont été stratifiées en fonction du ratio d'élèves moins performants aux élèves plus performants dans l'échantillon et affectées aléatoirement aux traitements incitatifs à l'intérieur des strates. (La proportion d'écoles affectées aux primes d'encouragement était de 1/2 si le ratio d'élèves moins performants aux élèves plus performants dans l'échantillon était ≥ 2, de 1/3 si $1/2 \leq \text{ratio} < 2$ et de 1/4 pour toutes les autres écoles.) La procédure a été répétée pour l'affectation des écoles dans les groupes de traitement par la poste, où celles sélectionnées pour un traitement incitatif ont été réparties aléatoirement en un groupe de traitement conditionnel et un groupe de traitement inconditionnel. (La proportion affectée au traitement incitatif était de 2/3 si le ratio était ≥ 2; de 1/3 pour $0 \leq \text{ratio} < 1/2$ et de 1/6 pour toutes les autres écoles.) Toutes les estimations du degré de signification présentées dans le texte tiennent compte du plan d'échantillonnage par grappe de l'expérience sur les primes d'encouragement.

Le tableau 1 montre les tailles des échantillons émis sur le terrain à chaque vague pour les diverses combinaisons de traitement incitatif et de mode de collecte, après exclusion des cas dont l'inadmissibilité était connue, à cause d'un déménagement à l'étranger ou du décès ($n = 13$ à la vague 2; $n = 3$ à la vague 3). Les cas inadmissibles à la vague 4 ne sont pas identifiés dans les données, mais leur nombre est vraisemblablement faible. Sont également exclus de l'analyse les répondants de la vague 1 pour lesquels aucun numéro de téléphone n'était connu au moment de l'affectation aux modes de collecte pour la vague 2, comme il est décrit plus haut, et 117 élèves plus performants affectés au mode de collecte par téléphone à la vague 2 qui ont répondu par la poste et ont été subséquemment affectés au mode de collecte par la poste.

Tableau 1 Taux conditionnels et cumulatifs de réponse

Vague	Taux de réponse %	Élèves plus performants					Élèves moins performants				
		$T_xT_xP_x$	$T_uT_uP_u$	$P_xP_xP_x$	$P_uP_uP_u$	$P_cP_uP_u$	$T_xP_xP_x$	$T_uP_uP_u$	$P_xP_xP_x$	$P_uP_uP_u$	$P_cP_uP_u$
2	Conditionnel	76,82	80,91	78,23	86,45	82,32	65,21	70,41	64,93	75,00	71,35
	(*n* émis)	(2,075)	(728)	(3,262)	(1,004)	(967)	(1,282)	(811)	(807)	(608)	(569)
3	Conditionnel	69,13	73,17	73,07	81,91	81,36	59,09	70,93	63,36	71,93	70,20
	(*n* émis)	(1,594)	(589)	(2,551)	(868)	(794)	(836)	(571)	(524)	(456)	(406)
	Cumulatif	53,11	59,20	57,16	70,82	66,94	38,53	49,94	41,14	53,95	50,09
4	Conditionnel	72,21	85,61	76,11	85,65	86,82	63,16	74,26	65,36	75,30	81,34
	(*n* émis)	(1,101)	(431)	(1,863)	(711)	(645)	(494)	(404)	(332)	(328)	(284)
	Cumulatif	38,31	50,69	43,48	60,66	58,03	24,34	36,99	26,89	40,63	40,60

Nota : Taux de réponse selon la définition normalisée RR1 de l'AAPOR. Les groupes de traitement sont désignés par T = téléphone, P = poste, x = témoin, u = prime d'encouragement inconditionnelle, c = prime d'encouragement conditionnelle. Par exemple, $T_x T_x P_x$ indique que l'échantillon a été affecté au groupe témoin interviewé par téléphone aux vagues 2 et 3, et au groupe témoin interviewé par la poste à la vague 4. Les taux de réponse conditionnels sont conditionnés sur la réponse à la vague précédente. La base du calcul est le nombre de cas émis dans l'échantillon, qui exclut les non-répondants à la vague précédente et les cas inadmissibles. Les taux de réponse cumulatifs correspondent au pourcentage de répondants de la vague 1 qui restent dans l'échantillon de répondants. La base pour le calcul est le nombre de cas émis dans l'échantillon de la vague 2, à l'exclusion de trois élèves plus performants inadmissibles à la vague 3 (1 cas $P_x P_x P_x$ et 2 cas $P_c P_u P_u$).

Le tableau 1 donne aussi les taux de réponse d'une vague sur l'autre ainsi que cumulatifs observés (RR1 de l'AAPOR). Les taux sont présentés selon le niveau de performance des élèves et la combinaison séquentielle mode de collecte/traitement incitatif. Les taux de réponse d'une vague sur l'autre pour l'échantillon d'élèves plus performants affectés au groupe témoin interviewé par téléphone à la vague 2 et transférés au groupe témoin interviewé par la poste à la vague 4 (colonne 1) étaient, par exemple, 76,82 %, 69,13 % et 72,21 %. Les nombres émis de cas ont diminué, pour passer de 2 075 à 1 101 au cours des trois vagues, parce que les cas non répondants n'ont pas été repris dans l'échantillon lors des vagues subséquentes.

4. Mesures des résultats et méthodes

L'analyse est fondée sur l'échantillon de répondants de la vague 1, puisque l'affectation aux traitements expérimentaux s'appuie sur l'information recueillie au cours de la première vague et que les caractéristiques correspondantes des non-répondants de la vague 1 sont inconnues. Nous nous concentrons par conséquent sur l'attrition, conditionnellement à la réponse à la vague 1. Il s'agit de l'aspect de la non-réponse qui est particulier aux enquêtes par panel, bien qu'il faille évidemment reconnaître que les caractéristiques de l'attrition sont conditionnées par celles de la réponse à la vague 1. Le taux de réponse à la vague 1 (RR1 de l'AAPOR) était de 54,80 %, après exclusion de cinq cas dont l'inadmissibilité était connue (Russell et Phelps 2001). La présente section décrit les mesures des résultats et les méthodes utilisées pour évaluer les hypothèses quant aux effets des primes d'encouragement.

4.1 Attrition

Afin de tester l'effet des primes d'encouragement sur l'attrition, nous avons estimé la probabilité d'attrition sous la forme d'une fonction des variables du plan expérimental [mode téléphonique (*tel*), primes d'encouragement inconditionnelles (*unc*), primes d'encouragement conditionnelles (*cond*), élèves moins performants (*la*)] et leurs interactions]. Pour chacune des trois vagues expérimentales ($t = 2, 3, 4$), nous avons estimé un modèle probit distinct de la probabilité d'attrition, en utilisant dans chaque cas l'échantillon de la vague 1 comme base :

$$\begin{aligned}\Pr(attrition_{it}) = F(&\beta_{0t} + \beta_{1t} tel_i + \beta_{2t} unc_i + \beta_{3t} cond_i \\ &+ \beta_{4t} la_i + \beta_{5t} tel_i * unc_i + \beta_{6t} la_i * tel_i \\ &+ \beta_{7t} la_i * unc_i + \beta_{8t} la_i * cond_i \\ &+ \beta_{9t} la_i * tel_i * unc_i + \varepsilon_i) \qquad (1)\end{aligned}$$

où F est la fonction lien probit. Ensuite, nous avons utilisé les coefficients et les erreurs-types estimés d'après le modèle pour calculer les probabilités prévues d'attrition sous diverses conditions de traitement et pour tester l'existence de différences due aux primes d'encouragement.

4.2 Non-réponse partielle

Afin de tester l'effet des primes d'encouragement sur la non-réponse partielle, nous avons estimé des modèles de comptage du nombre de réponses manquantes à une question, en utilisant toutes les questions non filtrées du questionnaire de base utilisé à la vague 2 ($n = 44$), 3 ($n = 48$) et 4 ($n = 46$), où toute réponse « Ne sais pas » a été comptée comme une valeur manquante. Nous avons utilisé la même spécification des prédicteurs que dans le modèle (1) pour estimer des modèles de régression binomiale négative distincts pour chacune des trois vagues expérimentales, conditionnellement à la réponse à la vague en question. (La surdispersion signifiait que les modèles de Poisson n'étaient pas ajustés aux données : la valeur P du test du rapport de vraisemblance de même moyenne et de même variance était 0,0000 pour toutes les vagues.) Les coefficients et les erreurs-types estimés pour ces modèles ont été utilisés pour calculer la non-réponse partielle prévue sous diverses conditions de traitement et pour tester le degré de signification des différences dues aux primes d'encouragement.

4.3 Biais d'attrition

Pour tester l'effet des primes d'encouragement sur le biais d'attrition, nous avons estimé la probabilité d'attrition en utilisant le modèle (1), mais en y incluant les caractéristiques des répondants à la vague (1) et leurs interactions avec les variables du plan expérimental comme prédicteurs. Nous avons estimé des modèles probit distincts pour l'attrition pour chacune des vagues expérimentales ($t = 2, 3, 4$) et pour chaque caractéristique, de nouveau en utilisant l'échantillon de répondants de la vague 1 (*v1*) comme base :

$$\begin{aligned}\Pr(attrition_{it}) = F(&\beta_{0t} + \beta_{1t} tel_i + \beta_{2t} unc_i + \beta_{3t} cond_i \\ &+ \beta_{4t} la_i + \beta_{5t} tel_i * unc_i + \beta_{6t} la_i * tel_i \\ &+ \beta_{7t} la_i * unc_i + \beta_{8t} la_i * cond_i \\ &+ \beta_{9t} la_i * tel_i * unc_i + \beta_{10} v1char \\ &+ \beta_{11} v1char_i * tel_i + \ldots \\ &+ \beta_{19} v1char_i * la_i * tel_i * unc_i + \varepsilon_i) \qquad (2)\end{aligned}$$

où β_{11} à β_{19} sont les coefficients des termes d'interaction des caractéristiques avec les variables du plan. Le

coefficient pour la caractéristique de répondant, β_{10}, fournit les renseignements sur la direction, l'ordre de grandeur et, en combinaison avec son erreur-type, la signification du biais d'attrition pour le groupe témoin des élèves plus performants interviewés par la poste et n'ayant pas reçu de prime d'encouragement. L'interaction de la caractéristique et les indicateurs de prime d'encouragement renseignent au sujet de la variation du biais d'attrition due aux primes d'encouragement. Nous avons calculé la signification de toutes les interactions présentées ici en suivant les recommandations formulées pour les modèles non linéaires par Norton, Wang et Ai (2004) à l'aide de la commande « predictnl » de la version 9 de Stata.

Les caractéristiques testées étaient le sexe, le type d'école, les résultats d'examens, l'activité courante (études à temps plein, emploi ou pas d'études, d'emploi ni de formation), l'expérience du chômage, la poursuite d'études en vue d'obtenir un certificat d'études professionnelles ou générales, la composition du ménage (vit avec les parents, un conjoint ou une conjointe, ni l'un ni l'autre) et un ensemble de questions attitudinales au sujet de l'emploi et de la formation. Les énoncés de toutes les questions sont présentés au tableau 6. Les caractéristiques choisies sont celles pour lesquelles on pouvait s'attendre à ce que les répondants et les non-répondants diffèrent, d'après des études antérieures de la non-réponse à la YCS ou à d'autres enquêtes et la théorie sur la non-réponse (Groves et Couper, 1998; Lynn, Purdon, Hedges et McAleese 1994).

4.4 Résultats présentés

Puisque les coefficients provenant des modèles non linéaires ne peuvent pas être interprétés de manière fondamentale (Long 1997), nous donnons les valeurs prévues fondées sur les estimations du modèle, plutôt que sur les coefficients. Sauf indication contraire, les résultats sont présentés pour le groupe des élèves plus performants. Pour donner une idée de la grandeur des différences de résultats entre les traitements, nous présentons les transformations des valeurs prévues, qui comparent chaque traitement avec le groupe de référence, c'est-à-dire le groupe témoin d'élèves plus performants interviewés par la poste.

5. Attrition, non-réponse partielle et biais d'attrition dans les groupes témoins

À titre de contexte de l'évaluation des effets des primes d'encouragement, la présente section fournit des renseignements sur la portée de l'attrition, de la non-réponse partielle et du biais d'attrition dans les groupes témoins, en mettant en relief les différences entre les vagues, les niveaux de performance des élèves et les modes de collecte. Tout au long de la discussion, le groupe des élèves plus performants interviewés par la poste et n'ayant pas reçu de prime d'encouragement constitue la catégorie de référence, à laquelle toutes les autres catégories de traitement sont comparées.

5.1 Attrition

Le taux cumulatif prévu d'attrition chez les élèves plus performants affectés au groupe témoin interviewé par la poste est passé de 21,77 % à la vague 2 à 56,53 % à la vague 4 (tableau 2, colonne 1). Pour les *élèves moins performants* (colonne 2), les taux d'attrition dans le groupe témoin interviewé par la poste étaient 61 % plus élevé à la vague 2, mais cet écart s'est amenuisé d'une vague à l'autre pour atteindre 29 % à la vague 4. L'écart selon le niveau de performance était néanmoins significatif pour chacune des trois vagues (valeur P de $\beta_4 = 0{,}0000$ pour $t = 2, 3, 4$). Dans le cas du *mode téléphonique* (colonne 3), le taux d'attrition dans le groupe témoin ne différait pas significativement à la vague 2, mais était supérieur de 9 % à la vague 3 (valeur P de $\beta_5 = 0{,}0034$ pour $t = 3$). Ce résultat est contraire à celui d'autres études, où la non-réponse est généralement plus faible pour le mode d'interview par téléphone, à cause du rôle joué par l'interviewer dans la persuasion des répondants à participer à l'enquête. L'une des raisons possibles du résultat opposé observé ici est qu'aussi bien pour le groupe interviewé par la poste que pour celui interviewé par ITAO, les efforts supplémentaires en vue d'obtenir les réponses auprès des non-répondants initiaux ont été faits par téléphone, de sorte que seul le groupe interviewé par la poste a été soumis à un traitement multimodes. Deuxièmement, à la vague 2, le fardeau de l'enquête (mesuré par la longueur de l'interview) était plus élevé pour les personnes interviewées par téléphone que celles approchées par la poste à cause des modules supplémentaires, ce qui pourrait avoir donné lieu à une non-réponse plus élevée à la vague 3 chez les premières que chez les secondes. Les taux de réponse cumulatifs prévus, qui ont servi de base pour le calcul des différences en pourcentage entre les groupes de traitement, sont présentés dans les trois premières colonnes du tableau 5.

5.2 Non-réponse partielle

Le nombre prévu de réponses manquantes à une question dans le groupe témoin des élèves plus performants interviewés par la poste était de 2,89 pour la vague 2, et tombait à 1,75 pour la vague 4 (tableau 3, colonne 1). Pour les *élèves moins performants* (colonne 2), le nombre prévu pour le groupe témoin était 21 % plus élevé à la vague 2, l'écart se creusant pour atteindre 45 % à la vague 4. Les différences selon le niveau de performance sont significatives pour les trois vagues ($P \leq 0{,}0001$ pour β_4, $t = 2, 3, 4$). Pour le *mode téléphonique* (colonne 3), le

nombre prévu était 4 % plus faible à la vague 2 et 12 % plus faible à la vague 3 ($P = 0{,}0000$ pour β_5, $t = 2, 3$) que pour le mode de collecte par la poste. Les nombres prévus de non-réponse partielle, utilisés comme base pour les calculs présentés au tableau 3, sont présentés dans les colonnes 4 à 6 du tableau 5.

5.3 Biais d'attrition

Dans le groupe témoin des élèves plus performants interviewés par la poste, la non-réponse était différentielle pour tous les domaines testés (tableau 4). Les échantillons de répondants surreprésentaient de manière significative les jeunes vivant avec leurs parents, poursuivant des études à temps plein ou étudiant en vue d'obtenir un certificat d'études générales. Les taux prévus d'attrition parmi ceux étudiant à temps plein dans le groupe témoin des élèves plus performants interviewés par la poste, par exemple, étaient 14 % plus faibles pour ceux qui n'étudiaient pas à temps plein à la vague 2, l'écart se creusant pour atteindre 17 % à la vague 4 ($P = 0{,}0000$ pour β_{10}, $t = 2, 3, 4$). Parallèlement, les échantillons de répondants sous-représentaient les garçons, les élèves fréquentant les écoles secondaires modernes, ceux dont les résultats d'examens étaient faibles ou inconnus, qui pensaient que les employeurs n'offraient pas de formation appropriée aux jeunes et que faire des plans pour l'avenir était une perte de temps, ceux employés à temps plein, ceux qui avaient connu le chômage et ceux qui étudiaient pour obtenir un certificat d'études professionnelles. Le biais était particulièrement prononcé en ce qui concerne les qualifications. Les élèves n'ayant aucun résultat d'examens ou des résultats très faibles étaient environ 50 % plus susceptibles d'avoir quitté l'échantillon aux vagues 3 et 4 que les membres de l'échantillon ayant de meilleurs résultats. De même, les répondants employés à temps plein étaient 17 % plus susceptibles que ceux qui ne travaillaient pas (dont la plupart poursuivaient des études) de décrocher à la vague 2, l'écart se creusant pour atteindre 22 % à la vague 4.

Le fait d'inclure des renseignements contextuels utilisés par l'YCS pour la pondération (sexe, type d'école, résultats des examens et région) dans les modèles n'a eu d'incidence sur le biais pour aucune des caractéristiques (pour chaque vague et pour chaque question, valeur $P > 0{,}05$ pour le test de Wald de l'égalité du coefficient β_{10} estimé avec et sans les caractéristiques contextuelles; données non présentées), sauf pour le biais en ce qui concerne les qualifications, lequel a diminué légèrement lorsque les renseignements contextuels ont été inclus.

L'ordre de grandeur du biais d'attrition est demeuré pour la plupart stable d'une vague à l'autre, sauf pour quelques caractéristiques. Dans l'échantillon témoin d'élèves plus performants interviewés par la poste, la sous-représentation des garçons s'est accrue d'une manière significative de la vague 2 à la vague 4 (valeur P pour le test de Wald de l'égalité de β_{10} entre les deux vagues = 0,0295; données non présentées). Pour certaines autres caractéristiques, le biais a diminué significativement d'une vague à l'autre. Le biais de non-réponse associé à la fréquentation d'une école moderne a diminué de la vague 3 à la vague 4 et le biais associé au fait de n'avoir aucune qualification a diminué de la vague 2 à la vague 3, puis de nouveau de la vague 3 à la vague 4.

Tableau 2 Effet des primes d'encouragement sur les taux d'attrition

		Groupes témoins		Primes d'encouragement		Primes selon l'aptitude		Primes selon le mode de collecte et l'aptitude	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Vague	P_x^{ha}	$\frac{P_x^{la} - P_x^{ha}}{P_x^{ha}}$	$\frac{T_x^{ha} - P_x^{ha}}{P_x^{ha}}$	$\frac{P_u^{ha} - P_x^{ha}}{P_x^{ha}}$	$\frac{(P_c^{ha} - P_x^{ha})/P_x^{ha}}{(4)^{ha}}$	$\frac{(4)^{la}}{(4)^{ha}}$	$\frac{(P_c^{la} - P_x^{la})/P_x^{la}}{(P_c^{ha} - P_x^{ha})/P_x^{ha}}$	$\frac{(T_u^{ha} - T_x^{ha})/T_x^{ha}}{(4)^{ha}}$	$\frac{(T_u^{la} - T_x^{la})/T_x^{la}}{(T_u^{ha} - T_x^{ha})/T_x^{ha}}$
2	21,77	0,6112	0,0650	-0,3777	0,4966	0,7602	0,9763	0,4669	0,8471
(Valeur *P*)		(0,0000)	(0,2268)	(0,0000)	(0,0142)	(0,5085)	(0,4332)	(0,0556)	(0,6810)
3	42,86	0,3734	0,0941	-0,3191	0,7066	0,6820	0,6743	0,4074	1,4275
(Valeur *P*)		(0,0000)	(0,0034)	(0,0000)	(0,0592)	(0,7834)	(0,8287)	(0,0057)	(0,0861)
4	56,53	0,2933	-	-0,3040	0,8402	0,6179	0,7340	0,6597	0,8338
(Valeur *P*)		(0,0000)	-	(0,0000)	(0,2244)	(0,2535)	(0,8177)	(0,0911)	(0,9265)

Nota : P = poste, T = téléphone, x = témoin, u = prime inconditionnelle, c = prime conditionnelle, ha = élèves plus performants, la = élèves moins performants. La colonne (1) donne le taux prévu d'attrition pour l'échantillon témoin d'élèves plus performants interviewés par la poste. Les autres colonnes donnent la variation proportionnelle des taux prévus. Les valeurs P des colonnes 2 à 4 représentent les erreurs-types des effets principaux dans le modèle probit; la colonne 5 donne les valeurs P pour un test de Wald de l'égalité des coefficients pour les primes conditionnelles et inconditionnelles; les colonnes 7 à 9 donnent les valeurs P pour les interactions pertinentes calculées en utilisant la commande « predictnl » de la version 9 de Stata, conformément à Norton et coll. (2004).

Tableau 3 Effet des primes d'encouragement sur la non-réponse partielle (nombres)

		Groupes témoins		Primes d'encouragement		Primes selon l'aptitude		Primes selon le mode de collecte et l'aptitude	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Vague	P_x^{ha}	$\frac{P_x^{la} - P_x^{ha}}{P_x^{ha}}$	$\frac{T_x^{ha} - P_x^{ha}}{P_x^{ha}}$	$\frac{P_u^{ha} - P_x^{ha}}{P_x^{ha}}$	$\frac{(P_c^{ha} - P_x^{ha})/P_x^{ha}}{(4)^{ha}}$	$\frac{(4)^{la}}{(4)^{ha}}$	$\frac{(P_c^{la} - P_x^{la})/P_x^{la}}{(P_c^{ha} - P_x^{ha})/P_x^{ha}}$	$\frac{(T_u^{ha} - T_x^{ha})/T_x^{ha}}{(4)^{ha}}$	$\frac{(T_u^{la} - T_x^{la})/T_x^{la}}{(T_u^{ha} - T_x^{ha})/T_x^{ha}}$
2	2,89	0,2068	-0,9579	0,1008	1,3849	2,4825	0,6927	0,1820	-0,4094
(Valeur *P*)		(0,0005)	(0,0000)	(0,0173)	(0,4790)	(0,1308)	(0,6472)	(0,6251)	(0,9202)
3	2,54	0,3879	-0,8828	0,1660	1,5599	1,6788	1,2445	-0,9526	-0,1378
(Valeur *P*)		(0,0001)	(0,0000)	(0,0049)	(0,2372)	(0,4339)	(0,6796)	(0,0442)	(0,3890)
4	1,75	0,4533	-	0,0085	17,5491	16,8405	0,4621	13,8706	2,3073
(Valeur *P*)		(0,0013)	-	(0,9262)	(0,2133)	(0,4724)	(0,6481)	(0,5049)	(0,4530)

Nota : *P* = poste, *T* = téléphone, *x* = témoin, *u* = prime inconditionnelle, *c* = prime conditionnelle, *ha* = élèves plus performants, *la* = élèves moins performants. La colonne (1) donne le nombre prévu de questions manquantes pour 44 questions non ramifiées à la vague 2, 48 à la vague 3 et 46 à la vague 4. Les autres colonnes donnent la variation proportionnelle des nombres prévus de non-réponse partielle. Les valeurs *P* des colonnes 2 à 4 représentent les erreurs-types des coefficients exponentiés du modèle de comptage; la colonne 5 représente les valeurs *P* d'un test de Wald de l'égalité des coefficients exponentiés pour les primes conditionnelles et inconditionnelles; les colonnes 7 à 9 représentent les valeurs *P* pour les interactions pertinentes calculées en utilisant la commande « predictnl » de la version 9 de Stata, conformément à Norton et coll. (2004).

Tableau 4 Biais d'attrition (groupe témoin des élèves plus performants interviewés par la poste)

	Vague 2	Valeur *P*	Vague 3	Valeur *P*	Vague 4	Valeur *P*
Sexe masculin	0,0807	(0,0000)	0,1330	(0,0000)	0,1474	(0,0000)
Type d'école						
Polyvalente 16	0,0196	(0,2645)	0,0102	(0,6178)	0,0259	(0,2060)
Polyvalente 18	-0,0197	(0,1966)	-0,0138	(0,4444)	-0,0200	(0,2650)
Sélective	-0,0188	(0,3661)	-0,0547	(0,0407)	-0,0213	(0,4577)
Moderne	0,2310	(0,0001)	0,2423	(0,0004)	0,1597	(0,0261)
Indépendante	-0,0142	(0,4639)	0,0147	(0,5245)	-0,0068	(0,7756)
Résultats des examens						
5 notes A à C ou plus	-0,0977	(0,1778)	-0,0866	(0,3060)	-0,1795	(0,0320)
1 à 4 notes A à C	0,0831	(0,2857)	0,0721	(0,4298)	0,1696	(0,0606)
5 notes D à G ou plus	0,0324	(0,8769)	0,0715	(0,7739)	0,1849	(0,4536)
1 à 4 notes D à G	-0,2177	(0,0000)	0,5714	(0,0000)	0,4347	(0,0000)
Aucun	0,7826	(0,0000)	0,5716	(0,0000)	0,4348	(0,0000)
Attitudes						
Les employeurs n'offrent pas de formation	0,0842	(0,0000)	0,0882	(0,0000)	0,0798	(0,0001)
La formation est plus importante que la rémunération	0,0108	(0,4808)	-0,0070	(0,6979)	-0,0062	(0,7370)
Les plans d'avenir sont une perte de temps	0,0656	(0,0959)	0,1457	(0,0015)	0,1371	(0,0030)
Information au sujet des possibilités	0,0034	(0,8431)	-0,0204	(0,3266)	-0,0236	(0,2549)
Suffisamment de soutien pour planifier l'avenir	0,0063	(0,6771)	0,0043	(0,8233)	-0,0105	(0,5848)
Activité courante						
Études à temps plein	-0,1371	(0,0000)	-0,1462	(0,0000)	-0,1728	(0,0000)
Travail à temps plein	0,1661	(0,0003)	0,1983	(0,0001)	0,2201	(0,0000)
Pas d'emploi, d'études ni de formation	0,0898	(0,1387)	0,1036	(0,1495)	0,1098	(0,1184)
Chômeur selon l'OIT	0,0112	(0,6272)	0,0573	(0,0421)	0,0475	(0,0879)
Chômeur au cours des 12 derniers mois	0,0246	(0,4216)	0,0731	(0,0523)	0,0891	(0,0146)
Études pour l'obtention d'un certificat d'études générales	-0,1173	(0,0000)	-0,1351	(0,0000)	-0,1341	(0,0000)
Études pour l'obtention d'un certificat d'études professionnelles	0,0677	(0,0001)	0,0882	(0,0000)	0,0721	(0,0003)
Mode de vie						
Vit avec les parents	-0,1348	(0,0111)	-0,1916	(0,0027)	-0,1033	(0,0986)
Vit avec un(e) conjoint(e)	0,0904	(0,4457)	-0,0441	(0,7475)	-0,1042	(0,4525)

Nota : Écart prévu entre les taux d'attrition fondés sur $\hat{\beta}_{10,t}$, c'est-à-dire prédiction pour chaque catégorie comparée à toutes les autres. Chaque entrée du tableau provient d'un modèle différent, comme il est expliqué dans le texte. Valeurs *P* fondées sur les erreurs-types estimées du coefficient pour la caractéristique dans le modèle probit.

Pour les élèves moins performants, les différences de grandeur du biais d'attrition étaient moindres (données non présentées). Le biais selon le sexe, c'est-à-dire l'écart entre les taux de non-réponse prévus chez les garçons et les filles, était inférieur de 12 % à celui observé pour les élèves plus performants à la vague 4 (la valeur *P* de l'interaction entre le niveau de performance et le sexe était 0,0425 pour $t = 4$) et le biais pour les jeunes travaillant à temps plein était

inférieur de 4 % à la vague 2 (valeur $P = 0{,}0269$ pour $t = 2$); le biais en fonction des attitudes relatives à la formation fournie par les employeurs était 9 % plus élevé à la vague 2 (valeur $P = 0{,}0056$); le biais selon que le répondant étudiait pour l'obtention d'un certificat d'études générales ou d'études professionnelles était supérieur à la vague 2 (22 % et 13 %), inférieur de 6 % et supérieur de 1 % à la vague 3, et inférieur à la vague 4 (81 % et 92 %).

Le biais d'attrition pour le mode de collecte par téléphone ne différait pas de celui pour le mode de collecte par la poste, sauf en cas de non-réponse différentielle selon le sexe : le biais était 7 % plus faible à la vague 2, 2 % plus faible à la vague 3 et 1 % plus élevé à la vague 4 (la valeur P de l'interaction entre le mode de collecte par téléphone et le sexe était $\leq 0{,}002$ pour $t = 2, 3, 4$).

6. Évaluation des hypothèses

Les résultats dont nous discutons ici sont résumés au tableau 2 (effets des primes d'encouragement sur l'attrition), tableau 3 (effets sur la non-réponse partielle), tableau 4 (effets sur le biais d'attrition) et tableau 5 (effet net sur la non-réponse totale et la non-réponse partielle).

*H*1 : *Effets des primes d'encouragement sur le taux d'attrition, le biais d'attrition et la non-réponse partielle.*

Les primes d'encouragement ont réduit l'attrition et augmenté la non-réponse partielle, mais n'ont pas eu d'incidence sur le biais d'attrition. Les primes d'encouragement inconditionnelles ont réduit l'*attrition* cumulative dans l'échantillon d'élèves plus performants interviewés par la poste (tableau 2, colonne 4) de 38 % (ce qui correspond à un écart de 8 points de pourcentage) à la vague 2, de 32 % à la vague 3 et de 30 % à la vague 4 (valeur P de $\beta_2 = 0{,}0000$ pour $t = 2, 3, 4$). Parallèlement, les primes d'encouragement ont augmenté la *non-réponse partielle* de 10 % à la vague 2 et de 17 % à la vague 3 (valeur P de $\beta_2 \leq 0{,}05$ pour $t = 2, 3$), mais n'ont eu aucun effet à la vague 4 (tableau 3, colonne 4). L'écart d'une vague à l'autre n'était toutefois pas significatif (voir *H*2).

Les primes d'encouragement ont eu un effet proportionnel sur l'attrition pour toutes les caractéristiques testées des répondants et, par conséquent, n'ont pas réduit le *biais d'attrition* : la valeur P de l'interaction des primes d'encouragement inconditionnelles et des caractéristiques des répondants était $> 0{,}05$ pour toutes les caractéristiques et les vagues (données non présentées). Faisait exception la proportion d'élèves dans les écoles « modernes » qui étaient sous-représentés dans chacune des trois vagues. (Les écoles modernes formaient la plus petite catégorie, ne représentant que 2,8 % de l'échantillon de répondants de la vague 1). Les primes d'encouragement inconditionnelles ont réduit ce biais de 60 %, 47 % et 78 % aux vagues 2, 3 et 4, respectivement (valeurs P de l'interaction des primes d'encouragement et de la variable école moderne $\leq 0{,}01$ pour $t = 2, 3, 4$).

Puisque les primes d'encouragement ont eu un effet positif sur la réponse totale et un effet négatif sur la réponse partielle, le tableau 5 donne l'effet net sur la quantité d'information recueillie dans l'enquête. Les avantages des primes d'encouragement en ce qui concerne la non-réponse totale l'emportent clairement sur le coût en non-réponse partielle. Pour chaque membre de l'échantillon émis à la vague 2, les taux prévus de réponse totale et de réponse partielle pour l'échantillon d'élèves plus performants interviewés par la poste impliquait qu'à la quatrième vague, 40 % de réponses valides à une question de plus ont été recueillies dans le cas de primes d'encouragement inconditionnelles que dans le groupe témoin. Pour les élèves moins performants, 50 % d'information supplémentaire a été recueillie grâce aux primes d'encouragement. Il s'agit toutefois d'une mesure grossière de l'effet net des primes d'encouragement, puisque, dans une analyse multivariée ou pour l'analyse des changements, des profils différents de réponse manquante d'une question à l'autre ou d'une vague à l'autre peuvent entraîner la suppression d'un grand nombre de cas par élimination par paire.

*H*2 : *Effets des primes d'encouragement d'une vague à l'autre.*

L'effet sur l'attrition a diminué quelque peu d'une vague à l'autre, tandis que les effets sur la non-réponse partielle et le biais d'attrition sont demeurés constants. Les primes d'encouragement ont réduit l'attrition de 38 % à la vague 2, de 32 % à la vague 3 et de 30 % à la vague 4 (tableau 2, colonne 4). Les effets étaient comparables aux vagues 2 et 3, mais significativement différents entre les vagues 2 et 4 et entre les vagues 3 et 4 (la valeur P d'un test de Wald de l'égalité de β_2 entre les vagues était $\leq 0{,}05$). Alors que l'effet relatif des primes d'encouragement a diminué, l'effet absolu a augmenté d'une vague à l'autre (-17 points de pourcentage à la vague 4, comparativement à -8 et à -14 aux vagues 2 et 3, voir le tableau 5). L'effet des primes d'encouragement sur la *non-réponse partielle* ne différait pas significativement d'une vague à l'autre (la valeur P de l'égalité de β_2 entre les vagues était $> 0{,}05$), quoique les nombres prédits de réponses manquantes à une question ont diminué d'une vague à l'autre. De même, les effets des primes

d'encouragement sur le *biais d'attrition* ne variaient pas d'une vague à l'autre.

*H*3 : *Primes d'encouragement conditionnelles comparativement à inconditionnelles.*
Les primes d'encouragement inconditionnelles ont eu un effet de réduction de l'attrition plus prononcé que les primes conditionnelles, mais des effets semblables sur la non-réponse partielle et le biais d'attrition. Dans le cas des élèves les plus performants, les primes conditionnelles utilisées à la vague 2 ont réduit l'attrition deux fois moins fort que les primes inconditionnelles (tableau 2, colonne 5) et l'écart entre les deux traitements était significatif (la valeur *P* d'un test de Wald de l'égalité de β_2 et β_3 était 0,0142). Simultanément, les primes conditionnelles ont augmenté la *non-réponse partielle* de 38 % de plus que les primes non conditionnelles (tableau 3, colonne 5), mais l'écart n'était pas significatif. Les primes conditionnelles ont réduit quelque peu le *biais d'attrition* pour une seule caractéristique, à savoir les membres de l'échantillon appartenant au groupe témoin qui étudiaient pour l'obtention d'un certificat d'études professionnelles à la vague 1 étaient 6,8 % plus susceptibles de décrocher de l'enquête que ceux ne faisant pas ce genre d'études. Dans le cas des primes d'encouragement conditionnelles, l'écart était de 6,4 % (la valeur *P* de l'interaction des primes conditionnelles avec la caractéristique était $\leq 0{,}05$ pour $t = 2$).

*H*4 : *Effets différentiels selon le mode de collecte.*
Les primes d'encouragement ont eu plus d'effets sur l'attrition et la non-réponse partielle dans le cas de la collecte par la poste que par téléphone, mais n'ont eu d'effet sur le biais d'attrition ni pour l'un ni pour l'autre mode. Dans le cas du mode de collecte par téléphone, les primes d'encouragement inconditionnelles ont eu un effet sur l'*attrition* égal à moins de la moitié de celui qu'elles ont eu dans le cas du mode de collecte par la poste pour le groupe des élèves plus performants (tableau 2, colonne 8). L'écart était significatif à la vague 3 (la valeur *P* de l'interaction entre le mode de collecte par téléphone et les primes d'encouragement inconditionnelles étaient de 0,0057), mais non à la vague 2. À la vague 3, les primes ont augmenté la *non-réponse partielle* de 5 % de moins dans le cas de la collecte par téléphone que dans celui de la collecte par la poste (la valeur *P* de l'interaction était 0,0442), mais l'écart à la vague 2 n'était pas significatif. Le manque d'effet des primes conditionnelles sur l'*attrition* ne variait pas selon le mode de collecte.

*H*5 : *Effets des changements de mode de collecte ou de traitement incitatif.*
Le changement du traitement incitatif ou du mode de collecte n'a pas eu d'effet durable. Le passage de primes conditionnelles aux primes inconditionnelles n'a eu aucun effet durable sur l'*attrition* ni sur la *non-réponse partielle* ($P > 0{,}05$ pour les tests de Wald de l'égalité de β_2 et β_3 pour $t = 3, 4$) et les effets après le changement de traitement étaient semblables à ceux observés pour l'échantillon affecté dès le départ à un traitement par prime d'encouragement inconditionnelle (tableaux 2 et 3, colonne 5). Le passage de la collecte par téléphone à la collecte par la poste n'a pas eu d'effet durable sur l'*attrition* ni sur la *non-réponse partielle* (valeur *P* de l'interaction pour le mode de collecte par téléphone et les primes inconditionnelles $> 0{,}05$ à $t = 4$) et les effets après le changement de mode de collecte ne différaient pas de ceux observés pour l'échantillon affecté dès le départ à la collecte par la poste et à des primes d'encouragement inconditionnelles (tableaux 2 et 3, colonne 8).

*H*6 : *Effets différentiels selon le niveau d'aptitude.*
Les effets des primes d'encouragement étaient les mêmes pour tous les niveaux de performance. Les écarts entre les effets proportionnels des primes inconditionnelles et conditionnelles sur l'*attrition* et la *non-réponse partielle* selon le niveau de performance n'étaient pas significatifs (les colonnes 6 et 7 des tableaux 2 et 3 donnent les valeurs *P* des interactions du niveau de performance avec chacun des traitements incitatifs), puisque les effets absolus étaient comparables. Par exemple, les effets des primes d'encouragement inconditionnelles ont réduit l'attrition à la vague 2 de 8 points de pourcentage chez les élèves plus performants et de 10 points de pourcentage chez les élèves moins performants. Toutefois, puisque le niveau de non-réponse dans le groupe témoin était 61 % plus élevé pour le groupe des élèves moins performants, l'effet absolu comparable implique un effet proportionnel plus faible, égal à 76 % seulement de l'effet pour les élèves plus performants.
De même, la différence entre les modes de collecte ne variait pas selon la performance (la colonne 9 des tableaux 2 et 3 donne les valeurs *P* pour l'interaction entre le niveau de performance, les primes d'encouragement inconditionnelles et le mode de collecte par téléphone) et l'absence d'effet sur le biais d'attrition ne différait pas pour les élèves moins performants (données non présentées).

Tableau 5 Effet net des primes d'encouragement sur la réponse totale et la réponse partielle

		TR cumulatif prévu (%)			N^{bre} moyen prévu de NRP			N^{bre} de questions valides par unité émise à la vague 2 : prime d'encouragement/témoin		
		v2	v3	v4	v2	v3	v4	v2	v3	v4
Élèves plus	P_x	78,23	57,14	43,47	2,89	2,54	1,75	-	-	-
performants	P_u	86,45	70,82	60,66	3,19	2,96	1,77	1,097	1,228	1,395
	P_c	82,32	66,80	57,91	3,30	3,20	2,01	1,042	1,152	1,324
	T_x	76,82	53,11	38,31	0,12	0,30	1,61	-	-	-
	T_u	80,91	59,20	50,69	0,12	0,25	1,80	1,053	1,116	1,317
Élèves moins	P_x	64,93	41,14	26,89	3,49	3,52	2,54	-	-	-
performants	P_u	75,00	53,95	40,63	4,37	4,51	2,91	1,130	1,282	1,498
	P_c	71,35	50,09	40,60	3,83	4,66	2,72	1,090	1,186	1,504
	T_x	65,21	38,53	24,34	0,50	3,48	2,35	-	-	-
	T_u	70,41	49,94	36,99	0,49	3,56	2,99	1,080	1,294	1,498

Nota : TR = taux de réponse, NRP = non-réponse partielle, *T* = téléphone, *P* = poste, *x* = témoin, *u* = prime d'encouragement inconditionnelle, *c* = prime d'encouragement conditionnelle. Les calculs sont fondés sur 44 questions non ramifiées à la vague 2, 48 à la vague 3 et 46 à la vague 4. Le nombre de questions avec réponse valide est calculé selon l'expression TR_4* (44-NRP_2 + 48-NRP_3 + 46-NRP_4).

Tableau 6 Énoncé des questions incluses dans l'analyse du biais de non-réponse

Variable	Énoncé de la question
Résultats des examens de 11e année	« Veuillez nous dire : a) quels sujets du CGES vous avez étudiés en 10e et en 11e années, b) pour quels sujets du CGES vous avez passé un examen et c) vos résultats au CGES (n'inscrivez aucun résultat de reprise obtenu en 11e année). »
Attitudes	« Voici certaines choses que des gens ont dit. Nous aimerions savoir ce que vous en pensez. Veuillez inscrire une croix dans l'une des cases pour chaque énoncé : D'accord, En désaccord, Ne sais pas. »
ATT : employeurs	D'accord : « La plupart des employeurs n'offrent pas aux jeunes le type approprié de formation au travail. »
ATT : formation/ rémunération	D'accord : « Dans ma recherche d'un emploi, je me préoccupe plus d'en trouver un avec offre de formation que de trouver celui qui paie le plus. »
ATT : plans	D'accord : « Je pense que faire des plans d'avenir est une perte de temps. »
ATT : information	D'accord : « Je sais comment obtenir des renseignements au sujet de futures possibilités de travail, de formation ou d'études. »
ATT : soutien	D'accord : « Je reçois suffisamment de soutien pour planifier mon avenir. »
Activité courante :	« Veuillez inscrire une croix dans une case pour nous indiquer votre activité principale à l'heure actuelle : a) sans emploi/chômeur, b) apprentissage moderne, stage de formation national, formation de la jeunesse ou autre formation financée par l'État, c) emploi à temps plein (plus de 30 heures par semaine), d) emploi à temps partiel (s'il s'agit de votre activité *principale*), e) études à temps plein dans une école ou un collège, f) m'occupe de la maison ou de la famille, g) fais autre chose (veuillez préciser). »
Études t.p.	Études à temps plein.
Emploi t.p.	Emploi à temps plein.
PEEF	Pas d'emploi, d'études ni de formation.
Chômeur selon l'OIT	En chômage et à la recherche d'un travail parmi les personnes économiquement actives (variable dérivée de l'YCS).
Chômeur	En chômage durant un ou plusieurs mois d'avril 1999 à mars 2000 : « Nous aimerions également savoir ce que vous avez fait au cours des derniers mois. Veuillez inscrire une croix dans l'une des cases pour chaque mois, pour nous montrer ce que vous avez fait pendant l'entièreté ou la plupart de chaque mois. » Les options de réponse sont les mêmes que pour l'activité courante, y compris « en vacances ».
Études (gén.)	Oui : « À l'heure actuelle étudiez-vous pour l'obtention d'un certificat de niveau GCSE, A/S ou A? ».
Études (prof.)	Oui : « À l'heure actuelle étudiez-vous pour l'obtention de tout certificat GNVQ (General National Vocational Qualifications)? » ou « À l'heure actuelle, étudiez-vous pour l'obtention de tout certificat NVQ (National Vocational Qualification) ou tout autre certificat d'études professionnelles, y compris les certificats BTEC, City & Guilds ou RSA? ».
Ménage	« Qui vit dans le même ménage que vous? a) père, b) beau-père, c) mère, d) belle-mère, e) vos propres enfants, f) frères et sœurs, g) autres personnes (veuillez indiquer le lien avec vous). »
Vit avec les parents	Vit avec une ou plusieurs des personnes suivantes : père,beau-père, mère ou belle-mère.
Vit avec un(e) conjoint(e)	Vit avec son petit ami, sa petite amie, son mari, sa femme ou un(e) conjoint(e).

7. Résumé et discussion

L'étude a fourni de nouvelles preuves des effets des paiements incitatifs continus dans une étude par panel à plusieurs modes de collecte. Nous avons testé les effets des primes d'encouragement sur l'attrition, la non-réponse partielle ou le biais d'attrition et vérifié si ces effets variaient d'une vague à l'autre. Nous avons vérifié si les primes d'encouragement conditionnelles et inconditionnelles avaient des effets semblables, si les effets des primes différaient selon le mode de collecte et le niveau d'aptitude, et si les changements de traitement incitatif ou de mode de collecte avaient un effet durable sur l'effet des traitements au cours des vagues subséquentes.

Les résultats montrent que les primes d'encouragement inconditionnelles réduisent significativement l'attrition et, quoiqu'elles accroissent aussi la non-réponse partielle, l'effet net sur la quantité d'information recueillie durant l'enquête est positive. Les primes d'encouragement ont des effets proportionnels pour une gamme de caractéristiques des répondants et, par conséquent, n'ont pas d'effet sur le biais d'attrition relatif à ces caractéristiques. La non-réponse partielle augmente davantage avec les primes d'encouragement inconditionnelles que conditionnelles et davantage en cas du mode de collecte par la poste que de celui par téléphone. Le biais d'attrition n'est affecté ni par le traitement incitatif ni par le mode de collecte. D'une vague à l'autre, les primes d'encouragement ont eu un effet quelque peu décroissant sur l'attrition, mais des effets semblables sur la non-réponse partielle. L'absence d'effets sur le biais d'attrition était aussi un élément constant d'une vague à l'autre. Les changements de traitement incitatif pour passer d'une prime conditionnelle à une prime inconditionnelle et le passage du mode de collecte par téléphone au mode de collecte par la poste n'ont pas d'effets sur les résultats observés lors de vagues ultérieures.

Les résultats impliquent que les primes d'encouragement offertes aux répondants constituent un moyen efficace de maintenir la taille de l'échantillon d'un panel et d'être certain de l'efficacité des estimations et de la faisabilité des analyses de sous-groupes. Parmi les élèves moins performants, au moins 50 % d'information supplémentaire a été recueillie durant les trois vagues expérimentales, pour ce qui est du nombre de questions à réponse valide par cas émis au départ. Les primes d'encouragement offertes étaient sûres, en ce sens que l'accroissement des taux de réponse n'a pas augmenté par inadvertance le biais de non-réponse relatif aux caractéristiques observées.

Les changements de traitement incitatif n'ont pas eu d'effet durable; cependant, dans la présente étude, le seul changement appliqué était une amélioration pour le répondant, c'est-à-dire le passage d'une prime conditionnelle à une prime inconditionnelle. Les attentes formées sur la base des traitements incitatifs antérieurs pourraient fort bien signifier que les changements ont un effet durable, si ceux-ci réduisent la valeur de la prime d'encouragement aux yeux du répondant (voir, Singer, Van Hoewyk et Maher 1998).

Les primes d'encouragement n'ont pas eu d'effet sur le biais d'attrition. Nous n'avons toutefois pas pu évaluer l'effet sur le biais de non-réponse à la vague 1. Idéalement, nous évaluerions à la fois la grandeur du biais dû à la non-réponse à la vague 1 et dû à l'attrition subséquente, et les effets des primes d'encouragement sur les deux. Il est possible que la non-réponse à la vague 1 soit plus nuisible en ce qui concerne le biais que l'attrition ultérieure, surtout dans les études telles que celle décrite ici, dont le taux de réponse initial est faible. Le cas échéant, l'effet des primes d'encouragement sur le biais à la vague 1 pourrait être plus important que tout effet sur le biais d'attrition. En outre, la discussion des effets des primes d'encouragement sur le biais d'attrition s'est concentrée entièrement sur les caractéristiques observées et, si les primes ne présentaient pas d'effet différentiel en ce qui concerne ces caractéristiques, elles pourraient néanmoins avoir des effets différentiels en ce qui concerne les facteurs inobservés. Le cas échéant, l'offre de primes d'encouragement aux répondants pourrait introduire un biais de sélection d'échantillon dans les estimations multivariées, si les caractéristiques inobservées déterminant l'absence de sensibilité aux primes d'encouragement sont corrélées aux résultats mesurés par l'enquête (Kennedy 2003). Par exemple, si la sensibilité aux primes d'encouragement dépend de préférences temporelles pour l'argent et que ce facteur détermine aussi la décision d'arrêter d'étudier et de travailler à la place, alors les modèles des déterminants des résultats en matière d'éducation produiront des estimations biaisées.

Enfin, il se dégage peu de preuves que l'échantillon de répondants est devenu moins sensible aux primes d'encouragement d'une vague à l'autre, à mesure que des membres de l'échantillon éventuellement moins déterminés à participer ont abandonné l'enquête. Ce constat concorde avec celui de Laurie (2007), qui rapporte qu'un accroissement de la valeur d'une prime d'encouragement dans la British Household Panel Survey a augmenté significativement la réponse, même après 14 vagues du panel, dont le taux de réponse annuel, de l'ordre de 95 %, était déjà élevé. Puisque des études antérieures ont montré que les effets d'une prime d'encouragement unique peuvent se transmettre d'une vague à l'autre (James 1997; Laurie et Lynn sous presse; Mack et coll. 1998), un test formel des effets marginaux des primes d'encouragement nécessiterait toutefois des comparaisons avec un groupe de traitement auquel n'est offert une prime d'encouragement qu'à la première vague.

Remerciements

Nous remercions Iain Noble, du Department for Children Schools and Families, d'avoir facilité l'accès aux données et formulé des commentaires au sujet d'une version antérieure de l'article, Tim Thair et Rory Fitzgerald, de leurs réponses patientes et constructives à nos questions sur les données, Noah Uhrig et Heather Laurie de leurs commentaires et Mark Bryan de ses conseils pour le calcul des interactions dans les modèles non linéaires. Les opinions exprimées sont celles des auteurs et ne représentent pas forcément celles du Department for Education and Skills.

Bibliographie

AAPOR (2006). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 4ème Édition : American Association for Public Opinion Research.

Armstrong, J.S. (1975). Monetary incentives in mail surveys. *Public Opinion Quarterly*, 39, 111-116.

Berk, M.L., Mathiowetz, N.A., Ward, E.P. et White, A.A. (1987). The effect of prepaid and promised incentives: Results of a controlled experiment. *Journal of Official Statistics*, 3, 449-457.

Church, A.H. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, 57, 62-79.

Davern, M., Rockwood, T.H., Sherrod, R. et Campbell, S. (2003). Prepaid monetary incentives and data quality in face-to-face interviews: Data from the 1996 survey of income and program participation incentive experiment. *Public Opinion Quarterly*, 67, 139-147.

Downs, B. (1999). *Incentive Use in Panel Surveys*. Internal Census Bureau Memorandum. Washington, DC : US Census Bureau.

Fox, R.J., Crask, M.R. et Kim, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for inducing response. *Public Opinion Quarterly*, 52, 467-491.

Gouldner, A. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25, 161–178.

Goyder, J. (1994). An experiment with cash incentives on a personal interview survey. *Journal of the Market Research Society*, 36, 360-366.

Groves, R., Singer, E. et Corning, A. (2000). Leverage-salience theory of survey participation: Description and an illustration. *Public Opinion Quarterly*, 64, 299-308.

Groves, R.M., et Couper, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York : John Wiley & Sons, Inc.

Hopkins, K.D., et Gullickson, A.R. (1992). Response rates in survey research: A meta-analysis of the effects of monetary gratuities. *Journal of Experimental Education*, 61, 52-62.

James, J.M., et Bolstein, R. (1990). The effect of monetary incentives and follow-up mailings on the response rate and response quality in mail surveys. *Public Opinion Quarterly*, 54, 1 septembre 1990, 346-361.

James, T.L. (1997). Results of wave 1 incentive experiment in the 1996 survey of income and program participation. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Alexandria, VA : American Statistical Association.

Kennedy, P. (2003). *A Guide to Econometrics*. Oxford : Blackwell.

Laurie, H. (2007). The effect of increasing financial incentives in a panel survey: An experiment on the british household panel survey, Wave 14. *ISER Working Paper* No. 2007-05. Colchester : University of Essex.

Laurie, H., et Lynn, P. (sous presse). The use of respondent incentives on longitudinal surveys. Dans *Methodology of Longitudinal Surveys*, (Éd. P. Lynn), Chichester : Wiley.

Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA : Sage.

Lynn, P., Purdon, S., Hedges, B. et McAleese, I. (1994). England and Wales Youth Cohort Study: An Assessment of Alternative Weighting Strategies. Employment Department Research Series YCS No. 30. Sheffield : Employment Department.

Mack, S., Huggins, V., Keathley, D. et Sundukchi, M. (1998). Do monetary incentives improve response rates in the survey of income and programme participation? *Proceedings of the Survey Research Methods Section*, American Statistical Association. Alexandria, VA : American Statistical Association.

Martin, E., Abreu, D. et Winters, F. (2001). Money and motive: Effects of incentives on panel attrition in the survey of income and program participation. *Journal of Official Statistics*, 17, 267-284.

Mason, R., Lesser, V. et Traugott, M.W. (2002). Effect of item nonresponse on nonresponse error and inference. Dans *Survey Nonresponse*, (Éds., R.M. Groves, D.A. Dillman, J.L. Eltinge, et R.J.A. Little), New York : John Wiley & Sons, Inc., 149-161.

Norton, E.C., Wang, H. et Ai, C. (2004). Computing interaction effects and standard errors in logit and probit models. *The Stata Journal*, 4, 154-167.

Rodgers, W. (2002). Size of incentive effects in a longitudinal study. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Alexandria, VA : American Statistical Association.

Russell, N., et Phelps, A. (2001). Youth Cohort Study Cohort 10 Sweep 1: Rapport technique. Londres : Taylor Nelson Sofres.

Ryu, E., Couper, M.P. et Marans, R.W. (2006). Survey incentives: Cash *vs*. in-kind; face-to-face *vs*. mail; Response rate *vs*. nonresponse error. *International Journal of Public Opinion Research*, 18, 89-106.

Shettle, C., et Mooney, G. (1999). Monetary incentives in U.S. government surveys. *Journal of Official Statistics*, 15, 231-250.

Singer, E., Hoewyk, J.V., Gebler, N., Raghunathan, T. et McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics*, 15, 217-230.

Singer, E., Van Hoewyk, J. et Maher, M.P. (1998). Does the payment of incentives create expectation effects? *Public Opinion Quarterly*, 62, 1 juin 1998, 152-164.

Singer, E., Van Hoewyk, J. et Maher, M.P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, 64, 1 juin 2000, 171-188.

Stratford, N., Simmonds, N. et Nicolaas, G. (2003). *National Travel Survey 2002*: *Report on Incentives Experiment*. Londres : National Centre for Social Research.

Teisl, M.F., Roe, B. et Vayda, M. (2005). Incentive effects on response rates, data quality, and survey administration costs. *International Journal of Public Opinion Research*, 18, 364-373.

Tzamourani, P., et Lynn, P. (1999). The Effect of Monetary Incentives on Data Quality - Results from the British Social Attitudes Survey 1998 Experiment. CREST document de travail No. 73. Oxford : University of Oxford.

Ward, R.K., Boggess, S., Selvavel, K. et McMahon, M.F. (2001). The use of targeted incentives to reluctant respondents on response rates and data quality. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Alexandria, VA : American Statistical Association.

Willimack, D.K., Schuman, H., Pennell, B.-E. et Lepkowski, J.M. (1995). Effects of a prepaid nonmonetary incentive on response rates and response quality in a face-to-face survey. *Public Opinion Quarterly*, 59, 1 mars 1995, 78-92.

Yu, J., et Cooper, H. (1983). A quantitative review of research design effects on response rates to questionnaires. *Journal of Marketing Research*, 20, 36-44.

Techniques d'enquête, juin 2008
Vol. 34, N° 1, pp. 131-140
Statistique Canada, N° 12-001-X au catalogue

Trouver l'équilibre entre les divers objectifs du plan d'échantillonnage de la National Health and Nutrition Examination Survey

Leyla Mohadjer et Lester R. Curtin [1]

Résumé

La National Health and Nutrition Examination Survey (NHANES) fait partie d'une série de programmes ayant trait à la santé parrainés par le National Center for Health Statistics des États-Unis. Une caractéristique unique de la NHANES est que tous les répondants de l'échantillon sont soumis à un examen médial complet. Afin de normaliser la façon dont ils sont effectués, ces examens ont lieu dans des centres d'examen mobiles (CEM). L'examen comprend des mesures physiques, des tests tels que l'examen de la vue et des dents, ainsi que le prélèvement d'échantillons de sang et d'urine pour des analyses biologiques. La NHANES est une enquête annuelle continue sur la santé réalisée auprès de la population civile des États-Unis ne résidant pas en établissement. Les principaux objectifs analytiques comprennent l'estimation du nombre et du pourcentage de personnes dans la population des États-Unis et dans des sous-groupes désignés qui présentent certaines maladies et certains facteurs de risque. Le plan d'échantillonnage de la NHANES doit permettre d'établir un juste équilibre entre les exigences liées à l'obtention d'échantillons annuels et pluriannuels efficaces et la souplesse requise pour pouvoir modifier les paramètres essentiels du plan afin de mieux adapter l'enquête au besoin des chercheurs et des décideurs qui élaborent les politiques en matière de santé. Le présent article décrit les défis associés à la conception et à la mise en œuvre d'un processus d'échantillonnage permettant d'atteindre les objectifs de la NHANES.

Mots clés : Échantillonnage à plusieurs degrés; échantillonnage par domaine; mesure pondérée de la taille; centres d'examen mobiles.

1. Introduction

La National Health and Nutrition Examination Survey (NHANES) fait partie d'une série de programmes liés à la santé parrainés par les Centers for Disease Control and Prevention des États-Unis par l'entremise du National Center for Health Statistics (NCHS). La NHANES est utilisée depuis plus de 45 ans pour évaluer l'état de santé et l'état nutritionnel de la population civile des États-Unis ne résidant pas en établissement. Les données recueillies servent à estimer la prévalence des principales maladies et des principaux facteurs de risque de maladie. Les données sur la nutrition permettent une surveillance temporelle de la population nationale en ce qui a trait à des facteurs tels que le régime alimentaire, le taux de cholestérol, l'hypertension, la carence en fer, l'anémie et l'obésité. La NHANES est également conçue en vue d'évaluer la relation entre le régime alimentaire, la santé et l'environnement, afin de pouvoir établir le lien entre les évaluations nutritionnelles et des affections comme la maladie cardiovasculaire, le diabète, l'hypertension et l'ostéoporose.

La collecte des données de la NHANES comprend au moins trois étapes, à savoir un questionnaire de sélection des ménages, une interview et un examen médical. L'objectif principal du questionnaire de sélection est de déterminer si un membre du ménage est admissible à l'interview et à l'examen médical. Le questionnaire de sélection vise à recueillir des renseignements de base sur la composition et les caractéristiques démographiques du ménage. L'interview est conçue pour recueillir des données au niveau du ménage, de la famille et de la personne sur les caractéristiques démographiques et socioéconomiques, la santé et les caractéristiques nutritionnelles. À la fin de l'interview, il est demandé au répondant de participer à un examen médical. Afin de normaliser la façon dont ils sont administrés et les protocoles, ces examens se déroulent dans un centre d'examen mobile (CEM) spécialement conçu et équipé. L'examen comprend des mesures physiques, des tests comme un examen des yeux et des dents, des mesures physiologiques et le prélèvement d'échantillons de sang et d'urine pour des analyses biologiques. Le site Web de la NHANES (http://www.cdc.gov/nchs/nhanes.htm) fournit des renseignements détaillés sur les composantes médicales de l'enquête.

L'élaboration d'un plan d'échantillonnage efficace a nécessité la résolution de plusieurs questions de conception particulières à la NHANES en plus de celles qui se posent habituellement en échantillonnage. Le présent article traite des aspects uniques et compliqués du plan d'échantillonnage de la NHANES. Néanmoins, nous estimons qu'il est utile de commencer par un résumé général de ce plan d'échantillonnage, ce que nous faisons ci-après, avant de discuter de ses caractéristiques uniques.

L'échantillon de la NHANES représente l'ensemble de la population civile ne résidant pas en établissement des États-Unis. Les militaires d'active et les personnes placées

1. Leyla Mohadjer, Westat, 1650 Research Blvd., Rockville, Maryland, États-Unis 20850. Courriel : LeylaMohadjer@Westat.com; Lester R. Curtin, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, Maryland, États-Unis 20782. Courriel : lrc2@cdc.gov.

en établissement ne font pas partie de la population d'inférence. L'échantillonnage de la NHANES ne se fait pas selon un plan à probabilité égale; les fractions d'échantillonnage sont établies de manière à « suréchantillonner » les Mexicano-Américains (et les Hispaniques dans les échantillons de 2007 et au-delà), les Américains de race noire, les Américains de race blanche et autres à faible revenu, les personnes de moins de 20 ans et les personnes de plus de 60 ans. Un plan d'échantillonnage à quatre degrés est utilisé. Les unités primaires d'échantillonnage (UPE), souvent appelées emplacements de collecte *(stands)*, sont sélectionnées à partir d'une base de sondage comprenant tous les comtés des États-Unis. Les UPE sont principalement des comtés uniques; dans quelques cas, des comtés adjacents sont fusionnés afin que la taille des UPE demeure supérieure à une taille minimale fixée. La base de sondage de la NHANES contient près de 3 000 UPE qui sont échantillonnées avec probabilité proportionnelle à une mesure de taille (PPT). Chaque échantillon annuel comprend 15 emplacements de collecte.

Le deuxième degré d'échantillonnage correspond à la sélection de segments de régions (*area segments*) constitués d'îlots de recensement ou de combinaison d'îlots. Comme les UPE ne sont pas toutes de même taille, la taille des segments et le nombre de segments par UPE présentent une certaine variabilité. Chaque segment est formé d'environ 150 ménages (ou unités de logements) en moyenne, environ 5 000 segments sont créés dans chaque UPE et 24 segments sont échantillonnés en moyenne. L'échantillonnage est conçu de façon à ce que la taille d'échantillon soit approximativement la même dans chaque UPE et la plupart des UPE comptent exactement 24 segments. Les segments sont également sélectionnés avec probabilité proportionnelle à la taille. Les mesures de taille des segments, lorsqu'elles sont combinées aux taux de sous-échantillonnage dans les segments, fournissent des nombres approximativement égaux de personnes échantillonnées (PE) par segment, quoique la variation relative de la charge de travail soit plus grande dans les segments que dans les UPE.

Le troisième degré d'échantillonnage consiste à sélectionner les ménages et les logements collectifs non institutionnels, comme les dortoirs. Dans une UPE donnée, après la sélection de segments, toutes les unités de logement (UL) comprises dans les segments échantillonnés sont répertoriées et un sous-échantillon de ménages et de logements collectifs compris dans les UL sont désignés pour une présélection afin d'identifier les PE éventuelles pour l'interview et l'examen médical. Les PE comprises dans les ménages ou les logements collectifs constituent le quatrième degré d'échantillonnage. Tous les membres admissibles d'un ménage sont répertoriés et un sous-échantillon de personnes est sélectionné. Les taux de sous-échantillonnage des ménages dans les segments et de personnes dans les ménages sont prédéterminés. La combinaison de la présélection et des taux d'échantillonnage différentiels fournit l'accroissement de taille d'échantillon pour les sous-domaines démographiques d'intérêt (âge, sexe, race/ethnicité et revenu). Ainsi, dans les 30 UPE dans lesquelles les données ont été recueillies durant le cycle de collecte de deux ans de 2005-2006, 716 segments ont été sélectionnés et 26 529 ménages ont été tirés pour la présélection. Après la présélection pour déterminer la composition selon l'âge, le sexe et la race/ethnicité et la situation de faible revenu, une ou plusieurs personnes ont été sélectionnées dans 6 372 ménages pour faire partie de l'échantillon. En tout, 12 862 personnes ont été sélectionnées, et parmi celles-ci, 9 950 ont répondu à l'interview et subi les examens.

Les examens de la NHANES requièrent du personnel hautement spécialisé, ainsi que l'analyse en laboratoire des échantillons prélevés. Par conséquent, la mise en œuvre des composantes de l'examen médical peut être très coûteuse. Afin de limiter les coûts et de réduire le fardeau de réponse, certaines composantes de l'examen médical ne sont administrées qu'à un sous-échantillon des répondants qui se présentent au CEM. Un seul algorithme de sous-échantillonnage sert à contrôler le degré de chevauchement entre les divers sous-échantillons afin qu'il soit possible d'analyser les corrélations entre les divers examens et composantes de laboratoire. L'affectation des PE aux sous-échantillons est déterminée entièrement avant qu'elles se présentent au CEM.

Les données recueillies dans le cadre des enquêtes de la NHANES ont joué un rôle extrêmement important dans l'obtention des renseignements nécessaires sur la santé et l'état nutritionnel de la population des États-Unis. Par conséquent, à partir de 1999, la NHANES est devenue une enquête annuelle permanente (Mohadjer et Khare 1998). Il est essentiel d'accorder beaucoup d'attention à l'élaboration et à la tenue à jour d'un plan d'échantillonnage efficace dans le cas d'une enquête aussi importante et complexe. Le présent article décrit les défis posés par la conception et la mise en œuvre d'un processus d'échantillonnage permettant d'atteindre les multiples objectifs de la NHANES. Il porte sur le plan d'échantillonnage utilisé jusqu'en 2006 (afin de répondre aux nouvelles exigences analytiques, certains aspects du plan ont été modifiés à partir de 2007).

La section 2 décrit les principaux objectifs de l'enquête et la section 3 donne un aperçu des facteurs les plus importants ayant une incidence sur le plan d'échantillonnage. La section 4 décrit les caractéristiques uniques du plan d'échantillonnage de la NHANES. Enfin, la section 5 résume brièvement l'article.

2. Principaux objectifs de la NHANES

La NHANES est une enquête annuelle permanente sur la santé réalisée auprès de la population civile ne résidant pas en établissement des États-Unis. Ses principaux objectifs sont : 1) estimer la prévalence nationale de certaines maladies et certains facteurs de risque, 2) estimer les distributions de référence dans la population nationale de certains paramètres de la santé et contaminants présents dans l'environnement, 3) décrire et étudier les raisons des tendances séculaires de certaines maladies et certains facteurs de risque, 4) contribuer à la compréhension des causes des maladies, 5) étudier l'évolution naturelle de certaines maladies, 6) étudier la relation entre le régime alimentaire, la nutrition, l'environnement, la génétique et la santé et 7) explorer les questions de santé publique naissantes.

3. Principaux facteurs ayant une incidence sur le plan d'échantillonnage

Comme nous l'avons mentionné plus haut, une caractéristique unique de la NHANES est qu'un examen médical complet est effectué dans les centres d'examen mobiles. En outre, le plan doit produire des tailles d'échantillon efficaces pour un grand nombre de sous-domaines de la population générale. Beaucoup de caractéristiques de la santé et de la nutrition diffèrent considérablement selon l'âge, le sexe et la race ou ethnicité, et dépendent aussi de la situation de revenu. Par conséquent, la plupart des analyses des données de la NHANES sont effectuées pour des groupes d'âge particuliers dans divers sous-groupes socio-économiques de la population. L'enquête est donc conçue afin de produire des tailles d'échantillon efficaces pour un très grand nombre de sous-domaines de la population des États-Unis.

En général, le plan d'échantillonnage de la NHANES doit permettre d'établir un juste équilibre entre les exigences liées à l'obtention d'échantillons de sous-domaine efficaces, d'une part, et d'une charge de travail pouvant être gérée par le personnel du CEM, d'autre part, tout en maintenant les taux de réponse aussi élevés que possible. Plus précisément, le plan d'échantillonnage de la NHANES vise à 1) obtenir des tailles préspécifiées d'échantillons autopondérés pour un ensemble d'environ 75 sous-domaines prédésignés, 2) produire des tailles d'échantillon par UPE donnant lieu à une charge de travail gérable pour les intervieweurs et le personnel du CEM, 3) obtenir des échantillons susceptibles de produire des taux de réponse élevés, 4) être aussi rentable que possible, 5) produire des échantillons annuels efficaces, 6) permettre le cumul d'échantillons au cours du temps, surtout pour les sous-domaines ou les maladies rares et 7) être souple afin de permettre la modification des paramètres clés, y compris les domaines d'échantillonnage et les taux d'échantillonnage en vue de répondre aux nouvelles questions en matière de santé.

La suite de la section est consacrée à un bref résumé de l'incidence de chacun de ces sept objectifs sur la conception et la mise en œuvre de la NHANES.

Sous-domaines de la NHANES - Le plan d'échantillonnage de la NHANES permet d'atteindre un niveau préspécifié de précision pour les données transversales et les comparaisons au cours du temps pour un ensemble de sous-domaines prédésignés. Plus précisément, 77 domaines d'échantillonnage (dans l'échantillon de 2006) sont définis en fonction de la race/ethnicité, le sexe, l'âge, le revenu et la situation de grossesse. Les Noirs, les Mexicains, les enfants très jeunes, les adolescents, les personnes âgées, les femmes enceintes et les personnes à faible revenu sont suréchantillonnés.

Quand les estimations de totaux d'univers pour l'ensemble de la population sont considérées comme étant de la plus haute importance, la meilleure estimation disponible du total de population est utilisée comme mesure de taille dans le processus d'échantillonnage. Dans le cas de la NHANES, où l'on s'intéresse à des sous-domaines de la population totale, une autre mesure de taille est nécessaire pour améliorer l'exactitude des estimations et permettre de mieux contrôler la taille de l'échantillon. La section 4 décrit la mesure de taille utilisée pour l'échantillonnage des UPE et des segments dans le cadre de la NHANES.

L'objectif du suréchantillonnage (en utilisant des probabilités de sélection différentielles) est de tirer un échantillon contenant des nombres proportionnellement plus élevés de membres de certains sous-domaines de population que n'en contient la population. Le but est d'obtenir des tailles d'échantillon adéquates pour faire des inférences pour des sous-domaines représentant une proportion relativement faible de l'univers d'intérêt total et de le faire de façon à réduire au minimum les variances, compte tenu du budget de l'enquête. Diverses stratégies de suréchantillonnage sont utilisées, selon le domaine d'intérêt. Par exemple, le suréchantillonnage des sous-populations minoritaires est réalisé par stratification des régions géographiques selon la concentration de ces groupes minoritaires et par sélection des segments à un taux plus élevé dans les régions à forte concentration. Par ailleurs, un grand échantillon de sélection peut être nécessaire pour le suréchantillonnage des personnes appartenant à des groupes d'âge particuliers. La sous-section sur les ratios des coûts qui figure plus bas décrit pourquoi les procédures de suréchantillonnage appliquées dans le cas de la NHANES diffèrent de celles habituellement utilisées pour de nombreuses enquêtes par sondage à base aréolaire.

Charge de travail des centres d'examen mobiles (*CEM*) - Les CEM sont constitués de quatre remorques spécialement conçues et équipées et contiennent tout l'équipement médical. Chaque remorque mesure environ 45 pieds de long et 10 pieds de large. Un camion tracteur détachable conduit les remorques d'un emplacement à un autre. Les CEM se rendent aux divers emplacements de collecte à travers le pays. Les remorques sont installées côte à côte et jointes par des passerelles fermées. L'espace à l'intérieur du CEM est divisé en salles pour permettre le respect de la vie privée durant les examens et les interviews. L'examen comprend diverses évaluations et mesures physiques et dentaires, des analyses de laboratoire et des interviews sur la santé.

Étant donné les difficultés logistiques associées à l'utilisation des CEM, pour chaque emplacement échantillonné, la taille d'échantillon doit être déterminée d'avance et considérée comme fixe, afin que la planification des opérations sur le terrain soit efficace et pratique. En outre, il est nécessaire d'établir un calendrier ferme pour chaque emplacement de collecte, afin que les rendez-vous puissent être pris pour les examens. Il est impossible de modifier le calendrier, car celui-ci doit être coordonné avec les visites du CEM à d'autres emplacements dont le calendrier est également préétabli.

Taux de réponse - L'obtention de taux de réponse élevés est une préoccupation dans le cas de presque toutes les enquêtes par sondage. Dans celui de la NHANES, le défi est particulièrement grand, étant donné la portée des interviews et des examens. L'offre d'une rémunération a été utilisée comme moyen d'améliorer les taux de réponse. En outre, la NHANES comprend un programme de relations communautaires important englobant des contacts avec les organismes locaux et les personnes dont il faut obtenir la coopération, ainsi que la couverture dans les médias locaux afin de joindre un aussi grand nombre que possible de PE. Dans le contexte des questions soulevées par le plan d'échantillonnage, une approche qui s'est avérée avoir un effet favorable sur les taux de réponse a été la sélection d'échantillons de plus grande taille dans les ménages échantillonnés. L'un des facteurs soupçonnés d'être à l'origine de l'accroissement des taux de réponse dans les ménages comptant plusieurs PE est que chaque personne reçoit un dédommagement pour son temps et sa participation et qu'il est généralement plus commode pour les membres du ménage de se rendre ensemble au CEM. Le tableau 1 donne le taux de réponse à la composante de l'examen des PE provenant des ménages dans lesquels une seule personne a été sélectionnée comparativement au taux pour les PE provenant des ménages où plusieurs personnes ont été sélectionnées. Comme le montre le tableau, les taux de réponse augmentent d'environ 4 à 7 % selon le type de ménage.

Par conséquent, la NHANES est conçue en vue de maximiser le nombre de PE par ménage. Cette approche est faisable dans le cas d'études de ce genre où l'échantillon est constitué d'un grand nombre de sous-domaines. Autrement dit, l'effet de la mise en grappes dans les ménages n'est pas très préoccupant, parce que la plupart des analyses sont faites dans des sous-domaines âge-sexe particuliers (ou dans certains groupes limités de sous-domaines) et que la mise en grappes dans les ménages est généralement faible au niveau du sous-domaine. Le nombre moyen de PE sélectionnées par ménage (dans les ménages où au moins une PE a été sélectionnée) dans les domaines définis d'échantillonnage varie de 1 à 1,24 dans l'échantillon pour 1999 à 2006. Le regroupement des domaines, afin de réduire leur nombre entre 12 et 15, en fonction de l'âge et/ou de la race/ethnicité donne des nombres moyens variant de 1,01 à 1,37 PE par ménage. Par conséquent, un certain niveau de mise en grappes existe dans la mesure où des domaines groupés sont utilisés pour l'analyse. Il convient de souligner que l'échantillon de PE est utilisé fondamentalement pour l'analyse au niveau des PE (par exemple, statistiques sur la santé et la nutrition). La mise en grappes des PE est évidemment plus importante aux niveaux de la famille et du ménage. Cependant, des variables au niveau du ménage ou de la famille sont utilisées pour ce genre d'analyse (par exemple, niveau de poussière dans le ménage, revenu familial ou assurance familiale). Consulter Curtin et Mohadjer (2008) pour une discussion de l'effet de la mise en grappes et des probabilités inégales de sélection des sous-domaines sur les niveaux de précision des diverses estimations.

Tableau 1
Examen des taux de réponse selon le nombre de PE dans le ménage, par type de ménage, dans l'échantillon de la NHANES pour 1999 à 2006

Type de ménage	Nombre de PE sélectionnées par ménage		Taux de réponse (%)	
	Une PE	Deux PE ou plus	Une PE	Deux PE ou plus
Noir/mexicain	4 892	20 222	76,5	82,3
Autre à faible revenu[1]	1 362	3 349	77,6	84,5
Autre pas à faible revenu	5 597	15 508	68,8	72,6

[1] Le groupe Autre comprend toutes les PE qui ne sont ni noires ni mexicaines. Le seuil de faible revenu est fixé à 130 % du seuil de pauvreté.

Ratio des coûts - Dans les échantillons d'enquête aréolaire, le coût des opérations de collecte des données sur le terrain comprend le coût d'établissement des listes d'unités de logements, la présélection des ménages afin de repérer les répondants admissibles et la réalisation de l'interview pour recueillir les données. Dans le cas de la NHANES, la phase de collecte des données comprend

l'interview auprès du ménage ainsi que l'examen au CEM. La NHANES requiert de l'équipement médical, du personnel et des processus de laboratoire hautement spécialisés. Par conséquent, le coût d'un examen est très élevé comparativement à d'autres coûts d'enquête. En fait, le coût de l'établissement des listes et de la présélection ne représente que de 3 % à 4 % du coût de l'interview et de l'examen. Ce ratio des coûts (le coût de l'interview et de l'examen relativement au coût de l'établissement des listes et de la présélection) a une incidence très importante sur le plan d'échantillonnage de la NHANES.

Comme nous l'avons mentionné plus haut, une méthode de suréchantillonnage doit être appliquée à un grand nombre de sous-domaines prédésignés de l'enquête afin d'atteindre les tailles requises d'échantillon. Pour les populations minoritaires, il est possible de réduire considérablement la présélection en suréchantillonnant les régions à forte concentration de groupes minoritaires. En général, un plan optimal est élaboré en déterminant l'effet, sur les coûts et la variance, de diverses procédures d'échantillonnage et en choisissant celle qui réduit la variance au minimum pour un coût fixé. Lors de l'évaluation des compromis entre les coûts et la variance, supposons qu'une stratégie de suréchantillonnage particulière réduise le nombre de ménages à inscrire sur la liste et à soumettre à la présélection, tout en accroissant la variance de la plupart des statistiques. Les économies réalisées grâce à la réduction des coûts d'établissement de la liste et de présélection pourraient être utilisées pour accroître la taille de l'échantillon et, donc, réduire la variance. Cependant, dans la NHANES, le listage et la présélection des ménages représentent une très faible fraction du coût, de sorte que les économies relatives à ces opérations doivent être très importantes pour justifier un accroissement modéré de la variance. Par conséquent, les procédures de suréchantillonnage établies pour l'enquête reflètent le ratio des coûts de la NHANES et diffèrent de celles appliquées dans le cas d'enquêtes aréolaires typiques.

Échantillons annuels et pluriannuels - Afin de faciliter les couplages éventuels avec d'autres enquêtes de grande portée, de maintenir la souplesse du plan d'échantillonnage et de permettre la production d'estimations annuelles pour de grands sous-domaines, la NHANES est devenue une enquête annuelle permanente à partir de 1999. Les déplacements requis aux États-Unis pour obtenir des échantillons annuels représentatifs de la population nationale posent un défi de taille. Trois CEM, dont, en tout temps, deux sont stationnés dans des UPE et le troisième se déplace, fonctionnent selon un calendrier minutieusement établi afin de répondre aux exigences du plan de l'étude.

Dans le cas de toute enquête, la capacité de faire des inférences significatives dépend à la fois de la précision des estimations proprement dites et de la précision des estimations de la variance des estimations utilisées dans l'analyse. L'une des principales limites de l'échantillon annuel de la NHANES est le petit nombre d'UPE (15 par année), qui donne un petit nombre de degrés de liberté pour l'estimation ainsi que l'analyse, si bien que les estimations de la variance par rapport au plan sont relativement imprécises. En outre, les tailles effectives d'échantillon pour la plupart des sous-domaines dans les échantillons annuels sont trop faibles. La plupart des analyses au niveau du sous-domaine devront être faites en cumulant un certain nombre d'échantillons annuels afin d'obtenir une précision et une puissance statistique suffisantes pour les comparaisons. Les procédures suivies pour combiner les échantillons annuels doivent être relativement simples et adaptées aux progiciels du commerce afin de maximiser l'utilité pour une grande variété d'utilisateurs des données de la NHANES. Par conséquent, il est essentiel de concevoir un plan d'échantillonnage qui permet de cumuler efficacement les échantillons annuels au cours des années.

Plan d'échantillonnage souple - Un objectif clé de la NHANES est d'explorer les questions d'actualité en matière de santé publique. L'enquête doit être souple afin que l'on puisse l'adapter à l'évolution des exigences et aux nouveaux défis. Donc, le plan d'échantillonnage doit tenir compte à la fois du besoin d'échantillons de sous-domaine efficaces et de souplesse afin de pouvoir modifier les paramètres clés de l'enquête. Jusqu'à présent, le plan d'échantillonnage existant de la NHANES a permis d'intégrer certaines modifications des définitions des sous-domaines et des taux d'échantillonnage, lorsque ces modifications ont été apportées après la sélection des UPE. Toutefois, dans des circonstances extrêmes, des changements importants dans les définitions des sous-domaines ou les exigences relatives aux tailles d'échantillons nécessiteraient la sélection d'un nouvel échantillon d'UPE.

4. Caractéristiques uniques du plan d'échantillonnage de la NHANES

Les facteurs décrits à la section 3 ont joué un rôle important dans l'élaboration du plan d'échantillonnage et ont donné lieu à certaines caractéristiques qui sont propres au plan de la NHANES. Ces caractéristiques uniques sont les suivantes : 1) mesure de taille pondérée des UPE et des segments, 2) échantillons annuels et pluriannuels efficaces, 3) nombre maximisé de personnes échantillonnées par ménage, 4) tailles d'échantillon contrôlées pour les UPE, 5) attribution progressive de l'échantillon de l'UPE, 6) méthodes spéciales pour tenir compte de la détérioration de l'efficacité du plan de sondage optimal au cours du temps et 7) méthodes spéciales afin de réduire le risque de divulgation de données par identification géographique.

Les paragraphes qui suivent décrivent brièvement les caractéristiques uniques du plan d'échantillonnage de la NHANES.

Mise en grappes et mesures de taille - Dans la NHANES, la taille d'échantillon doit être suffisamment grande pour produire une charge de travail efficace dans chaque UPE, compte tenu du temps et du coût de déplacement d'un CEM entre deux emplacements de collecte et du temps nécessaire pour monter le CEM et le démonter pour le déplacement. Selon l'expérience acquise lors de réalisations antérieures de la NHANES, un nombre moyen de 340 PE examinées est un nombre approximativement optimal qui produit le nombre maximal d'UPE, tout en maintenant la taille d'échantillon dans chaque région suffisamment grande pour justifier les coûts du déménagement du CEM. En outre, dans le cas de la NHANES, les UPE sont habituellement définies comme des comtés individuels afin de réduire le temps de déplacement des répondants pour se rendre au CEM et donc accroître la probabilité d'obtenir des taux de réponse élevés.

L'échantillon de la NHANES est conçu pour produire un échantillon autopondéré pour chaque sous-domaine échantillonné, tout en créant une charge de travail efficace dans chaque UPE. Les UPE et les segments sont sélectionnés avec une probabilité proportionnelle à une mesure de taille pondérée qui reflète la population de l'UPE dans les sous-domaines d'intérêt. La probabilité de sélection d'une UPE détermine le taux maximal auquel les personnes résidant dans cette UPE particulière peuvent être sélectionnées. Voir le document *Vital and Health Statistics*, *Series 2*, *No. 113*, *September 1992*, *CDC/NCHS*, consultable à http://www.cdc.gov/nchs/products/pubs/pubd/series/sr02/120-101/120-101.htm pour une description des mesures de taille utilisées dans le cadre de la NHANES.

Échantillons annuels et pluriannuels et stratification - Un moyen de réaliser des échantillons annuels représentatifs de la population nationale consiste à sélectionner un échantillon indépendant d'UPE. Étant donné le nombre limité d'UPE de la NHANES et le fait que ces dernières sont sélectionnées avec probabilité proportionnelle à la taille, cette approche donnerait vraisemblablement lieu à un chevauchement important des UPE d'une année à l'autre. Le chevauchement des échantillons, même au niveau des UPE, pourrait entraîner une perte de précision des estimations calculées d'après les données de l'enquête si les échantillons de plusieurs années de référence sont combinés (à cause de l'accroissement de la mise en grappes dans l'échantillon). Donc, au lieu d'échantillonner les UPE indépendamment chaque année, il a été décidé de sélectionner pour la NHANES un échantillon de six ans, à partir d'une structure hiérarchique de strates principales et secondaires (décrite plus loin), puis à affecter une UPE de chacune des strates principales à chaque année. Cette structure hiérarchique de l'échantillon de six ans évite le chevauchement des UPE non autoreprésentatives durant les six années.

L'échantillon de six ans de la NHANES est sélectionné selon un plan stratifié à deux UPE par strate qui a été élaboré en ayant pour objectif principal l'efficacité de l'échantillon de six ans, ainsi que celle des échantillons pluriannuels. Le plan de stratification est conçu de façon que les UPE formant les échantillons annuels et pluriannuels soient réparties uniformément en fonction de certaines caractéristiques géographiques et démographiques.

Le plan d'échantillonnage de la NHANES comprenait (jusqu'à 2006 inclusivement) 18 UPE autoreprésentatives. Ces UPE variaient de celles qui étaient autoreprésentatives pour les échantillons annuels à celles qui l'étaient pour les échantillons de trois ans ou de six ans. Ces UPE ont été affectées de telle façon que le nombre d'UPE autoreprésentatives soit le même chaque année, les UPE autoreprésentatives des échantillons de trois ans étant espacées de trois ans. Chaque strate principale comprenait six strates secondaires et une UPE a été sélectionnée dans chacune de ces strates finales. Dans chaque strate principale, les strates secondaires ont été appariées pour créer des pseudo-strates. Chaque paire a été assignée aléatoirement à l'étude avec un intervalle de trois ans. L'affectation des paires aux ensembles particuliers d'années de référence et l'affectation des années de référence dans les paires a été faite aléatoirement dans la première strate principale et le même schéma a été suivi dans toutes les autres.

Ce plan de stratification a donné un échantillon de 72 UPE non autoreprésentatives qui permettent de produire des estimations annuelles et pluriannuelles efficaces sans compromettre l'efficacité des estimations sur six ans. L'échantillon de six ans est obtenu selon un plan d'échantillonnage d'une UPE par strate secondaire (ou un plan d'échantillonnage de deux UPE par pseudo-strate) et chaque échantillon annuel, selon un plan d'échantillonnage d'une UPE par strate principale. En plus, ce plan offre la souplesse nécessaire pour répondre aux changements d'exigences d'échantillonnage (si un nouvel échantillon doit être sélectionné), puisque pendant les trois premières années, l'échantillon correspond à un plan d'échantillonnage d'une UPE par pseudo-strate.

Nombre maximisé de personnes échantillonnées par ménage - Après avoir obtenu l'échantillon de ménages présélectionnés, on sélectionne un échantillon de personnes qui seront interviewées et soumises à l'examen médical. La liste de tous les membres admissibles d'un ménage est dressée et un sous-échantillon de personnes est sélectionné en fonction du sexe, de l'âge, de la race ou de l'ethnicité et du revenu (toutes les femmes enceintes sont sélectionnées avec certitude). Les PE sont sélectionnées à des taux établis

afin d'être certain que soient réalisées les tailles d'échantillons cibles par sous-domaine.

L'échantillon de PE est sélectionné de façon à maximiser le nombre moyen de personnes sélectionnées par ménage afin d'accroître le taux global de réponse à l'enquête. Si l'on recourait à des sélections aléatoires indépendantes pour les sous-domaines, dans la plupart des cas, une seule personne serait sélectionnée par ménage et la taille moyenne d'échantillon par ménage serait assez faible, à peine supérieure à 1. Par conséquent, au lieu d'une randomisation non limitée, on utilise une procédure de pseudorandomisation afin de maximiser le nombre de PE par ménage. Consulter Waksberg et Mohadjer (1991) pour une description de l'approche.

Tailles contrôlées d'échantillon par UPE - La taille d'échantillon dans chaque UPE (emplacement de collecte) qui est effectivement générée d'après un échantillon autopondéré dans chaque domaine est basée sur un certain nombre d'hypothèses, dont la distribution par âge et par race/ethnicité de la population de l'UPE. Ces hypothèses ne sont vérifiées qu'approximativement. Une fois calculées, les tailles d'échantillon sont traitées comme des quotas et le nombre de PE dans chaque emplacement de collecte est forcé de correspondre étroitement au quota. On procède ainsi afin de s'assurer que les opérations sur le terrain soient gérables et efficaces. Il est nécessaire d'établir un calendrier ferme pour chaque emplacement de collecte afin que des rendez-vous puissent être donnés pour les examens des PE. Le calendrier tient naturellement compte du nombre prévu de PE à chaque emplacement de collecte. Comme nous l'avons mentionné plus haut, il est difficile de modifier le calendrier d'un emplacement, puisqu'il doit être coordonné avec les visites du CEM à d'autres emplacements dont le calendrier est également préétabli.

Il n'y a aucun moyen de savoir d'avance si le quota assigné pour un emplacement particulier est plus faible ou plus élevé que celui que l'on obtiendrait à partir d'échantillons autopondérés dans les divers domaines. L'incertitude tient en partie au fait que la mesure de taille utilisée pour sélectionner l'échantillon est basée sur le recensement décennal le plus récent et n'est donc peut-être pas à jour. Le problème est compliqué davantage par les variations des taux de réponse d'un emplacement de collecte à l'autre, ainsi que par la variation d'échantillonnage du nombre de PE identifiées. Par conséquent, il est nécessaire d'utiliser une méthode d'échantillonnage qui peut produire des échantillons un peu plus grands ou un peu plus petits que ceux résultant de l'application des taux d'échantillonnage autopondérés.

Affectation séquentielle de l'échantillon dans chaque emplacement de collecte - Afin de réaliser l'objectif susmentionné, on sélectionne dans chaque emplacement de collecte un échantillon initial en appliquant des taux d'échantillonnage supérieurs de 50 % à ceux requis pour obtenir les tailles d'échantillon cibles dans chaque domaine. Chaque échantillon initial d'emplacement de collecte est ensuite divisé en un groupe de sous-échantillons. Chaque sous-échantillon est un sous-échantillon systématique de l'échantillon initial, où les ménages sont ordonnés selon le numéro de segment et un numéro de série provisoire, basé sur les caractéristiques géographiques, avant le sous-échantillonnage. Donc, chaque sous-échantillon recoupe l'ensemble des segments, sauf si l'on est limité par la taille d'échantillon.

En règle générale, le sous-échantillon de 50 % (c'est-à-dire le sous-échantillon A) est le premier qui est attribué aux intervieweurs. Le rendement pour ce sous-échantillon est surveillé et utilisé pour projeter les estimations du nombre total prévu de PE lorsque la sélection de ce sous-échantillon sera achevée. D'après ces chiffres, des sous-échantillons supplémentaires sont attribués au besoin. L'échantillon est surveillé quotidiennement afin de déterminer si l'attribution de sous-échantillons supplémentaires est nécessaire.

Le problème opérationnel que pose la procédure de surveillance du rendement de l'échantillon tient au fait qu'elle ne permet pas de contrôler entièrement les tailles d'échantillon des sous-domaines. La distribution des sous-domaines diffère, dans une certaine mesure, des nombres prévus d'après les données de recensement les plus récentes (utilisées pour calculer les taux d'échantillonnage). Les leçons tirées de la NHANES indiquent qu'il faut s'attendre à certains changements de population qui auront une incidence sur les tailles d'échantillon. D'autres facteurs qui influent sur le rendement d'échantillon de sous-domaine sont les profils de non-réponse et de sous-dénombrement dans les emplacements de collecte. Une option en vue de corriger le déficit (ou l'excédent) dans les tailles d'échantillon de sous-domaine consiste à modifier les taux d'échantillonnage pour les futurs emplacements de collecte. Cependant, de tels changements augmenteront l'hétérogénéité des poids d'échantillonnage, donc auront un effet indésirable sur la précision des estimations au niveau du sous-domaine et ne sont pas conseillés, sauf dans des circonstances extrêmes.

Traitement de la détérioration de l'efficacité du plan optimal au cours du temps dans un échantillon étroitement contrôlé - Dans le cas des échantillons aréolaires, la pratique habituelle consiste à dresser la liste de tous les ménages dans les segments d'échantillon et à appliquer un taux d'échantillonnage préspécifié aux ménages énumérés. Cette approche donne à tous les ménages la probabilité de sélection souhaitée. Par exemple, si le taux d'échantillonnage est de 50 %, alors la moitié des unités de logement énumérées dans les segments seront incluses dans l'échantillon. Si le nombre d'unités de logement a triplé à cause de

constructions neuves (c'est-à-dire des unités de logement construites depuis le recensement décennal le plus récent), le même taux d'échantillonnage produira un nombre trois fois plus grand d'interviews et d'examens médicaux que le nombre prévu au départ. Des changements aussi spectaculaires de la taille des segments n'est pas surprenante lorsque la période de collecte des données est ultérieure de plusieurs années au recensement décennal le plus récent pour lequel des fichiers de données sont disponibles.

Dans le cas de la NHANES, les tailles d'échantillons ne peuvent pas varier fortement, à cause des calendriers établis pour les CEM. Le sous-échantillonnage dans les UPE pour essayer d'obtenir des échantillons de même taille dans toutes les UPE n'est pas recommandé non plus, car il introduirait des facteurs de pondération inégaux qui réduiraient l'efficacité de l'échantillon.

Le programme de la NHANES a utilisé deux méthodes pour mettre à jour les mesures de taille des segments, à savoir 1) la création de segments de constructions neuves et 2) l'échantillonnage à deux phases pour mettre à jour la mesure de taille. Une troisième approche consistant à acheter des listes d'adresses commerciales pour mettre à jour la mesure de taille dans un plan d'échantillonnage à deux phases est à l'étude.

Sous l'approche des constructions neuves (Bell, Mohadjer, Montaquila et Rizzo 1999), les unités qui viennent d'être construites sont exclues des segments de région et de nouveaux segments sont créés en se basant sur l'information du U.S. Census Bureau sur les permis émis pour des constructions neuves depuis le recensement décennal le plus récent. Les segments de constructions neuves comprennent des grappes de permis de bâtir émis durant un ou plusieurs mois contigus par un bureau d'octroi de permis de bâtir. Les fichiers de la Building Permits Survey menée par le Census Bureau servent de sources de données sur le nombre de permis de bâtir résidentiels émis par les bureaux d'octroi de permis de bâtir.

L'échantillonnage à deux phases est utilisé dans un certain nombre d'applications statistiques. L'une d'elles est la mise à jour d'une base de sondage lorsque l'échantillon doit être sélectionné en fonction d'une mesure de taille, mais qu'une estimation fiable de la mesure de taille n'est pas disponible. Selon cette approche, un échantillon plus grand d'unités (de segments dans le cas de la NHANES) est sélectionné. Une valeur mise à jour de la mesure de taille est alors recueillie pour cet échantillon plus grand (également appelé échantillon de première phase). L'échantillon final d'unités (de segments) est sélectionné à partir de l'échantillon de première phase en utilisant la mesure de taille mise à jour.

À partir de 2000, la mesure de taille des segments de la NHANES a été mise à jour (pour les emplacements de collecte pour lesquels cette mise à jour semblait nécessaire) en utilisant une méthode d'échantillonnage à deux phases (Montaquila, Bell, Mohadjer et Rizzo 1999). Dans ces cas, les personnes chargées de dresser les listes de logements se rendent dans l'emplacement de collecte pour obtenir un dénombrement des unités de logement (UL) dans chaque segment de l'échantillon de première phase. Partant de ces dénombrements, une mesure de taille mise à jour reflétant le ratio du nombre actuel d'UL au nombre prévu d'UL est calculée pour chaque segment de première phase. L'échantillon final de segments est alors sélectionné par sous-échantillonnage des segments de première phase en utilisant la mesure de taille mise à jour.

Risque de divulgation des données par identification géographique - De nos jours, les questions de confidentialité et le risque de divulgation des données posent de réels défis aux organismes qui parrainent les enquêtes. La capacité d'identifier les répondants à une enquête, d'après des combinaisons uniques de variables disponibles dans un seul fichier de données ou par couplage de diverses bases de données, est un grave sujet de préoccupation. Il en est particulièrement ainsi de la NHANES, étant donné la grande quantité de renseignements de nature délicate recueillis sur chaque personne échantillonnée et le petit nombre d'UPE dans l'échantillon. Par conséquent, le risque de divulgation est évalué sur deux fronts, à savoir la divulgation géographique et la divulgation d'après les caractéristiques individuelles. Diverses méthodes (diffusion limitée ou suppression de données) sont utilisées par la NHANES pour masquer les caractéristiques individuelles posant un grand risque de permettre d'identifier des personnes qui font partie de l'échantillon. Les éléments de données délicats, à diffusion limitée ou non diffusés, sont consultables dans un centre de données de recherche. À l'heure actuelle, seules des estimations nationales peuvent être produites d'après les fichiers de microdonnées à grande diffusion, et les analyses géographiques détaillées doivent être faites dans le Centre de données de recherche.

Bien que l'on ne puisse produire que des estimations nationales, l'estimation directe des erreurs d'échantillonnage pour ces estimations nécessite la diffusion des variables de plan, comme les identificateurs de strate et d'UPE. Habituellement, ces variables indiquent qu'un groupe de personnes échantillonnées vivent toutes dans le même comté, mais n'identifie pas le comté. La divulgation géographique est une cause de souci particulière, car, dans le cas de la NHANES 1) le nombre d'UPE est petit, 2) les UPE sont limitées géographiquement à un comté et 3) de nombreuses démarches de relations communautaires sont menées dans chaque UPE en vue d'améliorer les taux de réponse. Le programme de relations communautaires comprend la prise de contact avec divers organismes et

individus à chaque emplacement de collecte pour essayer d'obtenir leur appui et l'utilisation des médias (journaux, télévision et radio) afin de rejoindre un nombre aussi grand que possible de PE. Il est, par conséquent, relativement facile de déterminer les comtés compris dans l'échantillon de la NHANES. La composition raciale ou ethnique d'un comté, ainsi que la situation de région statistique métropolitaine ou non métropolitaine fournissent des renseignements suffisants pour apparier correctement une liste de comtés connus à des groupes identifiés comme formant une grappe dans un comté dans le fichier de données à grande diffusion. Pour limiter la divulgation géographique, on recourt à des méthodes de permutation probabiliste des enregistrements au deuxième degré d'échantillonnage (permutation des segments) afin de créer des unités à variance masquée. L'objectif est de réduire le risque d'identifier des individus en masquant leur emplacement. Consulter Park, Dohrmann, Montaquila, Mohadjer et Curtin (2006) pour une description des procédures de permutation appliquées à l'échantillon de la NHANES.

5. Sommaire et conclusion

Une caractéristique unique de la NHANES est l'examen médical complet effectué dans les CEM. En outre, l'enquête est conçue de manière à produire des tailles d'échantillon efficaces pour un grand nombre de sous-domaines de la population des États-Unis, puisque la plupart des analyses des données de la NHANES sont faites pour des groupes d'âge particuliers, dans divers sous-groupes socioéconomiques de la population. Donc, le plan d'échantillonnage de la NHANES doit établir un équilibre entre les exigences liées à l'obtention d'échantillons de sous-domaine efficaces, d'une part, et d'une charge de travail efficace pour le personnel d'examen du CEM, d'autre part, tout en maintenant les taux de réponse aussi élevés que possible. En outre, le plan doit être aussi rentable que possible, produire des échantillons annuels efficaces et permettre le cumul des échantillons au cours du temps pour les sous-domaines ou les maladies rares. De surcroît, le plan doit être souple pour permettre de modifier les paramètres clés, y compris les domaines d'échantillonnage, et les taux d'échantillonnage afin de répondre aux nouvelles questions en matière de santé.

Les exigences susmentionnées se traduisent par un plan d'échantillonnage très complexe dont certaines caractéristiques sont propres à la NHANES. En particulier, l'échantillon courant est conçu afin de produire des échantillons annuels et pluriannuels efficaces. La NHANES utilise des UPE pondérées et des mesures de taille de segment pour produire des échantillons autopondérés pour chaque sous-domaine, tout en produisant une charge de travail efficace dans chaque UPE. Une fois que les tailles d'échantillon sont calculées, elles sont traitées comme des quotas. Les tailles d'échantillon sont strictement contrôlées dans chaque UPS afin que les opérations sur le terrain soient gérables et efficaces. Un très grand échantillon de présélection est utilisé afin de suréchantillonner la plupart des sous-domaines d'âge et de revenu, et le suréchantillonnage des régions à forte concentration est utilisé pour certains sous-domaines minoritaires très rares. L'échantillon de PE est sélectionné selon une méthode pseudo-aléatoire afin de maximiser le nombre moyen de personnes sélectionnées par ménage, parce que cela a semblé accroître le taux global de réponse lors des enquêtes précédentes.

Les défis décrits dans le présent article ont trait aux principaux aspects de la NHANES. Cette dernière possède de nombreux autres caractéristiques uniques dont il faut tenir compte lors de l'analyse des données. Par exemple, chaque échantillon annuel ne contient qu'un très petit nombre d'UPE, mais les données recueillies dans ces UPE ne le sont pas aléatoirement d'une saison à l'autre. En particulier, s'il existe une interaction entre la saison et la région géographique pour une variable d'intérêt, le plan d'échantillonnage actuel de la NHANES ne permettra pas de l'estimer. Étant donné le petit nombre d'UPE dans chaque cycle de diffusion des données, tout couplage contextuel des données au niveau géographique doit se faire au centre de données de recherche du NCHS. Comme le nombre de sous-échantillons de la NHANES est élevé, il convient de prendre tout spécialement soin d'utiliser le poids de sous-échantillon approprié; ainsi, les estimations des cas de diabète non diagnostiqués doivent être calculées en utilisant le poids spécial pour le test effectué à jeun.

Afin de faciliter la bonne utilisation des CEM pour la collecte des données, aucun effort n'a été fait en vue de répartir aléatoirement l'échantillon d'UPE entre les périodes dans les échantillons annuels. Cependant, la dimension temporelle joue un rôle important dans certains indicateurs de la santé, telle que la nutrition. De surcroît, l'analyse des données sur la nutrition peut aussi être influencée par la nature complexe du plan d'échantillonnage et de la collecte des données. Des poids de sondage spéciaux construits pour les deux jours de collecte des données du rappel alimentaire de 24 heures tiennent compte de la variation du nombre d'examens selon le jour de la semaine. Un tutoriel sur Internet est développé à l'heure actuelle afin de faciliter l'analyse des données sur la nutrition de la NHANES. Un tutoriel général concernant l'analyse fondée sur le plan de sondage des données de la NHANES peut être obtenu à http://www.cdc.gov/nchs/tutorials/.

Remerciements

Les auteurs remercient le rédacteur associé et les examinateurs de leurs suggestions et commentaires constructifs qui leur ont permis d'améliorer considérablement l'article.

Bibliographie

Bell, B., Mohadjer, L., Montaquila, J. et Rizzo, L. (1999). Creating a frame of newly constructed units for household surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Curtin, L.R., et Mohadjer, L. (2008). Design trade-offs for the National Health and Nutrition Examination Survey. *Proceedings of the Ninth Conference on Health Surveys Research Methods*, à paraître.

Montaquila, J., Bell, B., Mohadjer, L. et Rizzo, L. (1999). A methodology for sampling households late in a decade. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Montaquila, J., Mohadjer, L. et Khare, M. (1998). The enhanced sample design of the future National Health and Nutrition Examination Survey (NHANES). *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Park, I., Dohrmann, S., Montaquila, J., Mohadjer, L. et Curtin, L.R. (2006). Reducing the risk of data disclosure through area masking: Limiting biases in variance estimation. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Waksberg, J., et Mohadjer, L. (1991). Automation of within-household sampling. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

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 23, No. 4, 2007

Non-Bayesian Multiple Imputation
Jan F. Bjørnstad 433

Discussion
Ray Chambers 453
Ralf T. Münnich 455
Chris Skinner 463
Seppo Laaksonen 467
Daniel Thorburn 477

Rejoinder
Jan F. Bjørnstad 485

The Interdependence of Determinants for the Strength and Direction of Social Desirability Bias in Racial Attitude Surveys
Volker Stocké 493

The Perils of Interpreting Age Differences in Attitude Reports: Question Order Effects Decrease with Age
Bärbel Knäuper, Norbert Schwarz, Denise Park and Andreas Fritsch 515

Dependent Interviewing and Seam Effects in Work History Data
Annette Jäckle and Peter Lynn 529

The Influence of End-Users on the Temporal Consistency of an International Statistical Process: The Case of Tropical Forest Statistics
Alan Grainger 553

Book and Software Reviews 593

Editorial Collaborators 603

Contents

Volume 24, No. 1, 2008

The Morris Hansen Lecture 2004
Bridging the Gap: Moving to the 1997 Standards for Collecting Data on Race and Ethnicity
Jennifer H. Madans 1

Discussion
Clyde Tucker 13
Robert B. Hill 17

A Model-free Bayes Analysis of Stratification, Clustering and Regression in Finite Population Survey Data
Murray Aitkin 21

On the Covariance Between Related Horvitz-Thompson Estimators
John Wood 53

Probability Based Estimation Theory for Respondent Driven Sampling
Erik Volz and Douglas D. Heckathorn 79

Survey Experiences and Later Survey Attitudes, Intents and Behaviour
Lars R. Bergman and Robert Brage 99

Temporal Disaggregation and Seasonal Adjustment
Tommaso Proietti and Filippo Moauro 115

New Approaches to Creating Data for Economic Geographers
Matthew Freedman, Julia Lane and Marc Roemer 133

Book and Software Reviews 157

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 35, No. 4, December/décembre 2007

Bent JØRGENSEN & Peter X.K. SONG
Stationary state space models for longitudinal data....461

Petra BŮŽKOVÁ & Thomas LUMLEY
Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables....485

Jae Kwang KIM & Jay J. KIM
Nonresponse weighting adjustment using estimated response probability....501

Jinhong YOU & Gemai CHEN
On inference for a semiparametric partially linear regression model with serially correlated errors....515

Alexander R. DE LEON & K.C. CARRIÈRE
General mixed-data model: extension of general location and grouped continuous models....533

Cinzia CAROTA
A family of power-divergence diagnostics for goodness-of-fit....549

Chiung-Yu HUANG, Jing QIN & Fei ZOU
Empirical likelihood-based inference for genetic mixture models....563

Xingwei TONG, Chao ZHU & Jianguo SUN
Semiparametric regression analysis of two-sample current status data, with applications to tumorigenicity experiments....575

Jesse FREY
Distribution-free statistical intervals via ranked-set sampling....585

Heng LIAN
Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces....597

Forthcoming papers/Articles à paraître....613

Volume 36 (2008): Subscription rates/Frais d'abonnement....614