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

Techniques d’enquête

Techniques d’enquête
43-1

Date de diffusion : le 22 juin 2017

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

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

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

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête

Nº 12-001-XPB au catalogue

Une revue
éditée
par Statistique Canada

Juin 2017

•

Volume 43

•

Numéro 1

Statistique Canada Statistics Canada

Canada

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

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

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée en version électronique deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préférablement en Word au rédacteur en chef, (statcan.smj-rte.statcan@canada.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca/techniquesdenquete).

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 43, numéro 1, juin 2017

Table des matières

Article sollicité Waksberg

Articles réguliers

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.
∞

Techniques d'enquête, juin 2017
Vol. 43, nº 1, p. 1
Statistique Canada, nº 12-001-X au catalogue

Série Waksberg d'articles sollicités

La revue *Techniques d'enquête* a mis sur pied une série de communications sollicitées en l'honneur de Joseph Waksberg, qui a fait de nombreuses contributions importantes à la méthodologie d'enquête. Chaque année, un éminent chercheur est choisi pour rédiger un article pour la série de communications sollicitées de Waksberg. L'article examine les progrès et l'état actuel d'un thème important dans le domaine de la méthodologie d'enquête et reflète l'agencement de théorie et de pratique caractéristique des travaux de Waksberg.

Ce numéro de *Techniques d'enquête* commence par le quinzième article de la série du prix Waksberg. Le comité de rédaction remercie les membres du comité de sélection, composé de Louis-Paul Rivest (président), Tommy Wright, Kirk Wolter et J.N.K. Rao, d'avoir choisi Don A. Dillman comme auteur de l'article du prix Waksberg de cette année.

Communication sollicitée pour le prix Waksberg 2016

Auteur : Don A. Dillman

Don A. Dillman, Ph.D., est professeur émérite du département de sociologie et directeur adjoint de la recherche au Centre de recherche en sciences sociales et économiques de la Washington State University, où il est membre du corps professoral depuis 1969. Son livre de 1978 sur les méthodes d'enquête par courrier et par téléphone, qui en est à sa quatrième édition sous le titre *Internet, Phone, Mail and Mixed-Mode surveys: The Tailored Design Method*, a servi de guide pour la réalisation d'enquêtes partout dans le monde depuis près de 40 ans. De 1991 à 1995, il a occupé le poste de méthodologiste d'enquête principal au U.S. Census Bureau, assurant la supervision de la conception des procédures de collecte de données pour le recensement décennal, ce qui lui a valu le prix Roger Herriot de 2000 pour l'innovation en statistique fédérale. À la Washington State University, il administre un programme de recherche actif en collecte de données à mode mixte. En 2017, son équipe de recherche a remporté le prix des innovateurs Warren J. Mitofsky de l'American Association for Public Opinion Research pour avoir créé la méthodologie de promotion par le Web décrite dans ce numéro de *Techniques d'enquête*.

Techniques d'enquête, juin 2017
Vol. 43, nº 1, pp. 3-34
Statistique Canada, nº 12-001-X au catalogue

Inciter les participants aux enquêtes à mode mixte à répondre sur le Web : les promesses et les défis

Don A. Dillman[1]

Résumé

La collecte de données par sondage axée sur le Web, qui consiste à prendre contact avec les enquêtés par la poste pour leur demander de répondre par Internet et à retenir les autres modes de réponse jusqu'à un stade ultérieur du processus de mise en œuvre, a connu un essor rapide au cours de la dernière décennie. Le présent article décrit les raisons pour lesquelles cette combinaison novatrice de modes de prise de contact et de réponse aux enquêtes était nécessaire, les principales étant la diminution de l'efficacité de la téléphonie vocale et l'élaboration plus lente que prévu de méthodes de collecte de données par courriel/Internet uniquement. Les obstacles historiques et institutionnels à cette combinaison de modes d'enquête sont également examinés. Vient ensuite une description de la recherche fondamentale sur l'utilisation des listes d'adresses postales aux États-Unis, ainsi que les effets de la communication auditive et visuelle sur la mesure par sondage, suivie d'une discussion des efforts expérimentaux en vue de créer une méthodologie axée sur le Web comme remplacement viable des enquêtes à réponse par téléphone ou par la poste. De nombreux exemples d'usage courant ou prévu de la collecte de données axée sur le Web sont fournis. L'article se termine par une discussion des promesses et des défis considérables qui résultent du recours plus important aux méthodes d'enquête axées sur le Web.

Mots-clés : Enquêtes; mode mixte; axé sur le Web; poste; téléphone; échantillonnage fondé sur les adresses; communication visuelle; taux de réponse; différences de mesure.

1 Introduction

Au début du XXI[e] siècle, la conception des enquêtes a connu une évolution étonnante, mais cruciale, caractérisée par un large recours à des méthodes de collecte de données axées sur le Web, c'est-à-dire l'envoi d'une invitation par la poste à des échantillons du grand public en vue d'obtenir leurs réponses à un questionnaire principalement par Internet plutôt que sur papier. Les méthodes axées sur le Web visent à remplacer aujourd'hui les procédures axées sur l'envoi par la poste, qui consistent à essayer d'obtenir des réponses à un questionnaire papier par la poste avant d'utiliser d'autres modes de réponse, comme l'interview par téléphone ou sur place. Les méthodes axées sur le Web sont maintenant employées dans les enquêtes gouvernementales officielles et pour remplacer les enquêtes téléphoniques par composition aléatoire (CA).

Ainsi, le programme de l'*American Community Survey*, qui est la source principale de renseignements sur les ménages américains au niveau des états et des régions, a commencé à utiliser en 2013 une approche de collecte des données axée sur le Web qui comprend la possibilité de répondre par la poste, par interview téléphonique ou par interview sur place, à une étape ultérieure du processus de mise en œuvre. Les plans sont maintenant établis pour appliquer ce genre de méthodologie pour le recensement décennal des États-Unis de 2020. La collecte de données axée sur le Web après une demande initiale par la poste est également utilisée partout dans le monde. Le Recensement du Japon de 2015 (Statistics Japan 2015), ainsi que les recensements du Canada (Statistique Canada 2016) et de l'Australie (Australian Bureau of Statistics 2016) de 2016 en sont des exemples. Les enquêtes auprès des ménages réalisées en Suisse (Roberts, Joye et Staehli

1. Don A. Dillman, Washington State University. Courriel : dillman@wsu.edu.

2016) et la *Community Life Survey* du Royaume-Uni (United Kingdom Cabinet Office 2016), qui est en train de passer de l'interview sur place à la réponse par Internet, en sont d'autres. En outre, la *U.S. College Graduates Survey*, réalisée tous les deux à trois ans par la *National Science Foundation*, a achevé le passage de la collecte de données par la poste et par téléphone à une approche de collecte axée sur le Web, suivie par les deux autres modes de collecte de données (Finamore et Dillman 2013). Ces exemples ne représentent que quelques-uns des programmes d'enquête importants à travers le monde qui appliquent maintenant cette méthodologie.

Un certain nombre de considérations, dont les problèmes en apparence insolubles liés aux enquêtes téléphoniques par CA et le fait que les listes d'adresses résidentielles du service postal ou les listes d'enregistrement nationales fournissent maintenant la couverture la plus complète des ménages, ont incité à recourir aux méthodes de collecte de données axées sur le Web. En outre, il n'existe aucun moyen acceptable de tirer des échantillons probabilistes d'adresses de courriel des ménages en vue de prendre contact avec ces derniers. Même si ces adresses pouvaient être échantillonnées, il est peu probable que l'on obtiendrait des taux de réponse raisonnables en se limitant à la prise de contact par courriel (Lozar, Bosnjak, Berzelak, Haas et Vehovar 2008).

L'importance actuelle de la prise de contact par la poste est étonnante, même si preuve a été faite à la fin du XX^{e} siècle qu'il était possible d'obtenir des taux de réponse raisonnables aux enquêtes par la poste (Dillman 2000). Jusqu'à récemment, les bases de sondage fondées sur les adresses postales étaient, pour la plupart, non disponibles et inadéquates. De surcroît, l'existence générale d'une option téléphonique avant la fin des années 1990 signifiait que l'envoi postal était utilisé peu fréquemment pour les enquêtes gouvernementales, sauf les recensements officiels.

Mon propos ici est d'abord d'exposer, dans la section 2, les raisons qui ont poussé les enquêteurs à élaborer et à adopter des méthodologies axées sur le Web partout dans le monde. Ensuite, dans la section 3 et la section 4, je décris les travaux de recherche qui ont non seulement permis d'appliquer des méthodologies axées sur le Web, mais aussi de les rendre plus efficaces afin de produire des estimations fiables des opinions et des comportements des populations sondées à travers le monde.

Ces travaux de recherche ont montré, voir la section 5, que les méthodologies axées sur le Web sont assez prometteuses pour ce qui est de l'amélioration des taux de couverture et de réponse, tout en réduisant les différences de mesure entre les modes, ainsi que le coût total des enquêtes. Ils ont aussi révélé que l'utilisation de ces méthodes comporte de nombreux écueils décrits à la section 6, allant de la confiance des répondants dans Internet à la pléthore d'appareils disponibles aujourd'hui pour répondre à ce type d'enquête. Le présent article met avant tout l'accent sur les grandes promesses et les nombreux défis associés aux méthodes d'enquête par sondage axées sur le Web. Un résumé et une conclusion sont présentés à la section 7.

2 Pourquoi la collecte de données axée sur le Web est nécessaire

Fondamentalement, prendre contact avec les ménages ou les particuliers selon un mode, tel l'envoi par la poste ou l'appel téléphonique, pour leur demander de répondre par un autre mode, n'est pas une méthode idéale de collecte des données. Il existe forcément une certaine friction entre la réception d'une lettre ou

d'un appel téléphonique et l'obligation de passer à un mode de réponse différent. Ce passage aura vraisemblablement, en soi, une incidence négative sur les taux de réponse. Donc, il n'est pas étonnant que les difficultés à mener des enquêtes à mode de collecte unique par téléphone ou par courriel/Web soient la raison fondamentale de rechercher une autre solution.

2.1 L'efficacité décroissante des enquêtes par téléphone

Au milieu du XX^{e} siècle, la plupart des méthodologistes considéraient les interviews en face à face comme le seul moyen acceptable de mener des enquêtes par sondage (par exemple Parten 1950; Kerlinger 1965). En outre, l'échantillonnage des ménages et la réalisation d'enquêtes auprès de ceux-ci représentaient un processus lent et coûteux, et étaient par conséquent limités principalement aux enquêtes de portée nationale ou couvrant d'autres grandes régions.

Même si des données étaient parfois obtenues par sondage téléphonique (Nathan 2001), l'essor du téléphone comme seul moyen de recueillir les réponses aux enquêtes n'a eu lieu qu'au début des années 1970, comme il est décrit en détail dans Nathan (2001). Les trois premiers ouvrages sur les méthodes d'enquêtes par téléphone, dont les dates de parution se sont succédées rapidement, établissaient des perspectives d'études de marché (Blankenship 1977), d'enquêtes sur les populations des états et des populations spéciales (Dillman 1978) et d'enquêtes sur les populations nationales (Groves et Kahn 1979). L'utilisation des méthodes de collecte de données par téléphone a progressé rapidement en raison de la présence croissante du téléphone dans les ménages et de l'élaboration de la procédure de Mitofsky-Waxberg en vue d'utiliser des méthodes de composition aléatoire (CA) pour sélectionner les ménages. De surcroît, étant donné la baisse du prix des appels interurbains, des enquêtes téléphoniques par CA ont remplacé la plupart des interviews sur place (Dillman 2005).

De 1997 à 2012, le *Pew Research Center* (2012), un important organisme menant des enquêtes sociales par téléphone aux États-Unis, a fait état d'une baisse des taux de réponse aux enquêtes par CA, ces taux étant passés de 35 % à environ 9 %. Plus récemment, Dutwin et Lavrakas (2016) ont effectué une analyse des taux de réponse par téléphone pour neuf organisations. Ces auteurs ont constaté une baisse des taux de réponse par téléphone fixe, lesquels sont passés de 15,7 % en 2008 à 9,3 % en 2015, tandis que les taux de réponse par téléphone mobile sont passés durant cette période de 11,6 à 7,0 %. Ils ont également signalé que cette baisse d'environ 40 % de la réponse résultait moins d'un accroissement des refus que d'une augmentation de la non-réponse et de l'utilisation de répondeurs, de 10 points de pourcentage pour les lignes fixes et de 24 points de pourcentage pour les téléphones mobiles.

Cependant, ces résultats ne représentent que la partie émergée de l'iceberg pour ce qui est des changements touchant le téléphone. Ce dernier a évolué, pour passer d'un appareil ménager, ou téléphone fixe, partagé par tous les membres du ménage à un instrument sans fil à possession individuelle, facilement transportable d'un lieu à l'autre. Aux États-Unis, la moitié des ménages et 60 % de ceux comptant des enfants n'utilisent aujourd'hui que la téléphonie sans fil (Blumberg et Luke 2017). Parallèlement, la présence de téléphones mobiles et (ou) fixes dans les ménages a atteint un sommet inégalé d'au moins 95 % dans la plupart des pays européens (Mohorko, de Leeuw et Hox 2013) et de 97 % aux États-Unis (Blumberg

et Luke 2017). Une conséquence de la proportion croissante de téléphones mobiles est que l'échantillonnage des ménages est devenu beaucoup plus difficile. Il est possible d'inclure les numéros de téléphone mobile dans les bases de sondage par CA. Toutefois, il est également devenu nécessaire de consacrer des minutes d'interview précieuses à la confirmation d'une gamme de renseignements, y compris le nombre et le type de téléphones dans un ménage afin de déterminer les probabilités de sélection des ménages.

En outre, il faut savoir si la personne qui répond au téléphone est un adulte, et sélectionner un répondant approprié. Qui plus est, le « problème du moment inopportun » lié aux lignes fixes, où le répondant est, par exemple, interrompu quand il prépare le repas ou n'a pas le temps de parler, a pris de l'ampleur, car il faut maintenant chercher à savoir si la personne qui répond au téléphone conduit une automobile ou accomplit une autre tâche posant un problème sérieux de sécurité. Dans le cas d'interviews téléphoniques pour lesquels de fortes pressions existent en vue de maintenir la durée à quelques minutes seulement, l'ajout d'éléments de ce genre réduit la capacité de poser d'autres questions. Bref, un effet important des changements concernant la propriété, la réglementation et l'utilisation des téléphones est que leur usage pour d'importantes opérations de collecte des données devient de plus en plus difficile.

Les téléphones fixes et mobiles posent les uns et les autres un défi plus important. De moins en moins de personnes conversent vocalement par téléphone. Il s'agit d'un énorme changement par rapport à l'époque où le moyen fondamental de communication pour les discussions d'affaires, le maintien des relations sociales et la coordination rapide des activités quotidiennes se faisaient principalement par téléphonie vocale. Les courriels et les messages texte ont largement remplacé cet usage. Parler par téléphone à un intervieweur est de moins en moins au diapason des autres aspects de la vie quotidienne des gens.

Les répondeurs prennent maintenant la plupart des appels vocaux sur les téléphones tant fixes que mobiles. Ne pas répondre au téléphone n'est plus considéré comme une grossièreté. Les appels souhaités des enfants et d'autres proches peuvent être repérés par le destinataire de l'appel au moyen de sonneries spéciales. Les appels provenant de numéros particuliers peuvent également être bloqués ou, sur les téléphones intelligents, balayés. En outre, les numéros de téléphone fixes ainsi que mobiles sont maintenant transportables entre divers types de téléphone et indicatifs régionaux aux États-Unis, et diverses règles fédérales s'appliquent à la composition automatique des numéros de téléphone.

Un autre nouveau problème lié aux téléphones tient au fait que les prises de contact répétées nécessaires pour obtenir des taux de réponse raisonnables pour tous les types d'appareils deviennent de moins en moins efficaces. De plus en plus souvent, les intervieweurs n'ont qu'une seule chance, qui ne dure que quelques secondes, de persuader les personnes de répondre à leurs questions. L'apparition du numéro de téléphone et (ou) de la source de l'appel sur l'écran d'identification du téléphone rend de plus en plus probable l'évitement des appels de suivi. En outre, la pléthore d'appels de télémarketing et de collecte de fonds a créé un contexte dans lequel de moins en moins de personnes sont disposées à répondre au téléphone, et encore moins à être interviewées. Une difficulté supplémentaire associée aux téléphones mobiles est que, étant donné leur usage, il est beaucoup plus probable que les demandes de participer à une enquête arrivent quand le destinataire de l'appel est en plein milieu d'activités d'affaires ou de travail qui ne sont pas propices à prendre le temps de répondre à une interview.

Le déclin des interviews téléphoniques par CA a été ralenti un certain temps en raison d'études qui ont montré que les campagnes intensives de rappel en vue d'augmenter les taux de réponse n'amélioraient pas l'exactitude des résultats (Keeter, Miller, Kohut, Groves et Presser 2000), et d'autres qui ont donné à penser qu'il n'existait pas de lien étroit entre l'existence d'une erreur due à la non-réponse (différences entre les répondants et les non-répondants) et les taux de réponse (Groves et Peytcheta 2008). L'important investissement des organisations dans le matériel et les logiciels téléphoniques, ainsi que dans le personnel spécialisé, qui souvent n'avait jamais procédé à d'autres types de collecte de données d'enquête, a également incité ces organisations à continuer d'utiliser le téléphone. Cependant, la baisse persistante des taux de réponse par téléphone observée ces dernières années par Dutwin et Lavrakas (2016) et les préoccupations concernant les mesures ont rendu moins crédible la réalisation d'enquêtes téléphoniques simples.

2.2 L'émergence plus lente que prévu des enquêtes par courriel/Web seulement

Au milieu des années 1990, période où a débuté la tendance à la baisse des taux de réponse par téléphone, les enquêtes par Internet, le remplacement prévu, ont commencé à prendre rapidement de l'expansion (Dillman 2000, chapitre 11). Pourtant, deux décennies plus tard, leur usage pour les enquêtes auprès de l'ensemble de la population demeure limité.

Aux États-Unis et dans de nombreux autres pays développés, la pénétration d'Internet dans les ménages dépasse maintenant 85 %, taux plus élevé que celui observé pour le téléphone au moment de l'essor rapide des enquêtes téléphoniques au début des années 1970 (Nathan 2001). L'absence d'Internet dans certains ménages (par exemple 41 % des adultes américains de 65 ans et plus, et 26 % des personnes titulaires uniquement d'un diplôme d'études secondaires ou ne l'ayant pas obtenu) demeure préoccupante (Anderson et Perrin 2016), mais ce problème diminue d'année en année. Les compétences d'utilisation d'Internet sont aujourd'hui essentielles au processus d'éducation, aux activités organisationnelles et à l'accès aux services au consommateur. Cependant, les obstacles à l'obtention de réponses par Internet demeurent énormes dans le cadre des enquêtes auprès des ménages lorsque la prise de contact se fait uniquement par courriel.

Pour les adresses de courriel, il n'existe aucun algorithme d'échantillonnage des ménages ou de la population générale donnant une chance non nulle connue d'être sélectionné pour participer à une enquête, comme cela était le cas des appels par composition aléatoire pour les enquêtes téléphoniques. Il n'existe pas de présentation normalisée des adresses de courriel, contrairement à nos numéros de téléphone à 10 chiffres qui identifient un indicatif régional, un central, et les 10 000 numéros possibles dans chaque central. Les membres des ménages sont également susceptibles de posséder plusieurs adresses de courriel, de sorte que les probabilités de rejoindre des ménages particuliers ou d'autres unités de l'échantillon ne peuvent pas être calculées. De surcroît, certains membres de la population la plus versée en informatique, c'est-à-dire les jeunes adultes, a acquis la réputation de réduire au minimum son utilisation des systèmes habituels de courrier électronique. Ces jeunes adultes privilégient fortement Facebook, Snapchat et d'autres applications de messagerie instantanée pour communiquer avec leurs amis et leurs connaissances.

En outre, pour les échantillons aléatoires d'adresses de courriel existantes, les taux de réponse par Internet sont vraisemblablement aussi faibles, voire plus faibles, que ceux obtenus pour les enquêtes téléphoniques aujourd'hui (Lozar et coll. 2008). Et les répondants comptent probablement un nombre disproportionnellement élevé de personnes plus jeunes et plus instruites, en dépit du fait que bon nombre de jeunes gens s'appuient sur d'autres moyens de connexion électronique, ce qui en fait des utilisateurs occasionnels seulement du courrier électronique classique. Sur les ordinateurs personnels, les boîtes de réception sont habituellement des espaces encombrés où le nombre de courriels non sollicités et non souhaités est plus important que ne l'était auparavant le nombre d'appels téléphoniques indésirables. Qui plus est, les courriels sont souvent survolés et supprimés en se basant sur la source seulement ou après n'avoir lu que quelques mots du message d'accompagnement.

L'évolution des technologies informatiques contribue aussi à la non-réponse aux enquêtes par Internet. Les téléphones intelligents que l'on peut mettre dans son sac à main ou dans sa poche ont maintenant une puissance informatique beaucoup plus grande que celle des ordinateurs de bureau au moment où a débuté la réalisation d'enquêtes par Internet (par exemple Friedman 2016). Leur présence constante à portée de la main fait qu'ils sont devenus les premiers appareils de réponse utilisés pour déceler et écarter les demandes indésirables. Certains utilisateurs peuvent attendre de répondre aux demandes de participation à une enquête jusqu'à ce qu'ils aient accès à un ordinateur portable ou un ordinateur de bureau doté d'un clavier complet. Cependant, pour certaines personnes, le téléphone intelligent est maintenant le principal, voire le seul, appareil pour répondre à tous leurs courriels.

Quand le mode principal d'enquête était le téléphone, les intervieweurs pouvaient habituellement aider le répondant à se concentrer sur les questions de l'enquête et le guider tout au long de l'interview. Sur les ordinateurs de bureau, les portables et maintenant les tablettes qui sont utilisés au bureau ou à domicile, le répondant peut souvent atteindre un niveau élevé de concentration mentale. En revanche, à l'ère du téléphone intelligent où les personnes contactées se déplacent vraisemblablement d'un endroit à un autre, il semble un peu moins probable qu'elles arrivent à se concentrer sur le contenu de l'enquête. Or, il est évident que la proportion d'enquêtes auxquelles il est répondu sur un téléphone intelligent augmente relativement à l'ensemble de celles auxquelles il est répondu par Internet (Couper, Antoun et Mavletova, sous presse). Toutefois, il ne semble pas y avoir de preuve que la réception de demandes d'enquête par Internet sur un téléphone intelligent accroisse la réponse totale à l'enquête, et cela pourrait en fait la réduire. De plus, les taux d'abandon sont nettement plus élevés pour les téléphones intelligents que pour les ordinateurs de bureau et les ordinateurs portables.

Les craintes quant aux conséquences d'une tentative de réponse à une enquête électronique est un autre facteur qui limite l'efficacité potentielle de la réalisation d'enquêtes par courriel/Internet. La facilité et le faible coût de l'envoi d'un nombre massif de demandes d'enquête par courriel a accru la probabilité que les personnes reçoivent des demandes en provenance d'organisations dont elles ne savent rien. En outre, la crainte que ce genre de demande puisse provenir de sources imitant des commanditaires légitimes et qu'il s'agisse d'une tentative de livraison d'un maliciel, d'un logiciel de rançon et (ou) de la collecte de données à d'autres fins répréhensibles. Donc, les personnes désireuses et capables de répondre à des enquêtes par

Internet légitimes pourraient ne pas vouloir prendre ce genre de risque. Pour nombre d'entre elles, Internet est un endroit qui fait peur, où « le consommateur doit prendre garde ».

Pour toutes ces raisons, il n'est guère étonnant que les enquêtes à prise de contact par courriel/réponse par Internet, dont le coût est faible, ne soient pas devenues la méthode de choix pour réaliser les sondages aléatoires du grand public nécessaires pour élaborer les politiques publiques. Même si l'on pouvait résoudre la question difficile du tirage d'échantillons probabilistes, de nombreux problèmes, incluant les technologies informatiques, les circonstances dans lesquelles les répondants éventuels reçoivent les demandes d'enquête, et le manque de confiance à l'égard des émetteurs de la demande d'enquête et de la façon dont les données pourraient être utilisées, limitent la capacité qu'a ce mode d'enquête de remplacer le téléphone.

3 Surmonter les obstacles à l'acceptation des plans de collecte de données à modes mixtes

3.1 Les obstacles historiques aux modes mixtes

L'utilisation de plus d'un mode d'enquête, comme moyen de prendre contact et (ou) de poser les questions, était rare à la fin du XX^{e} siècle. Faire accepter les plans de collecte de données à modes mixtes à quelque fin que ce soit a été un lent processus. Avant les années 1990, l'obstacle le plus important était simplement le manque d'avantages perçus. Les taux de réponse aux enquêtes sur place, par téléphone et par la poste étaient considérés suffisamment élevés pour juger l'utilisation d'un deuxième ou d'un troisième mode de collecte comme étant inutile. Une exception importante était le cas des enquêtes où l'on appliquait une méthode moins coûteuse au début du processus de collecte des données, mais où des méthodes d'interview sur place étaient nécessaires pour obtenir des taux de réponse supérieurs à 90 %. Les recensements décennaux des États-Unis au cours de la période de 1970 à 1990, qui comprenaient un questionnaire envoyé par la poste suivi d'une interview sur place et, dans certains cas, d'appels téléphoniques en est un exemple.

Un autre obstacle au mélange des modes d'enquête tenait au fait que la technologie de collecte des données de l'époque rendait difficile la mise en œuvre simultanée de plusieurs modes dans une seule enquête. L'absence d'ordinateurs en réseau et de logiciels signifiait que, pour utiliser un deuxième mode de collecte des données, il fallait achever la collecte des données selon un mode avant de transférer le travail à une unité de collecte de données distincte chargée de mettre en œuvre un deuxième mode (Dillman, Smyth et Christian 2009, chapitre 8). Une revue antérieure de l'utilisation du téléphone dans les enquêtes à modes mixtes à la fin des années 1980 a montré que peu de ces enquêtes avaient été réalisées, à part l'envoi d'une lettre d'introduction avant une interview par téléphone ou sur place prévue (Dillman et Tarnai 1988).

Au cours des années 1990, il est devenu évident qu'il fallait élaborer de nouvelles méthodes d'enquête. Les taux de réponse, surtout aux interviews sur place et par téléphone, avaient commencé à baisser (Brick et Williams 2013). Les problèmes de couverture se multipliaient également, à mesure que les immeubles à

logements multiples verrouillés et les ensembles résidentiels protégés empêchaient d'atteindre de nombreux ménages en personne. En outre s'est amorcé le long déclin inexorable de la couverture des ménages par les lignes téléphoniques fixes, si bien qu'aujourd'hui environ la moitié des ménages américains ne possèdent plus ce genre de connexion.

L'intérêt accordé à l'usage coordonné de plusieurs modes de collecte des données en vue d'améliorer les taux de réponse a attiré l'attention sur des problèmes d'interview dont on n'avait pas tenu compte antérieurement en raison des obstacles pratiques au mélange des modes de collecte. Par exemple, les répondants interviewés donnaient souvent des réponses socialement souhaitables de sorte que les estimations des comportements désirables, par exemple « avoir voté aux dernières élections », étaient supérieures à la réalité. En outre, les estimations des comportements indésirables, par exemple fumer de la marijuana ou avoir des relations sexuelles en dehors du mariage, étaient plus faibles (de Leeuw 1992). Des différences ont également été constatées entre les réponses aux questions d'enquête par la poste, d'une part, et aux enquêtes par téléphone et sur place, d'autre part, où les répondants donnaient davantage de réponses positives extrêmes aux questions d'opinion (de Leeuw 1992; Tarnai et Dillman 1992). Des travaux de recherche avaient également donné à penser que les répondants étaient plus susceptibles de choisir les premières catégories de réponse dans les enquêtes par la poste (effet de primauté), et les dernières catégories dans les enquêtes téléphoniques (effet de récence) (Krosnick et Alwin 1987). Par conséquent, les commanditaires d'enquête ont eu de plus en plus de difficulté à soutenir que les interviews par téléphone, voire même sur place, étaient des modes d'enquête supérieurs.

Les enquêtes à modes mixtes ont été proposées comme une solution possible, quoiqu'imparfaite, aux problèmes des modes d'enquête individuels. Cinq types d'enquêtes à modes mixtes ont été définis, allant de la collecte des mêmes données auprès de différents membres d'un échantillon à l'utilisation d'un mode uniquement pour inviter à répondre par un autre mode (Dillman 2000, page 219). Le principal avantage de la combinaison de divers modes semblait tenir aux améliorations des taux de couverture et de réponse que l'on pouvait obtenir. La principale difficulté reconnue était l'existence éventuelle de différences de mesure dues à l'utilisation de différents modes de réponse.

Un article crucial publié par de Leeuw (2005) a déclenché une évolution importante des idées au sujet des enquêtes à modes mixtes. L'auteure y énonçait une gamme de combinaisons possibles acceptées des modes d'enquête et donnait des preuves d'un recours croissant aux enquêtes à modes mixtes. Elle soulignait également le passage d'un débat sur le meilleur mode d'enquête pour une étude particulière à un débat sur la façon d'utiliser divers modes ensemble et de produire de meilleurs résultats.

Une évolution contextuelle avait également lieu, alors que les sociétés modernes partout dans le monde commençaient à passer d'activités requérant l'intervention de personnes (par exemple obtenir de l'argent auprès des caissiers dans les banques, s'adresser à un agent pour faire des réservations de voyage et acheter des biens dans les magasins et sur catalogue) à des activités autogérées (Dillman 2000). Mais les chercheurs n'avaient pas encore répondu à la question de savoir si les interviews par téléphone pourraient persister face à ces tendances d'autogestion.

3.2 Les obstacles institutionnels à l'usage conjoint de divers modes d'enquête

Combiner divers modes d'enquête et, en particulier, abandonner les moyens privilégiés de poser les questions différemment selon le mode d'enquête faisait l'objet de grandes hésitations, voire même d'une opposition pure et simple (Dillman 2000). Les nouveaux moyens de recueillir des données d'enquête qui ont vu le jour au cours du dernier tiers du XX^{e} siècle ont eu pour conséquence, entre autres, une assez grande spécialisation du personnel de collecte des données. Certaines organisations réalisaient des enquêtes selon un mode seulement. Il était fréquent pour certains employés préposés à la collecte des données et leurs organisations de n'effectuer que des enquêtes téléphoniques, et dans une moindre mesure, des enquêtes par la poste. Quelques grandes entreprises étaient dotées d'unités d'échantillonnage et de collecte des données sur place. Il existait une tendance à vouloir réaliser des enquêtes selon le mode qu'un groupe connaissait le mieux. Cette tendance a été exacerbée à la fin des années 1990 quand les organisations spécialisées dans les enquêtes par Internet uniquement ont commencé à voir le jour.

En outre, différents styles de libellé des questions en fonction du mode d'enquête utilisé sont apparus. Les intervieweurs avaient tendance à garder en suspens la catégorie « ne sait pas », ne l'offrant que si le répondant opposait une objection. Les concepteurs des questionnaires sur papier et par Internet utilisaient souvent des présentations de question de type « cocher toutes les réponses pertinentes » pour faciliter la réponse, mais la lourdeur de cette présentation au téléphone a mené à l'utilisation de présentations à choix forcé uniquement consistant à obtenir une réponse après que chaque item individuel était présenté. Le problème qui se posait aux concepteurs d'enquête était de savoir s'il fallait optimiser la présentation des questions en fonction du mode ou essayer de maintenir le même stimulus pour tous les modes (Dillman et Christian 2005).

L'un des facteurs étayant la tendance à s'en tenir à ce que les enquêteurs connaissaient le mieux était le constat que les enquêtes à mode unique représentaient la meilleure option dans de nombreuses situations, et ce pour chacun des modes. Les interviews sur place constituaient le seul moyen d'obtenir une couverture adéquate pour certaines enquêtes nationales, comme la *Current Population Survey*, qui produit des estimations du taux d'emploi. Les enquêtes téléphoniques par CA étaient le meilleur moyen de réaliser des sondages électoraux et d'autres enquêtes transversales auprès des ménages. L'envoi par la poste était le moyen le plus adéquat de procéder à des enquêtes régionales et locales pour lesquelles on ne disposait que d'adresses résidentielles. Les enquêtes à réponse vocale interactive étaient les plus pratiques pour de nombreuses enquêtes sur la satisfaction des clients quand des gens communiquaient avec des centres d'appel pour obtenir un service particulier. Et les enquêtes par Internet sont devenues la méthodologie de choix pour les enquêtes auprès de la clientèle et d'autres situations pour lesquelles des adresses de courriel avaient été recueillies précédemment.

Cette situation a marqué le début de l'intérêt pour la « conception sur mesure », c'est-à-dire la reconnaissance que certains modes de collecte des données conviennent mieux que d'autres pour des enquêtes sur des populations, des sujets et des situations de collecte particuliers. Cette tendance de la conception des enquêtes est aujourd'hui plus prononcée qu'elle ne l'était au tournant du siècle. Il est manifeste que choisir le seul mode de collecte des données jugé le meilleur pour une enquête particulière est une approche de plus en plus inadéquate, en raison des effets négatifs sur la couverture, les taux de réponse et l'erreur due à la non-réponse.

À la fin du XX^{e} siècle, l'incertitude était grande quant à la direction que prendraient les méthodes de collecte des données. Les chances de pouvoir continuer à s'appuyer uniquement sur les enquêtes sur place ou par téléphone étaient ténues. Les problèmes de couverture et les coûts augmentaient considérablement, et il semblait peu probable que les taux de réponse s'amélioreraient dans le cas des enquêtes vocales par téléphone. L'intérêt pour le remplacement de ces méthodes d'interview par Internet était grand, mais au tournant du siècle, la moitié seulement des ménages américains étaient dotés d'ordinateurs, et un nombre encore plus faible avaient accès à Internet (Dillman 2000).

4 L'élaboration et la mise à l'essai de la collecte de données à modes mixtes axée sur le Web

Au cours de la première décennie du XXI^{e} siècle, l'idée d'utiliser plusieurs modes d'enquête pour communiquer avec les particuliers et obtenir les réponses aux questionnaires semblait être une question qu'il était temps d'examiner en profondeur (par exemple Tourangeau 2017; de Leeuw, Villar, Suzer-Gurtekin et Hox 2017). En outre, les technologies de l'information qui avaient apporté Internet offraient la possibilité de gérer de manière efficace et efficiente l'usage simultané de plusieurs modes de collecte des données, ce qui éliminait le principal obstacle pratique à la réalisation d'enquêtes à modes mixtes.

La bonne élaboration de méthodes axées sur le Web signifiait qu'il fallait chercher à résoudre plusieurs questions en même temps afin de savoir si une telle approche serait efficace. Ces questions allaient de la réponse aux problèmes de couverture des ménages et de la compréhension de la façon dont la communication visuelle différait de la communication auditive, à l'élargissement de notre réflexion théorique au sujet des éléments qui influencent les gens à répondre aux demandes d'enquête.

Une grande question sans réponse était celle de savoir si l'autoadministration des questionnaires pouvait remplacer l'intervention de l'intervieweur, et si les résultats seraient meilleurs ou pires. Une difficulté confondante était que d'importantes différences existaient entre les modes d'enquête en ce qui concerne les taux de couverture et de réponse, ainsi que le biais lié à la façon dont les personnes réagissent à leur utilisation, chaque mode étant peut-être meilleur dans certaines situations et pire dans d'autres.

4.1 Les listes d'adresses résidentielles des services postaux américains fournissent aujourd'hui une excellente couverture des ménages

Puisque le service postal des États-Unis fournit, par l'entremise de vendeurs, des listes d'adresses résidentielles complètes, il est possible d'envoyer des demandes par la poste à presque toutes les résidences aux États-Unis (Harter, Battaglia, Buskirk, Dillman, English, Mansour, Frankel, Kennel, McMichael, McPhee, Montaquila, Yancey et Zukerberg 2016). Ces listes d'adresses résidentielles informatisées sont fournies sans les noms, de la même façon que les listes téléphoniques pour la CA ne contiennent pas de noms. L'absence de noms n'est pas un obstacle à l'obtention d'une réponse auprès des ménages, comme l'a montré une série d'études sur les taux de réponse en se servant des listes d'adresses du recensement décennal des États-Unis (Dillman 2000, chapitre 9). Qui plus est, l'envoi par la poste n'est pas adressé à une seule personne dans les ménages, dont les membres sont moins liés les uns aux autres qu'à l'époque où les taux

de mariage étaient plus élevés. L'utilisation des listes peut aussi permettre une sélection plus exacte des répondants, puisqu'elle n'oblige pas à contourner les limites des envois par la poste associés à un seul membre du ménage.

L'une des premières études à grande échelle en vue d'évaluer l'utilisation d'un échantillon fondé sur les adresses (EFA) avec collecte des données par la poste était celle de Link, Battaglia, Frankel, Osborn et Mokdad (2008). Ces auteurs ont constaté, pour un questionnaire du *Behavior Risk Factor Surveillance System* (BRFSS) de 2005, qu'un questionnaire envoyé par la poste à un échantillon EFA produisait des taux de réponse significativement plus élevés que ceux obtenus pour l'échantillonnage par CA dans cinq des six états visés par l'enquête. Les auteurs ont conclu, avec une mise en garde appropriée, que le potentiel réel de l'échantillon EFA pourrait tenir à la facilitation des enquêtes à modes mixtes qui comprennent également un suivi par téléphone, et ont vivement recommandé de poursuivre l'étude.

D'autres travaux de recherche menés à ce moment-là ont montré que les échantillons EFA offraient un taux de couverture très élevé qui s'améliorait à mesure que les adresses de type urbain remplaçaient les adresses moins précises, telles que les routes rurales (O'Muircheartaigh, English et Eckman 2007; Battaglia, Link, Frankel, Osborn et Mokdad 2008). En outre, une série d'études ont montré qu'une enquête par la poste en deux étapes avec échantillon EFA (présélection des ménages concernant la présence d'enfants d'âge scolaire, suivie par un questionnaire détaillé sur un enfant particulier) produit de meilleurs résultats qu'une approche par CA en deux étapes, les taux de réponse étant nettement plus élevés (Brick, Williams et Montaquila 2011; Williams, Brick, Montaquila et Han 2014).

Ces études ont joué un rôle important dans l'établissement des attributs de grande couverture de l'échantillonnage fondé sur les adresses en tant qu'option de remplacement de l'échantillonnage par CA. Cependant, elles ne vérifiaient pas si les ménages contactés pouvaient être persuadés de répondre par Internet à des demandes envoyées par la poste.

4.2 La détection et la correction des différences de mesure entre les enquêtes à modes visuel et auditif

Une autre préoccupation, assez différente, limitant l'intérêt pour l'échantillonnage fondé sur les adresses avec utilisation d'un questionnaire imprimé et (ou) d'un questionnaire sur Internet était que les réponses aux questions seraient vraisemblablement différentes de celles fournies par téléphone. Cette préoccupation était double. Premièrement, sans intervieweur, les répondants ne pouvaient pas recevoir d'encouragement supplémentaire quand ils n'étaient pas capables de répondre à une question ou hésitaient à le faire, ni des éclaircissements s'ils comprenaient mal les questions. Deuxièmement, il existait de longue date des preuves que la désirabilité sociale et la tendance à être d'accord (acquiescement) étaient plus importantes pour les réponses aux questionnaires téléphoniques qu'aux questionnaires à remplir soi-même (envoyés par la poste) (de Leeuw 1992). En règle générale, l'avantage qu'offrait la présence d'un intervieweur était considéré comme excédant l'éventuel biais dû à ce dernier.

Un sondage commandité par la *Gallup Organization* en 1999 a fait voir ces différences sous un nouvel angle. L'essai a révélé que, dans une interview, la réponse à des stimuli reçus auditivement, par téléphone ou par réponse vocale interactive, produisait des réponses plus positives que celles données à des stimuli

transmis visuellement, au moyen d'un questionnaire envoyé par la poste ou par Internet (Dillman 2002; Dillman, Phelps, Tortora, Swift, Kohrell, Berck et Messer 2009).

Des découvertes sur la façon dont l'information visuelle est traitée publiées par Palmer (1999), Hoffman (2004) et Ware (2004), ont fourni des notions théoriques sur les actions distinctes qui ont lieu lorsque les yeux captent l'information et que le cerveau la traite pour comprendre ce qui se trouve sur la page ou sur l'écran. L'application de ces concepts a permis de comprendre les raisons pour lesquelles les questionnaires à remplir soi-même produisaient souvent des réponses différentes des questionnaires avec intervieweur, comme l'avait révélé l'étude Gallup. Les répondants sont guidés à travers les questionnaires visuels par de multiples langages qui communiquent une signification. Ces langages comprennent des symboles, des nombres et la composition graphique (taille, espacement, couleur, symétrie, régularité, etc.) qui ont une incidence sur la façon dont l'information sur les pages imprimées ou les pages Internet est parcourue, groupée mentalement et interprétée (Dillman 2007, pages 462 à 497; Tourangeau, Couper et Conrad 2004). Des travaux de recherche supplémentaires ont montré que l'on pouvait augmenter considérablement l'observation des instructions d'enchaînement en modifiant les symboles, la taille de la police, la brillance de la police (Redline et Dillman 2002; Christian et Dillman 2004), ainsi que le placement de ces instructions d'enchaînement par rapport aux choix de réponses (Redline, Dillman, Dajani et Scaggs 2003; Dillman, Gertseva et Mahon-Haft 2005).

Une autre cause importante de différences de mesure entre les modes est devenue évidente : les questions étaient souvent libellées différemment pour chaque mode et présentées selon des structures différentes (Dillman et Christian 2005). Ainsi, dans les sondages d'opinion concernant une liste d'items, les chercheurs avaient depuis longtemps l'habitude de poser des questions à choix forcé individuelles dans les enquêtes téléphoniques, mais ils les convertissaient souvent en une présentation de type « cocher toutes les réponses pertinentes » pour les items regroupés sur les questionnaires envoyés par la poste (Smyth, Dillman, Christian et Stern 2006). Cette pratique avait été transposée aux enquêtes par Internet. De nouvelles études ont montré qu'une présentation à choix forcé pour les modes visuel ainsi qu'auditif rapprochait considérablement les réponses des répondants (Smyth, Dillman, Christian et McBride 2009). La recherche a également montré que les réponses aux questions ouvertes des enquêtes par la poste et par Internet étaient comparables si une construction visuelle similaire était utilisée pour les deux modes d'enquête (Smyth, Christian et Dillman 2008). En outre, on a appris que des variations dans la présentation des questions sur échelle (par exemple étiquetage complet *vs* étiquetage des points extrêmes) produisaient des différences de réponse très importantes entre les divers modes visuels (Christian, Parsons et Dillman 2009).

En vue d'éliminer les différences de mesure entre ces modes, on a proposé une construction selon un mode unifié, c'est-à-dire l'utilisation du même énoncé et de la même présentation visuelle des questions de l'enquête (Dillman 2000). Une construction unifiée pouvait être réalisée facilement pour de nombreux types de questions (par exemple présenter les catégories « ne sait pas » à tous les répondants plutôt qu'uniquement à ceux qui ne choisissent pas l'une des réponses offertes), comme cela se faisait généralement dans les interviews par téléphone. Cependant, dans d'autres situations, une construction variable selon le mode était pratique et réduisait également les erreurs, par exemple l'enchaînement automatique à la question appropriée suivante sur Internet et par téléphone. Cette forme de présentation ne peut être réalisée pour l'enchaînement des items sur les questionnaires papier où toutes les options doivent être imprimées, parce qu'il est impossible de savoir d'avance comment une personne répondra à ces items.

Le principal apport de la construction à mode unifié a été de réduire l'inquiétude que de multiples modes de réponse à une enquête favoriseraient des différences de mesure. Il existe toutefois des preuves convaincantes que la réponse par téléphone à des échelles d'opinion utilisant des quantificateurs vagues est systématiquement plus susceptible de produire des réponses à l'extrémité positive de l'échelle et le choix moins fréquent des catégories intermédiaires que dans le cas des questionnaires sur Internet ou envoyés par la poste (Christian, Dillman et Smyth 2008). Cette différence semble être due au fait que la présentation visuelle des catégories de réponse intermédiaires est plus visible et, par conséquent, plus accessible aux répondants que quand ces mêmes catégories sont lues au téléphone, processus qui rend les catégories finales plus dominantes dans l'esprit des répondants, Dillman et Edwards (2016).

La façon dont les personnes fournissent des réponses socialement désirables à certaines questions est une autre différence que ne résout pas la construction à mode unifié. Toutefois, les questionnaires à remplir soi-même (visuels) sont généralement considérés comme produisant des réponses plus honnêtes.

L'accumulation d'études sur les problèmes liés à la conception de questionnaires visuels *vs* auditifs a fourni aux concepteurs d'enquête des outils essentiels, qui permettent d'éliminer partiellement les différences de mesure susceptibles de réduire les avantages de couverture et de réponse des enquêtes à modes mixtes. La pratique d'une conception de questionnaire à mode unifié a été un élément fondamental de l'élaboration initiale et de la mise à l'essai de la méthodologie axée sur le Web décrite plus bas.

4.3 L'élaboration séquentielle d'une méthodologie axée sur le Web efficace

Une série de dix essais de procédures de collecte des données axées sur le Web a été réalisée par une équipe de chercheurs de l'Université de l'État de Washington entre 2007 et 2012 à l'occasion de cinq collectes de données distinctes. Le plan qui sous-tendait ces expériences était de s'appuyer sur les enseignements tirés des premiers essais pour concevoir et mettre en œuvre les essais ultérieurs. Toutes les comparaisons expérimentales ont été faites en se servant de l'équivalent d'un questionnaire papier de 12 pages, contenant de 50 à 70 questions numérotées, qui demandaient de 90 à 140 réponses possibles. Ces questionnaires étaient conçus afin d'équivaloir à un questionnaire d'interview de 20 à 30 minutes. Les études portaient sur divers sujets, à savoir la participation et la satisfaction communautaires, l'usage des technologies de l'information, les effets économiques et sociaux de la récession de 2008, les attitudes de consommation d'énergie, et la compréhension de la qualité et de la gestion de l'eau. Les chercheurs ont fait varier les sujets afin de réduire les préoccupations quant à l'effet du sujet sur les taux de réponse et la qualité des données.

Les populations étudiées variaient d'une région rurale dans les états de l'Idaho et de Washington, ainsi que d'enquêtes à l'échelle des états de Washington, de Pennsylvanie et d'Alabama réalisées par l'Université de l'État de Washington, à des enquêtes auprès des résidents du Nebraska et de l'état de Washington envoyées par l'Université du Nebraska et les mêmes enquêtes envoyées aux deux états par l'Université de l'État de Washington. Les procédures de mise en œuvre variaient, mais comprenaient de 4 à 5 prises de contact par la poste, avec l'option d'un questionnaire à réponse par la poste offerte à la 3^{e} ou 4^{e} prise de contact. Un petit incitatif monétaire symbolique a été envoyé avec la première demande de réponse, et dans certains cas, un petit incitatif a été envoyé avec le questionnaire papier quand celui-ci était retenu jusqu'à la 3^{e} ou à la 4^{e} prise de contact. Les procédures détaillées pour chacune des études sont décrites ailleurs (Smyth,

Dillman, Christian et O'Neill 2010; Messer et Dillman 2011; Messer 2012; Edwards, Dillman et Smyth 2014; Dillman, Smyth et Christian 2014).

Lors du premier essai dans une région rurale de l'Idaho et de l'état de Washington, 55 % des ménages ont répondu au traitement axé sur le Web, avec 74 % de ces réponses fournies par Internet. Cet essai a également révélé que le taux de réponse était significativement plus élevé (63 %) si l'on incluait un questionnaire imprimé et que l'on offrait directement un choix de modes (Smyth et coll. 2010). Malheureusement, près de 80 % des réponses étaient alors fournies sur papier, proportion trop élevée pour justifier le coût de la mise en place d'une collecte de données par Internet. Étant donné cet effet et la promesse initiale d'obtenir plus de la moitié des réponses par Internet dans le cas du traitement axé sur le Web, un terme a été mis à l'expérience sur la méthodologie avec choix. Ce premier essai nous a également permis de constater que le traitement axé sur le questionnaire papier avec l'offre d'une option par Internet reportée au dernier contact ne produisait que 2 % de réponses sur Internet. Compte tenu de ce résultat, le suivi par Internet a été interrompu après que deux essais additionnels aient donné des résultats similaires. En outre, les résultats de cette première étude en région rurale nous encourage à retenir la méthodologie axée sur le Web avec suivi papier en vue d'un essai supplémentaire auprès de populations à l'échelle de l'état.

Sur l'ensemble des dix expériences couvrant cinq états, les enquêtes axées sur le Web ont produit un taux de réponse moyen de 43 %, variant de 31 % à 55 % (figure 4.1). Les comparaisons avec le traitement par envoi postal seulement ont produit un taux de réponse moyen de 53 %, avec une fourchette de 38 % à 71 %. En moyenne, 60 % des réponses aux traitements axés sur le Web ont été reçues par Internet. Dans l'une des études, les traitements expérimentaux ont montré que l'incitatif inclus dans la demande axée sur le Web accroissait considérablement la réponse par Internet, pour passer de 13 % à 31 %, soit environ 18 points de pourcentage (Messer et Dillman 2011). Bien qu'une comparaison avec la méthode de CA n'était incluse dans aucune des expériences, les résultats des procédures axées sur le Web étaient sans aucun doute beaucoup plus élevés que ceux que l'on aurait obtenus par téléphone pour ces longs questionnaires, si une comparaison avait été effectuée.

Une comparaison de la non-réponse partielle a été faite dans le cas de trois des expériences pour déterminer si les taux de non-réponse partielle étaient plus élevés pour les questionnaires de suivi envoyés par la poste que pour les réponses par Internet obtenues pour ces groupes de traitement. Dans le cas de l'étude régionale de 2007 et de deux études à l'échelle de l'état de 2009, les questionnaires papier de suivi produisaient des taux de non-réponse partielle plus de deux fois plus élevés, 8,2 % *vs* 3,6 %, que les questionnaires remplis sur Internet. Cependant, la comparaison de la non-réponse partielle globale pour les groupes de traitement axé sur le Web (réponses par Internet plus par la poste) et pour le groupe du traitement axé sur la réponse par la poste seulement n'a révélé pratiquement aucune différence entre les groupes, les taux étant de 5,3 % et 5,7 %, respectivement. Les auteurs ont supposé que les premières réponses par Internet étaient fournies par de « meilleurs » répondants, tandis que les réponses ultérieures par la poste provenaient de répondants ayant moins de capacités, plus âgés et ayant fait moins d'études (Messer, Edwards et Dillman 2012).

Les taux de réponse étaient significativement plus faibles pour les populations ne connaissant vraisemblablement pas l'Université de l'État de Washington – le commanditaire de ces études –, en

particulier celles répondant par Internet. Par exemple, les taux de réponse n'étaient que de 12 % et de 11 % en Pennsylvanie et en Alabama, respectivement, comparativement à 28 % dans l'état de Washington (Messer 2012). Une étude sur la gestion de l'eau réalisée par l'Université du Nebraska et par l'Université de l'État de Washington a permis de mieux comprendre ce phénomène grâce à l'envoi de demandes de réponse à des ménages dans l'autre état. La réponse pour le traitement axé sur le Web était inférieure de 6,1 points de pourcentage chez les résidents de l'état de Washington et de 14,7 points de pourcentage chez les résidents du Nebraska lorsque l'enquête était réalisée par l'université située en dehors de l'état (Edwards et coll. 2014). Cette baisse touchait presque entièrement les réponses par Internet, qui sont passées de 32 % à 26 % dans l'état de Washington et de 38 % à 23 % au Nebraska, lorsque les demandes de réponse provenaient de l'université de l'état opposé. Nous avons supposé que la réponse par Internet est plus sensible que la réponse par la poste au manque de familiarité avec le commanditaire de l'enquête et de confiance en celui-ci.

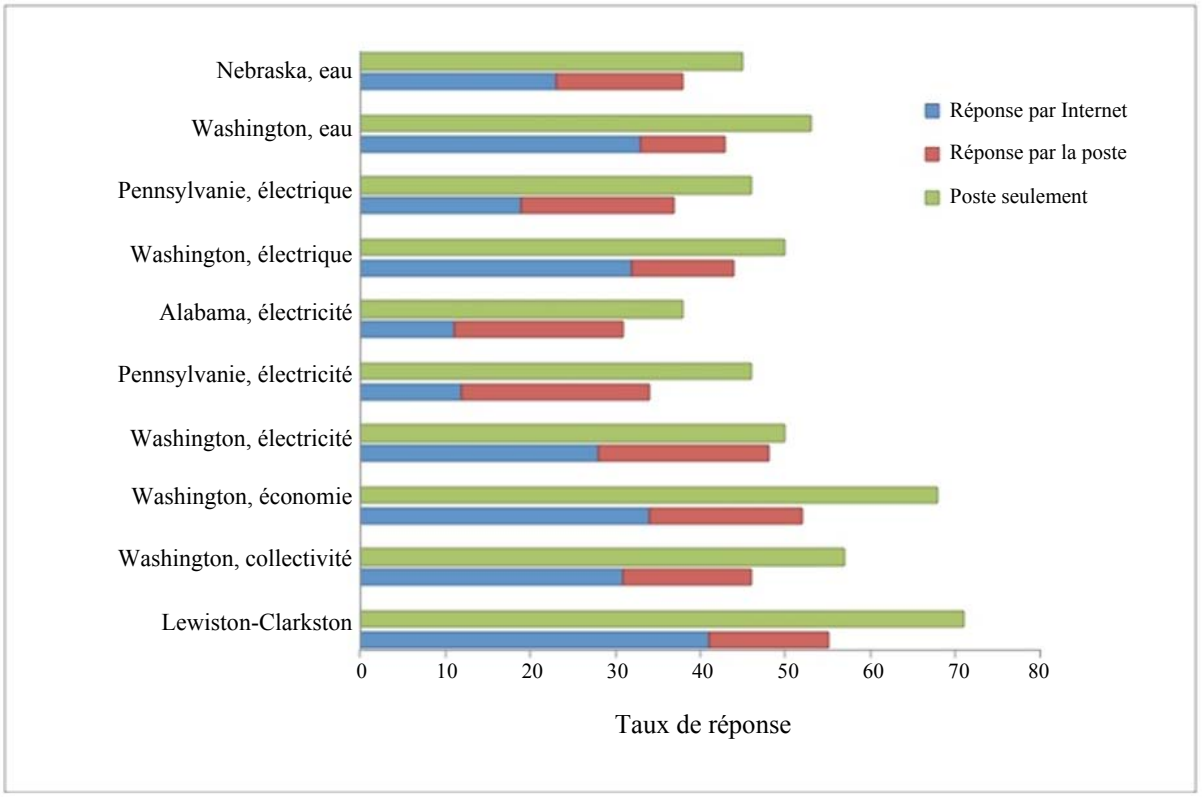

Figure 4.1 Taux de réponse pour le traitement par la poste seulement *vs* taux de réponse pour le traitement axé sur le Web avec proportions des réponses reçues par chaque mode (Dillman et coll. 2014, chapitre 11).

Les dix comparaisons de ces études de la poussée vers le Web ont révélé que les personnes qui répondaient par Internet dans les groupes de traitement axé sur le Web différaient considérablement de celles qui répondaient plus tard au questionnaire envoyé par la poste. Par exemple, les personnes qui répondaient par Internet étaient plus jeunes, avaient fait plus d'études, avaient un revenu plus élevé et étaient moins susceptibles de vivre seules (Messer et Dillman 2011). Toutefois, regroupés, les répondants par Internet et

par la poste issus des groupes de traitement axé sur le Web étaient démographiquement assez semblables aux répondants des groupes de traitement axé sur la réponse par la poste seulement. L'étude a abouti à la conclusion que les personnes enclines à répondre par Internet pourraient aussi être convaincues de répondre dans le cas du traitement par la poste seulement. Cette constatation a été renforcée par le fait qu'un suivi avec questionnaire papier à la demande de réponse par Internet seulement améliorait significativement les taux de réponse, tandis qu'un suivi par Internet à une approche de traitement par la poste seulement ne produisait pas de réponses par Internet qui augmentaient significativement la réponse globale.

Même si les réponses des groupes de traitement axé sur le Web et d'envoi par la poste seulement étaient assez semblables, les données non pondérées présentaient un biais de non-réponse pour certaines catégories démographiques. Les répondants étaient plus instruits et comptaient plus d'enfants au foyer que ceux qui répondaient à l'*American Community Survey* (discutée plus en détail plus bas) qui fait appel aujourd'hui à des interviews par la poste, par Internet, par téléphone et sur place pour obtenir un taux de réponse de près de 97 % et sur laquelle on s'appuie pour produire les statistiques officielles pour tous les états américains. Ce genre de comparaison dépassait le cadre et la portée des expériences susmentionnées, et des études plus poussées devront être réalisées pour comprendre la nature de ces différences. De surcroît, le coût par répondant ne s'est pas avéré plus faible pour les réponses par Internet, parce que les coûts de prise de contact étaient à peu près les mêmes pour les méthodes axées sur le Web et d'envoi par la poste seulement, tout en produisant moins de répondants (Messer et Dillman 2011). Cette situation évoluera vraisemblablement à mesure que l'utilisation d'Internet continuera de s'étendre à un plus grand nombre de personnes et de domaines de la vie.

Globalement, cet ensemble coordonné d'études a montré clairement que la méthodologie axée sur le Web en vue d'obtenir des réponses par Internet aux enquêtes auprès des ménages est très prometteuse. Il est également évident que les questionnaires de suivi papier amélioreraient la représentation des personnes incapables ou non désireuses de répondre par Internet.

4.4 Essais supplémentaires d'enquêtes axées sur le Web pour d'autres populations et situations

Ces dernières années, les stratégies de collecte de données axées sur le Web ont pris de l'ampleur et elles sont appliquées à grande échelle aux enquêtes menées par les administrations publiques, les universités et le secteur privé dans de nombreux pays. Leur usage s'est également étendu au-delà des populations générales et englobe aujourd'hui des situations d'enquête où les demandes de réponse par Internet ne sont pas limitées à la prise de contact par la poste. En outre, certaines enquêtes comportent jusqu'à trois modes de prise de contact et trois modes de réponse, l'intention étant d'obtenir des taux de réponse très élevés, en poussant autant de répondants que possible à opter pour Internet, en vue de réduire les coûts d'enquête.

En 2013, l'*American Community Survey* a été convertie d'une séquence de demande de réponse par la poste-par téléphone-sur place à un démarrage avec réponse par Internet, suivi par les trois procédures de prise de contact et de réponse restantes (United States Census Bureau 2014, chapitre 7). La loi exige que les citoyens américains répondent à l'*American Community Survey* (anciennement le questionnaire détaillé du recensement décennal). Par conséquent, le taux de réponse global pour les logements occupés était d'environ 97 %. Des essais de stratégies de poussée vers le Web ont débuté en 2011, quand une première expérience

a confirmé que la stratégie axée sur le Web produisait des taux de réponse par Internet considérablement plus élevés (28 % *vs* 10 %) qu'une stratégie de « choix » offrant également la réponse par la poste à la première prise de contact (Tancreto 2012). En 2013, 28 % des réponses provenant des logements occupés ont été faites par Internet, 22 %, par la poste, 6 %, par téléphone, et 43 %, par interview sur place. Donc, environ 51 % des réponses auto-administrées ont été fournies par Internet, proportion qui est passée à 58 % en 2015. Des essais sont maintenant en cours en vue d'appuyer les plans de convertir le Recensement décennal de 2020 à des méthodes axées sur le Web avec un suivi similaire.

Le Recensement du Japon a été converti à une méthodologie axée sur le Web en 2015 (City of Sapporo 2015). La réponse en ligne était d'environ 37 %, dont un tiers provenant de téléphones intelligents, qui sont très répandus au Japon. Les autres réponses ont été obtenues au moyen de questionnaires envoyés par la poste et de visites d'agents recenseurs. Le Recensement de l'Australie de 2016 et le Recensement du Canada de 2016 ont également été réalisés en appliquant une méthodologie axée sur le Web. Bien que les résultats définitifs ne soient pas encore disponibles, on sait qu'au Canada, 68 % des ménages ont répondu par Internet, 20 % par la poste, et 10 % supplémentaires durant la visite d'un agent recenseur, ce qui donne un taux de réponse global de 98 % (Statistique Canada 2016). La proportion de réponses par Internet au Recensement du Canada est la plus élevée que je connaisse pour une enquête auprès des ménages axée sur le Web. Dans certaines régions du Canada, un questionnaire papier a été joint à la demande, afin d'offrir un choix de mode de réponse aux répondants. Le taux élevé de réponse par Internet (68 %) et de réponse par Internet plus par la poste (88 %) est prometteur pour l'usage d'une méthodologie axée sur Internet dans ce pays et peut-être d'autres où la pénétration d'Internet est forte.

La *National Child Health Survey*, récemment élaborée pour remplacer une enquête auprès des ménages par CA aux États-Unis, prévoit la présélection d'un échantillon d'enfants fondée sur des adresses, puis la sélection d'un enfant pour l'obtention de renseignements détaillés sur les questions de santé. Cependant, au lieu d'utiliser deux collectes de données par la poste distinctes, l'équipe du programme a testé en 2015 la possibilité de réduire le processus à une seule étape, dans laquelle l'ordinateur utilise des critères pour directement sélectionner un enfant et administrer un questionnaire thématique sur la santé. Cette procédure a pour objectif d'améliorer le processus de réponse par la poste en deux étapes élaboré pour la *National Child Education Survey*. Les résultats d'un pré-essai réalisé en 2015 étaient prometteurs et la procédure fait maintenant l'objet d'une deuxième phase d'essai.

La *Residential Energy Consumption Survey* des États-Unis, réalisée pendant de nombreuses années par l'*Energy Information Administration* par interview sur place auprès des ménages, est en train d'être convertie en une enquête axée sur le Web. Cette enquête est intéressante, parce qu'elle combine un incitatif monétaire avec la demande initiale de réponse par Internet, et offre aussi un incitatif consécutif à la réponse. L'incitatif consécutif à la réponse a été jugé particulièrement important en raison des économies qu'il permettait de réaliser en ne devant pas envoyer des intervieweurs sur place auprès des ménages non répondants (Biemer, Murphy, Zimmer, Berry, Deng et Lewis 2015).

Les enquêtes axées sur le Web ne font pas toutes appel à l'échantillonnage fondé sur les adresses. Le programme de la *National Survey of College Graduates* (NSCG) de 2010 a commencé à échantillonner les personnes qui avaient déclaré être diplômées d'un collège durant l'*American Community Survey* de l'année précédente et leur ont demandé de remplir le questionnaire de la NSCG, qui est réalisée tous les deux ans

(Finamore et Dillman 2013). Les adresses postales, ainsi que les numéros de téléphone, étaient disponibles principalement pour les ménages dans lesquels les diplômés avaient vécu l'année précédente. Avant 2010, les ménages étaient sélectionnés d'après le questionnaire détaillé du recensement décennal (rempli pour la dernière fois en 2000) et l'enquête était réalisée par téléphone, par la poste et, dans certains cas, par interview sur place. En 2010, des comparaisons ont été réalisées entre trois méthodes, à savoir pousser les personnes à répondre par téléphone, les pousser à répondre par la poste et les pousser à répondre par Internet, suivi par l'utilisation des deux autres modes. Chacun de ces trois traitements a été suivi d'un dernier rappel téléphonique donnant la possibilité de répondre par ce mode ou par l'un des deux autres. Deux résultats se sont avérés particulièrement importants. Premièrement, les trois taux de réponse ne différaient l'un de l'autre que de quelques points de pourcentage, variant de 74 % à 77 %, pour cette enquête à participation volontaire. Cependant, la stratégie de poussée vers le Web, pour laquelle 53 % de personnes ont répondu par Internet, s'est révélée beaucoup moins coûteuse que les autres, 48 $ par répondant *vs* 66 $ pour la réponse par la poste d'abord et 75 $ pour la réponse par téléphone d'abord. La conclusion a été que les résultats de chaque procédure représentaient assez bien l'échantillon original.

Une enquête à participation volontaire récente menée auprès des conjointes et conjoints des militaires américains a servi à comparer une stratégie axée sur le Web avec une stratégie axée sur la réponse par la poste. La méthodologie axée sur le Web a produit un taux de réponse significativement plus élevé, 33 % *vs* 28 %, avec 87 % des réponses à la méthode axée sur le Web reçues par Internet (McMaster, LeardMann, Speigle et Dillman 2016). La stratégie axée sur le Web a également été nettement moins coûteuse, soit 61 $ par répondant *vs* 89 $.

Le succès des stratégies axées sur le Web pour les études des diplômés des collèges et des militaires peut avoir des explications différentes. Tous les participants à la NSCG possédaient au moins un diplôme d'études collégiales de quatre ans. Les participants à la *Family Study of Military Members* étaient aussi relativement jeunes. Selon les auteurs de la dernière étude, le fait que les militaires utilisent beaucoup Internet pour communiquer avec leur conjointe ou conjoint durant le déploiement pourrait expliquer la plus grande efficacité de la méthode axée sur le Web que des méthodes axées sur l'envoi par la poste.

De nombreux autres essais d'une méthodologie axée sur le Web ont eu lieu au cours de la dernière décennie. Une étude suisse a donné des taux de réponse d'environ 72 % des ménages sélectionnés à partir de listes suisses d'enregistrement, dont 44 % par Internet, 20 % par la poste et le reste par interview téléphonique ou sur place (Roberts et coll. 2016). Au Royaume-Uni, où les enquêtes statistiques nationales étaient réalisées beaucoup plus fréquemment par interview sur place que par interview téléphonique, il a été décidé récemment de convertir la *Community Life Survey* pour passer d'une stratégie d'interview sur place à une stratégie axée sur le Web suivie d'une réponse par la poste (United Kingdom Cabinet Office 2016). Cette décision a été prise afin de réduire les coûts, tout en augmentant la taille d'échantillon. Il reste à voir quels seront les résultats.

L'utilisation de méthodes axées sur le Web par le secteur privé pour étudier des populations particulières a également évolué. Nexant réalise maintenant des enquêtes auprès des clients des services publics de distribution de gaz et d'électricité par des méthodes axées sur le Web. Par le passé, les enquêtes téléphoniques étaient la méthode privilégiée. Les entreprises auprès des clients desquelles il faut mener les sondages peuvent fournir les adresses postales et les numéros de téléphone pour presque tous les clients et

les adresses de courriel pour 20 % à 40 % des ménages (Sullivan, Leong, Churchwell et Dillman 2015). Selon une procédure élaborée par Millar et Dillman (2011), un courriel est envoyé à ces ménages de manière à ce qu'il arrive peu de temps après la lettre de demande de participation qui contient un incitatif de 2 $, et est suivi d'un autre courriel trois jours plus tard, et de l'envoi d'un questionnaire papier en cas de non-réponse. Plusieurs essais ont produit des taux de réponse allant de 40 % à 80 %, avec utilisation d'Internet par 80 % à 90 % des répondants ayant reçu ces courriels supplémentaires après la prise de contact par la poste, comparativement à environ 35 % à 70 % de ceux pour lesquels on n'avait pas d'adresse de courriel. Les taux de réponse peuvent être haussés de 8 à 10 points de pourcentage grâce à un appel téléphonique de suivi aux personnes sans adresse de courriel, comparativement à 1 % ou 2 % pour celles possédant une telle adresse.

5 L'avenir est prometteur, mais des problèmes difficiles persistent

5.1 Des raisons d'être optimiste

Les travaux d'élaboration et de déploiement de méthodologies axées sur le Web pour la collecte de données d'enquête qui ont eu lieu au cours de la dernière décennie donnent des raisons d'être optimiste quant à la possibilité de recueillir des données d'enquête de plus grande qualité. Cet optimisme découle moins de l'excitation à l'idée d'appliquer une approche particulière pour prendre contact avec les personnes et les convaincre de répondre par Internet que d'une combinaison de facteurs.

L'échantillonnage fondé sur les adresses produit maintenant une excellente couverture des ménages et est propice à l'application de procédures de sélection des répondants. De grandes proportions des populations à étudier peuvent être contactées par un mode particulier (envoi par la poste) et invitées à répondre par un autre (Internet ou téléphone). La légitimité des commanditaires de l'enquête, qui sont inconnus de la personne qui reçoit la demande, peut être établie dans un envoi par la poste d'une manière qui ne peut pas être réalisée dans le cadre de demandes par courriels dont la plupart ne sont pas lus ou de demandes vocales par téléphone dont la plupart restent sans réponse.

La prise de contact par la poste permet aussi d'envoyer de petits incitatifs symboliques avec la demande, motivant ainsi la personne à passer de la lettre à l'ordinateur et d'y entrer une adresse URL (pour *Uniform Resource Locator*) et un mot de passe. Les prises de contact multiples par la poste donnent l'occasion d'expliquer plus complètement pourquoi une enquête est réalisée et comment les résultats seront utilisés. L'envoi d'un questionnaire papier comme autre option de réponse à l'occasion d'un contact ultérieur non seulement augmente les taux de réponses considérablement, mais amène aussi des types de ménages qui ne sont pas bien représentés parmi les réponses par Internet initiales. Plusieurs études ont également montré que la capacité des enquêtes axées sur le Web à convaincre de la moitié aux trois quarts des répondants à répondre rapidement par Internet, selon la base de sondage et les modes de prise de contact, peut réduire les coûts d'enquête.

Quand des adresses de courriel sont disponibles pour les unités échantillonnées, comme cela est maintenant le cas pour certaines populations étudiées, le renforcement par courriel (c'est-à-dire l'envoi d'un suivi rapide par courriel après la demande postale initiale afin de fournir un lien électronique qui permet au

destinataire de répondre plus facilement par Internet) s'est avéré améliorer considérablement la réponse par Internet. De même, quand on dispose de numéros de téléphone, un renforcement par téléphone peut être un moyen efficace d'améliorer la réponse. Le concept de l'utilisation de ces moyens de communication pour renforcer les prises de contact par la poste incite les enquêteurs à ne pas réfléchir simplement à des prises de contact indépendantes, mais à la façon dont chaque prise de contact devient partie intégrante d'une stratégie globale de réponse.

Comme l'ont montré les études sur l'*American Community Survey*, le Recensement du Canada, la *National Science Foundation* et Nexant, utiliser de multiples modes de prise de contact et de réponse donne la possibilité d'obtenir des taux de réponse que de nombreux commanditaires d'enquête croyaient être devenus impossible. La capacité de prendre contact avec les gens pour formuler des demandes répétées de répondre – et de le faire par différents modes – améliore la réponse aux enquêtes davantage que tout mode unique de prise de contact et (ou) de réponse.

En outre, s'appuyer plus sur l'autoadministration du questionnaire (par Internet et par la poste) représente une meilleure adaptation culturelle que le choix d'une conversation téléphonique vocale, de moins en moins en harmonie avec le comportement de communication habituel fortement axé sur les messages texte et les courriels. En outre, l'évolution des méthodes de construction des questionnaires pour passer de l'utilisation de différents énoncés et présentations des questions pour chaque mode en vue de créer le meilleur questionnaire possible pour chacun à une construction unifiée pour les divers modes aide à éviter les différences de mesure entre les divers modes de collecte de données d'enquête.

Au fil du temps, il paraît probable qu'une proportion croissante d'adultes seront désireux et capables de répondre aux enquêtes par Internet. Donc, les procédures de collecte de données axées sur le Web semblent concorder avec d'autres tendances sociétales privilégiant l'Internet plutôt que d'autres formes de communication.

La promesse des méthodes d'enquête axées sur le Web émane de leur capacité à réduire l'erreur d'enquête due à la couverture et à la non-réponse. En outre, notre meilleure compréhension de la façon dont la présentation visuelle par opposition à auditive des questionnaires influe sur les réponses et le recours à des méthodes de construction unifiée pour les divers modes permettent de réduire les différences de mesure et l'erreur. Il est probable que le nombre d'enquêtes faisant appel à des méthodes axées sur Internet augmentera.

6 Les défis de la collecte de données axée sur le Web

Malgré les possibilités qu'offrent les méthodes de collecte de données axées sur le Web, certaines incertitudes persistent quant à l'expansion continue des enquêtes par Internet. Ces préoccupations sont le sujet de la dernière partie du présent article.

6.1 La crainte de répondre sur Internet

Quand le recensement de l'Australie axé sur le Web a débuté en 2016, une série d'attaques par déni de service (DOS) sur le site ont incité le *Bureau of Statistics* à fermer le système par crainte des pirates. Les

attaques de ce genre sont conçues pour surcharger le trafic sur un serveur, afin de le rendre inaccessible aux utilisateurs prévus. Il ne s'agit que d'un des types d'attaque qui pourraient être perpétrés contre une enquête ou un utilisateur d'ordinateur particulier. D'autres comprennent l'envoi d'un maliciel (par exemple logiciel d'espionnage ou de rançon) conçu pour avoir accès à un ordinateur ou l'endommager quand l'utilisateur accède sans le savoir à ce maliciel en ouvrant des pièces jointes ou en cliquant sur des liens. Des courriels de hameçonnage peuvent aussi être envoyés. Ils sont conçus pour tendre un piège à la personne qui les reçoit, par exemple en donnant l'impression d'être envoyés par quelqu'un qu'elle connaît bien, et l'amener à ouvrir le message et à fournir des renseignements personnels. Ces diverses possibilités font que de nombreuses personnes ont des craintes concernant la sécurité du site Web, ou le manque de celle-ci, et la sécurité de l'information qu'elles fournissent en réponse aux demandes d'enquête par Internet. Le manque de confiance dans les enquêtes par Internet et les craintes que l'information puisse être conservée et utilisée à des fins autres que celle de l'enquête sont également des obstacles possibles à la réponse.

Les enquêtes à grande échelle, surtout celles bien connues du public, comme un recensement national qui comprend une campagne de communication générale préalable invitant à répondre, constituent une cible tentante pour ceux qui espèrent nuire au processus de réponse. Donc, même si le commanditaire de l'enquête est connu, la perception de risque peut être grande. Dans le cas du Recensement de l'Australie, les efforts de communication visaient à inviter chacun à répondre un « jour de recensement » particulier, ce qui a empiré la situation. Donc, en plus de devoir lutter contre la possibilité d'une cyberattaque, les commanditaires d'une enquête doivent aussi relever le défi de rétablir la confiance dans le système de collecte des données.

Les attaques intentionnelles sur des ordinateurs et des appareils individuels ou sur des enquêtes particulières sont probablement le plus grand péril pesant sur la réalisation d'enquêtes par Internet. Elles justifient aussi l'élaboration de multiples options quant au mode de réponse afin de ne pas dépendre entièrement d'Internet. Le recours à des méthodologies axées sur le Web comprenant plusieurs modes de réponse offre un certain degré de protection contre les attaques visant une enquête particulière, tout comme il offre maintenant une autre option aux personnes qui considèrent une réponse par Internet inacceptable. Dans le cas des très grandes enquêtes, comme les recensements nationaux, ne plus demander à tout le monde de répondre le même jour pourrait aussi réduire l'exposition à certains problèmes que peut poser Internet, ainsi que leur impact.

Il est difficile de prévoir si les progrès en matière de contrôle technologique et social annuleront les risques associés à l'utilisation de l'ordinateur. Pour le moment, il s'agit d'un problème menaçant le bon déroulement des enquêtes sur Internet que l'on ne peut ignorer.

6.2 Les téléphones intelligents et le problème « sac à main/poche »

Un deuxième problème, assez différent qui représente aujourd'hui un défi pour la collecte des données par Internet est l'utilisation de multiples appareils pour répondre. De plus en plus fréquemment, les gens transportent avec eux un appareil informatique – principalement un téléphone intelligent. À de nombreux égards, il s'agit d'un fait nouveau très positif. Puisque les gens transportent avec eux tout au long de la journée un appareil leur donnant la capacité de participer à une enquête, ils peuvent répondre aux demandes

de participation presque n'importe quand de n'importe où. Cette disponibilité constante met aussi en évidence ce que l'on peut décrire comme étant le problème de la poche ou du sac à main. Des préférences et probablement des limites de taille sont associées aux appareils que la plupart des gens sont disposés à transporter avec eux en vue de les utiliser dans les automobiles, dans les moyens de transport en commun, en travaillant et pendant leurs loisirs.

Des travaux de recherche récents ont montré que, si des proportions croissantes de la population répondent à des demandes par Internet sur leurs téléphones intelligents, la petite taille des écrans constitue un problème important. D'importants travaux de recherche résumés ailleurs (Dillman, Hao et Millar 2016) ont révélé que la proportion de réponses par téléphone intelligent a augmenté. En outre, il est difficile de poser de nombreux types de questions qui semblaient bien fonctionner pour d'autres modes d'enquête. Ainsi, Sarraf, Brooks, Cole et Wang (2015) ont montré que la présentation courante des questions, c'est-à-dire la question à gauche et les catégories de réponse présentées horizontalement à droite et l'échelle de quatre points placée en dessous aboutissait à l'abandon rapide du processus de réponse et à une augmentation spectaculaire des réponses manquantes. Dans un ensemble ultérieur d'expériences, Barlas et Thomas (2016) ont donné la preuve qu'il est avantageux de raccourcir les questions associées à une échelle. Ces travaux soulèvent la question de savoir s'il est souhaitable de poser des questions sur des échelles de sept points entièrement étiquetées, souvent privilégiées par le passé comme étant parfaites pour les enquêtes menées par des intervieweurs. Des travaux réalisés par Stern, Sterrett et Bilgen (2016) donnent à penser que les grilles – dans lesquelles une question générale établissant un ensemble de catégories de réponse est suivie par des listes d'items nécessitant chacun une réponse, un élément de base des questionnaires sur papier et par Internet, ne sont pas une présentation visuelle acceptable pour les téléphones intelligents.

Une excellente revue des études disponibles faite par Couper et coll. (2017) aboutit à la conclusion que les questionnaires remplis sur les téléphones mobiles sont associés à des taux de réponse plus faibles, des taux d'abandon plus élevés, et des temps d'achèvement plus longs que ceux des enquêtes par Internet remplies sur des ordinateurs personnels. Les auteurs soulignent que ces problèmes persistants pourraient être dus partiellement au fait que les spécialistes des enquêtes n'ont pas encore réussi à optimiser la conception pour les téléphones mobiles. Un autre facteur qui contribue à ces problèmes pourrait être la concurrence des demandes d'attention entre les téléphones intelligents et d'autres activités pendant que les gens vaquent à leurs activités quotidiennes.

L'un des défis que pose la conception de questionnaires pour les téléphones intelligents consiste à maintenir une construction unifiée des questions pour tous les modes d'enquête. Ce problème pourrait être particulièrement aigu quand les répondants à des enquêtes bien établies constatent que les structures, énoncés et présentations visuelles des questions utilisées précédemment sont modifiés unilatéralement en vue de leur utilisation sur les téléphones intelligents. Cette difficulté a été mentionnée par Mistichelli, Eanes et Horwitz du U.S. Census Bureau (2015). Il n'est pas encore certain que les concepteurs d'enquête sont disposés à modifier les façons utilisées de longue date de poser des questions, (par exemple questions sur les attitudes comportant moins de catégories et demander les items d'une série comme des items individuels plutôt que comme une liste d'items introduite par une question qui s'applique au groupe complet d'items à évaluer). S'il faut utiliser une construction unifiée pour les divers modes sur les téléphones intelligents, les

exigences que pose ce genre d'appareil seront vraisemblablement le principal déterminant de la façon dont les questions sont présentées dans tous les modes.

Le défi que doivent relever aujourd'hui les spécialistes des enquêtes en ce qui concerne les téléphones intelligents et les téléphones mobiles est aussi beaucoup plus profond que la façon de présenter les questions efficacement dans un plus petit espace sans devoir effectuer un déroulement horizontal et vertical. À l'époque des premières enquêtes, les intervieweurs qui se rendaient sur place pouvaient, par leur présence, obtenir l'attention complète du répondant. Dans le cas des courriels, des ordinateurs de bureau, des ordinateurs portables et des tablettes, on pourrait s'attendre à ce que les répondants répondent souvent, voire normalement, aux enquêtes au moment où ils sont le moins susceptibles d'être interrompus. Les téléphones intelligents, de par leur nature, sont des appareils d'interruption, puisqu'il est possible de recevoir des messages texte, des appels téléphoniques vocaux ou des courriels à tout moment, souvent pendant que l'on vaque physiquement à ses activités quotidiennes. La réponse à certaines enquêtes nécessite de consulter des dossiers auxquels on n'a pas accès quand on n'est pas à son domicile, ou de consulter un autre membre du ménage, ce qui semble plus difficile à faire si l'on essaie de répondre à un questionnaire en se déplaçant. La concurrence pour l'attention qui se produit avec ce genre d'appareil pourrait mener un enquêteur à inviter un répondant à ne pas répondre au questionnaire sur un téléphone intelligent et à lui demander de le faire plutôt sur son ordinateur portable ou à la maison. Le problème avec cette approche est que, pour un nombre important de personnes, le téléphone intelligent est peut-être leur seul ordinateur ou le seul qu'elles utilisent quotidiennement. En outre, il semble vraisemblable que plus on met d'obstacles à la réponse à un questionnaire « à l'instant présent », le moins les gens sont susceptibles d'y répondre.

Résoudre ces problèmes est l'un des plus grands défis qui se posent aux méthodologistes d'enquête aujourd'hui. Du côté positif, lorsque plusieurs modes de prise de contact sont utilisés et plusieurs moyens de répondre sont offerts, il semble plus facile d'orienter les répondants vers le moyen le plus efficace pour eux de répondre, ainsi que d'assurer le succès de l'enquête.

6.3 L'hésitation des commanditaires à entreprendre des enquêtes à modes de collecte mixtes et à modifier les procédures à mode unique

Un autre écueil des enquêtes axées sur le Web est associé au stress que de nombreuses organisations éprouvent à réaliser des enquêtes avec plusieurs modes de prise de contact et (ou) de réponse. Chaque mode de prise de contact et de réponse nécessite des compétences, du matériel et des logiciels spécialisés. Afin d'être efficace, l'enquête doit aussi être bien coordonnée afin de pouvoir résoudre de nombreux problèmes en une seule fois, comme il est décrit ailleurs (Dillman et coll. 2014, chapitre 11).

Les commanditaires d'enquête qui se sont spécialisés dans une forme particulière de collecte des données, ou qui souhaitent que les activités de collecte des données demeurent simples, pourraient être tentés d'éviter l'utilisation d'un deuxième ou d'un troisième mode de collecte. Il est peu probable que cela se produise si des taux de réponse élevés sont souhaités (par exemple un recensement national) ou qu'il existe un important incitatif économique à pousser les premiers répondants vers le Web (par exemple Biemer et coll. 2015). Cependant, le développement de logiciels prêts à utiliser a incité de nombreux enquêteurs à trouver des moyens de recourir uniquement à la collecte de données par Internet. Selon des études

antérieures, les résultats présenteront un biais important en faveur d'un niveau d'études et d'un revenu plus élevés si la collecte des données est limitée aux réponses par Internet uniquement (Rookey, Hanway et Dillman 2008; Messer et Dillman 2011). Au fil du temps, ce biais pourrait être réduit, mais il semble que cela ne se soit pas encore produit dans le cas des populations générales. Une autre source de faible taux de réponse et de biais éventuel se manifeste quand les enquêteurs n'obtiennent que les adresses de courriel pour une enquête proposée, ce qui élimine la possibilité d'une prise de contact préalable par la poste permettant d'inclure une prime en vue d'inciter les participants à répondre par Internet.

Même si le besoin est grand, apporter les changements appropriés prend du temps. Durant les années 1990, le *U.S. Census Bureau* a élaboré une stratégie de collecte des données comportant une lettre d'introduction préalable, un questionnaire papier, et une carte postale de suivi (Dillman, Clark et Sinclair 1995), qui a été appliquée pour les recensements de 2000 et de 2010. Après que la stratégie axée sur le Web pour l'ACS ait été lancée en 2013, le *Bureau* a continué de suivre cette approche. Le problème qu'elle posait était que la carte postale de suivi ne pouvait pas fournir de renseignements sur le mot de passe (toute personne ramassant la carte postale aurait pu les lire), créant donc l'attente que le répondant retourne à la lettre de demande d'enquête par Internet pour obtenir cette information. En outre, la séquence d'envoi d'une lettre préalable informant les gens qu'ils allaient recevoir une demande de répondre (par Internet), d'une seconde lettre leur demandant de se mettre en ligne en utilisant l'information fournie, puis d'une carte postale de rappel paraissait inutilement laborieuse. Par conséquent, une procédure où la lettre d'introduction préalable était abandonnée et la carte postale, remplacée par une lettre de suivi a été proposée. Elle a été adoptée en août 2015 après avoir été mise à l'essai par le *Census Bureau* (Clark et Roberts 2016) et a donné lieu à une augmentation importante de 2,5 points de pourcentage du taux de réponse par Internet et à une légère réduction des coûts globaux.

De nombreux autres problèmes sont associés au passage d'un concept à mode unique à l'adoption généralisée d'enquêtes axées sur le Web comportant de multiples modes de réponse. Par exemple, comment les chercheurs évitent-ils la frustration des personnes désireuses de participer et qui sont irritées qu'on leur dise qu'elles devront attendre quelques semaines pour que la demande arrive ? En outre, si les numéros de téléphone sont disponibles, un appel téléphonique pourrait être utilisé comme rappel de suivi avec encouragement à participer, au lieu d'essayer simplement d'interviewer les gens par téléphone. Des essais expérimentaux de ces options doivent avoir lieu.

6.4 Les effets des nouvelles découvertes et des innovations

Prévoir l'avenir est difficile. Quand l'interview par téléphone a pris son essor au cours des années 1970, les ordinateurs personnels n'existaient pas encore. Et presque personne ne pensait ou n'imaginait même que seulement deux décennies plus tard nous pourrions apporter avec nous presque partout grâce à des connexions sans fil le téléphone qui avait été jusque-là attaché à nos foyers et nos lieux de travail. Au début des enquêtes par Internet durant les années 1990, peu de personnes s'attendaient à ce que les ordinateurs de bureau gros et encombrants qui avaient commencé à être présents dans les foyers évolueraient vers des ordinateurs portables que les gens pourraient transporter avec eux d'un endroit à l'autre, ou à ce que cet appareil évoluerait plus tard vers des tablettes et des téléphones intelligents avec écrans tactiles et dotés les

uns et les autres d'une beaucoup plus grande puissance informatique que les ordinateurs de bureau et les ordinateurs portables originaux.

Dans une analyse récente, Friedman (2016) expose en détail les changements monumentaux touchant les capacités et la puissance des appareils personnels, qui sont considérés comme allant de soi par une part croissante de la population mondiale. Il fait remonter ces capacités à la croissance exponentielle de cinq composantes distinctes des ordinateurs d'aujourd'hui, à savoir 1) les circuits intégrés qui effectuent les calculs, 2) les unités de mémoire qui sauvegardent et extraient l'information, 3) les systèmes de réseau qui fournissent les communications dans les ordinateurs et entre ceux-ci, 4) les applications logicielles qui permettent à différents ordinateurs d'effectuer diverses tâches individuellement et ensemble, et 5) les capteurs qui détectent le mouvement, le langage, la lumière, les sons et d'autres caractéristiques de l'environnement et les transforment en données numérisées. Il fait remonter l'accélération rapide de ces éléments au développement de l'iPhone et aux innovations connexes qui ont eu lieu depuis 2007, et leur amalgamation en ce qu'il décrit comme étant la supernova (ou le nuage).

Ces développements n'étaient qu'à peine prévus, même par nombre des innovateurs qui les ont créés. Essayer d'imaginer l'avenir n'est pas plus facile aujourd'hui que par le passé. Ainsi, l'activation par la voix de recherches sur ordinateur remplace rapidement la frappe individuelle ou le balayage de commandes sur les téléphones intelligents. Aux États-Unis, 20 % des recherches dans Google sur un combiné avec système d'exploitation Androïd sont aujourd'hui entrées vocalement (The Economist 2017). Les gens peuvent aussi dicter des courriels et des messages texte, avec un certain succès. Les réponses activées par la voix représenteront-elles la prochaine vague de progrès pour les concepteurs d'enquête ? Il est facile d'imaginer qu'une personne soit interviewée par son téléphone intelligent. Et se pourrait-il que des traductions simultanées d'une langue à une autre, qui peuvent être effectuées aujourd'hui avec un certain succès, deviennent courantes dans les enquêtes ? Toutefois, c'est en ça que réside un défi fondamental, décrit par Friedman, à savoir la vitesse à laquelle les humains et les sociétés peuvent s'adapter à ces changements.

Nombre de répondants prospectifs qui intéressent les enquêteurs continuent d'utiliser des téléphones, tandis que d'autres se dépêchent furieusement d'adopter l'appareil informatique et de communication le plus avancé qu'ils trouvent pratique. Et d'autres encore hésitent tout simplement à utiliser un ordinateur. Les différences entre les capacités et les préférences des gens obligent les enquêteurs à ne se situer ni trop en avance ni trop à la traîne par rapport à la majorité des gens.

La question qui se pose est donc de savoir si les méthodes axées sur le Web sont simplement une autre phase de transition de la conception des enquêtes qui pourrait s'évanouir aussi rapidement qu'elle a pris de l'ampleur. La concentration sur la conception personnalisée et à mode mixte qui semble maintenant dominer la pensée des concepteurs d'enquête témoigne de la reconnaissance de l'hétérogénéité des populations dont les enquêteurs cherchent à décrire les opinions et les comportements.

Pendant un certain temps, il semblait que certains enquêteurs pensaient que la valeur des enquêtes à modes mixtes tenait au fait de donner aux participants le choix du mode qu'ils utiliseraient pour répondre à une demande. Cependant, cela n'est vrai qu'en partie. La puissance de réponse réelle des plans de collecte des données à modes mixtes en vue d'améliorer les taux de réponse tient à l'exécution efficace de multiples prises de contact. Chaque prise de contact donne l'occasion de fournir de nouveaux renseignements au sujet

de la demande d'enquête et, dans certains cas, d'atteindre des personnes avec lesquelles on ne peut pas prendre contact par d'autres modes d'enquête. Quand les demandes de participation à une enquête sont offertes par différents modes, il existe souvent une occasion d'améliorer la couverture (en atteignant des personnes qui ne peuvent pas être rejointes par un autre mode) et d'arriver à offrir à la personne des arguments persuasifs de participer à l'enquête. En outre, le jalonnement de ces prises de contact peut aider à motiver les gens à répondre (par exemple un renforcement par courriel de la lettre envoyée par la poste qui facilite la réponse).

7 Résumé et conclusion

La collecte de données axée sur le Web, qui débute par une demande de répondre sur Internet envoyée par la poste et est l'un des principaux faits nouveaux en matière de conception d'enquête du début du XXI^{e} siècle, offre aujourd'hui la promesse de réaliser des enquêtes plus rapides et moins coûteuses. De nombreux enquêteurs ont été surpris du recours que l'on fait à l'heure actuelle à une première prise de contact par la poste. Si les enquêtes par la poste ont souvent été utilisées pour recueillir des données d'enquête, nombreux sont ceux qui s'attendaient à ce que l'usage croissant d'Internet les fasse disparaître.

L'élément critique qui a poussé à reconsidérer un usage plus fréquent des méthodes de prise de contact par la poste est attribuable à Link et coll. (2008) et à Battaglia et coll. (2008). Ces travaux de recherche ont montré que les listes d'adresses résidentielles disponibles auprès du service postal des États-Unis fournissaient la meilleure couverture d'échantillonnage des résidences américaines et pourraient être utilisées pour procéder à des enquêtes par la poste efficaces auprès du grand public. Ces travaux ont été motivés par le puissant désir de trouver d'autres options que les enquêtes téléphoniques par CA qui faisaient face à une baisse continue des taux de réponse et à d'autres défis.

Une série d'études, qui a débuté en 2007, portait sur les moyens d'utiliser les prises de contact par la poste pour pousser les membres des ménages vers le Web à partir de ces listes d'adresses. Ces travaux visaient à combiner les réponses par Internet ainsi que sur papier. Ils s'appuyaient sur plusieurs années de recherches antérieures sur les différences de mesure entre les modes de collecte des données qui montraient que les réponses aux questionnaires en ligne et sur papier étaient assez semblables à condition d'utiliser des structures, énoncés et présentations visuelles similaires des questions pour les deux méthodes de collecte des données. Dix comparaisons expérimentales effectuées dans le cadre de ces études ont affiché un taux de réponse pour la méthode axée sur le Web de 43 % des ménages, desquels environ 60 % provenaient d'Internet et le reste, d'un suivi par la poste (Dillman et coll. 2014). Dans plusieurs pays, d'importants programmes d'enquête ont étudié et adopté les méthodes axées sur le Web qui s'appuient non seulement sur Internet et l'envoi par la poste, mais dont les protocoles comprennent maintenant un suivi par téléphone et (ou) sur place. L'objectif est d'obtenir des taux de réponse plus élevés et des données de meilleure qualité, ce que l'on n'imaginait plus être possible pour les enquêtes auprès des ménages il y a une décennie.

Nous sommes aujourd'hui dans une période de conception sur mesure dans laquelle différents plans d'enquête sont utilisés pour des sujets, des populations et des situations d'enquête différents. Cependant, il paraît probable que le recours à des méthodes de collecte de données axées sur le Web augmentera à travers

le monde industrialisé, à mesure que les commanditaires d'enquête chercheront à profiter du faible coût de la collecte de données sur Internet pour réduire le coût global des enquêtes courantes.

Cependant, ces méthodes posent des défis qui méritent notre attention. L'un est la menace que font peser sur les enquêtes et les répondants les maliciels, le hameçonnage et les attaques de serveur. Un autre est lié à l'utilisation accrue de téléphones intelligents qui pourraient nécessiter de modifier considérablement la façon dont les questions sont structurées et présentées aux répondants. En outre, l'hésitation des organisations et des particuliers à accepter et à maîtriser la plus grande complexité associée au passage d'enquêtes à mode unique à des enquêtes à modes mixtes est aussi un obstacle important.

L'histoire des méthodes d'enquête au cours des 75 dernières années comprend d'importantes transitions, de la dominance des interviews sur place au recours massif aux méthodes de téléphonie vocale, et maintenant aux enquêtes en ligne et à modes mixtes. Il reste à voir si les méthodes axées sur le Web – dont l'usage à titre de remplacement croît aujourd'hui – auront une présence durable, ou si elles finiront par laisser la place à la collecte de données par Internet seulement ou à d'autres procédures novatrices qui n'ont pas encore été conçues.

Bibliographie

Anderson, M., et Perrin, A. (2016). 13% of Americans don't use the Internet, Who are they? Accessible à l'adresse http://www.pewresearch.org/fact-tank/2016/09/07/some-americans-dont-use-the-internet-who-are-they/. Consulté le 2 mai 2017.

Australian Bureau of Statistics (2016). Making Sense of the Census. Accessible à l'adresse http://www.abs.gov.au/websitedbs/censushome.nsf/home/2016. Consulté le 2 mai 2017.

Barlas, F.M., et Thomas, R.K. (2017). Good questionnaire design: Best practices in the mobile era. *American Association for Public Opinion Research*, 19 janvier.

Battaglia, M.P., Link, M.W., Frankel, M.R., Osborn, L. et Mokdad, A.H. (2008). An evaluation of respondent selection methods for household mail surveys. *Public Opinion Quarterly*, 72(3), 459-469.

Biemer, P., Murphy, J., Zimmer, S., Berry, C., Deng, G. et Lewis, K. (2016). A test of Web/PAPI protocols and incentives for the residential energy consumption survey. Article non-publié présenté à la conférence annuelle de l'American Association for Public Opinion Research, 13 mai.

Blankenship, A.B. (1977). *Professional Telephone Surveys*. New York: McGraw-Hill Book Company.

Blumberg, S.J., et Luke, J.V. (2017). Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, janvier à juin 2016.

Brick, J.M., et Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *Annals of the American Academy of Political and Social Science*, 645(1), 36-59.

Brick, J.M., Williams, D. et Montaquila, J.M. (2011). Address-based sampling for subpopulation surveys. *Public Opinion Quarterly*, 75(3), 409-428.

Christian, L.M., et Dillman, D.A. (2004). The influence of symbolic and graphical language manipulations on answers to paper self-administered questionnaires. *Public Opinion Quarterly,* 68, 1, 57-80.

Christian, L.M., Dillman, D.A. et Smyth, J.D. (2008). The effects of mode and format on answers to scalar questions in telephone and Web surveys. Dans *Advances in Telephone Survey Methodology,* (Éds., J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link et R.L. Sangster). New York: Wiley-Interscience, 250-275.

Christian, L.M., Parsons, N.L. et Dillman, D.A. (2009). Designing scalar questions for Web surveys. *Sociological Methods and Research,* 37(3), 393-425.

City of Sapporo (2015). The Japanese government is conducting a Population Census. Accessible à l'adresse https://www.city.sapporo.jp/city/english/news/news201508_1e.html. Consulté le 1er octobre 2016.

Clark, S., et Roberts, A. (2016). Evaluation of August 2015 ACS mail contact strategy modification. *2016 American Community Survey Research and Evaluation Report Memorandum Series ACS16-0RER-13*.

Couper, M.P., Antoun, C. et Mavletova, A. (sous presse). Mobile Web surveys. Dans *Total Survey Error in Practice*, (Éds., P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker et B.T. West). New Jersey: Hoboken.

de Leeuw, E.D. (1992). Data quality in mail, telephone and face-to-face surveys. *TT-Publications Amsterdam*.

de Leeuw, E.D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics,* 21(2), 233-255.

de Leeuw, E., Villar, A., Suzer-Gurtekin, T. et Hox, J. (sous presse). How to design and implement mixed-mode surveys in cross National Surveys: Overview and guideline. Dans *Total Survey Error in Practice*, (Éds., P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker et B.T. West). New Jersey: Hoboken.

Dillman, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*. New York: John Wiley & Sons, Inc.

Dillman, D.A. (2000). Mail and Internet Surveys: The Tailored Design Method, 2nd Edition. New York: John Wiley & Sons, Inc.

Dillman, D.A. (2002). Navigating the rapids of change: Some observations on Survey Methodology in the early 21st century. *Public Opinion Quarterly,* 66(3), 473-494.

Dillman, D.A. (2005). Telephone surveys. Dans *Encyclopedia of Social Measurement,* (Éd., K. Kempf-Leonard), Volume 3. Londres, Royaume-Uni: Elsevier Press, 757-762.

Dillman, D.A. (2007). *Mail and Internet Surveys: The Tailored Design Method*. 2007 Update with New internet. Visual and Mixed-mode Guide. New Jersey: Hoboken.

Dillman, D.A., et Christian, L.M. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods,* 17(1), 30-52.

Dillman, D.A., et Tarnai, J. (1988). Administrative issues in mixed-mode surveys. Dans *Telephone Survey Methodology*, (Éds., R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II et J. Waksberg), New York: John Wiley & Sons, Inc., 509-528.

Dillman, D.A., et Edwards, M.L. (2016). Designing a mixed-mode survey. Dans *The SAGE Handbook of Survey Methodology,* (Éds., C. Wolfe, D. Joye, T.W. Smith et Y.-c. Fu), Sage Publications, Thousand Oaks, CA, 255-268.

Dillman, D.A., Clark, J.R. et Sinclair, M.D. (1995). Incidence des lettres de préavis, enveloppes-réponse affranchies et cartes de rappel sur les taux de réponse par la poste lors du recensement. *Techniques d'enquête,* 21, 2, 173-179. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1995002/article/14394-fra.pdf.

Dillman, D.A., Gertseva, A. et Mahon-Haft, T. (2005). Achieving usability in establishment surveys through the application of visual design principles. *Journal of Official Statistics,* 21(2), 183-214.

Dillman, D.A., Hao, F. et Millar, M.M. (2016). Improving the effectiveness of online data collection by mixing survey modes. Dans *The Sage handbook of Online Research Methods, 2nd Edition,* (Éds., N. Fielding, R.M. Lee et G. Blank). Sage Publications, Londres, 220-237.

Dillman, D.A., Smyth, J.D. et Christian, L.M. (2014). Internet, Phone, Mail and Mixed-Mode Surveys: The Tailored Design Method, 4th Edition. New Jersey: Hoboken.

Dillman, D.A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J. et Messer, B.L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research*, 38(1), 1-18.

Dutwin, D., et Lavrakas, P. (2016). Trends in telephone outcomes, 2008-2015. *Survey Practice*, 9(3). Accessible à l'adresse http://www.surveypractice.org/.

Edwards, M.L., Dillman, D.A. et Smyth, J.D. (2014). An experimental test of the effects of survey sponsorship on Internet and mail survey response. *Public Opinion Quarterly,* 78(3), 734-750.

Finamore, J., et Dillman, D.A. (2013). How mode sequence affects responses by internet, mail and telephone in the national survey of college graduates. Présentation à l'European Survey Research Association, Ljubljana, Slovénie, 18 juillet.

Friedman, T.L. (2016). Thank you for Being Late: An Optimist's Guide to Thriving in the Age of Accelerations. New York, Farrar: Straus and Giroux.

Groves, R.M., et Kahn, R.L. (1979). *Surveys by Telephone*. New York: John Wiley & Sons, Inc.

Groves, R.M., et Peytcheta, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, 72(2), 167-189.

Harter, R., Battaglia, M.P., Buskirk, T.D., Dillman, D.A., English, N., Mansour, F., Frankel, M.R., Kennel, T., McMichael, J.P., McPhee, C.B., Montaquila, J., Yancey, T. et Zukerberg, A.L. (2016). Address-base sampling. *American Association for Public Opinion Research* Task Force Report. Accessible à l'adresse http://www.aapor.org/getattachment/Education-Resources/Reports/AAPOR_Report_1_7_16_CLEAN-COPY-FINAL-(2).pdf.aspx, 140 pages.

Hoffman, D.D. (2004). Visual Intelligence: How we Create What we See. New York: Norton.

Keeter, S., Miller, C., Kohut, A., Groves, R.M. et Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly,* 64(2), 125-148.

Kerlinger, F.N. (1965). *Foundations of Behavioral Research*. New York: Holt, Rinehart and Winston.

Krosnick, J.A., et Alwin, D.F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly,* 51(2), 201-219.

Link, M.W., Battaglia, M.P., Frankel, M.R., Osborn, L. et Mokdad, A.H. (2008). A comparison of address-based sampling (ABS) versus random-digit dialing (RDD) for General Population Surveys. *Public Opinion Quarterly*, 72(1), 6-27.

Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I. et Vehovar. V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research,* 50(1), 79-104.

McMaster, H.S., LeardMann, C.A., Speigle, S. et Dillman, D.A. (2016). An experimental comparison of web-push vs. paper-only survey procedures for conducting an in-depth health survey of military spouses. BMC Medical Research Methodology.

Messer, B.L. (2012). Pushing households to the web: Results from Web+mail experiments using address based samples of the general public and mail contact procedures. Thèse de doctorat. Washington State University, Pullman.

Messer, B.L., et Dillman, D.A. (2011). Surveying the general public over the Internet using address-based sampling and mail contact procedures. *Public Opinion Quarterly,* 75(3), 429-457.

Messer, B.L., Edwards, M.L. et Dillman, D.A. (2012). Determinants of item nonresponse to Web and mail respondents in three address-based mixed-mode surveys of the general public. *Survey Practice,* 5(2), 1-9. Article accessible à l'adresse http://wwww.surveypractice.org/.

Millar, M.M., et Dillman, D.A. (2011). Improving response to Web and mixed-mode surveys. *Public Opinion Quarterly,* 75(2), 249-269.

Mistichelli, J., Eanes, G. et Horwitz, R. (2015). Centurion: Internet Data Collection and Responsive Design. Présentation au Federal Economic Statistics Advisory Committee, 12 juin.

Mohorko, A., de Leeuw, E. et Hox, J. (2013). Coverage bias in European telephone surveys: Developments of landline and mobile phone coverage across countries and over time. *Survey Methods: Insights from the Field*. Récupéré à partir de http://surveyinsights.org/?p=828.

Nathan, G. (2001). Méthodes de téléenquêtes applicables aux enquêtes-ménages – Revue et réflexions sur l'avenir. *Techniques d'enquête,* 27, 1, 7-34. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2001001/article/5851-fra.pdf.

O'Muircheartaigh, C., English, N. et Eckman, S. (2007). Predicting the relative quality of alternative sampling frames. *2007 Proceedings of the Survey Research Methods Section,* American Statistical Association, [CD ROM], Alexandrie, VA: American Statistical Association.

Palmer, S.E. (1999). *Vision Science: Photons to Phenomenology*. Londres: Bradford Books.

Parten, M. (1950). *Surveys, Polls and Samples*. New York: Harper and Brothers.

Pew Research Center (2012). Assessing the representativeness of public opinion surveys. Accessible à l'adresse http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/. Consulté le 24 octobre 2016.

Redline, C.D., et Dillman, D.A. (2002). The influence of alternative visual designs on respondents' performance with branching instructions in self-administered questionnaires. Dans *Survey Nonresponse,* (Éds., R. Groves, D. Dillman, J. Eltinge et R. Little), New York: John Wiley & Sons, Inc.

Redline, C.D., Dillman, D.A., Dajani, A. et Scaggs, M.A. (2003). Improving navigational performance in U.S. census 2000 by altering the visual languages of branching instructions. *Journal of Official Statistics,* 19(4), 403-420.

Roberts, C., Joye, D. et Staehli, M.E. (2016). Mixing modes of data collection in Swiss social survey: Methodological report of the LIVES-FORS mixed mode experiment. Document de travail 2016.48. Swiss National Centre of Competence in Research, a research instrument of the Swiss National Science Foundation.

Rookey, B.D., Hanway, S. et Dillman, D.A. (2008). Does a probability-based household panel benefit from assignment to postal response as an alternative to Internet-only? *Public Opinion Quarterly,* 72(5), 962-984.

Sarraf, S., Brooks, J., Cole, J. et Wang, X. (2015). What is the impact of smartphone optimization on long surveys? Présentation à l'American Association for Public Opinion Research Annual Conference, Hollywood, FL, 16 mai.

Smyth, J., Christian, L.M. et Dillman, D.A. (2008). Does 'Yes or No' on the telephone mean the same as check-all-that-apply on the Web? *Public Opinion Quarterly,* 72(1), 103-111.

Smyth, J.D., Dillman, D.A., Christian, L.M. et McBride, M. (2009). Open-ended questions in Web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly,* 73(2), 325-337.

Smyth, J.D., Dillman, D.A., Christian, L.M. et O'Neill, A.C. (2010). Using the Internet to survey small towns and communities: Limitations and possibilities in the early 21st century. *American Behavioral Scientist,* 53(9), 1423-1448.

Smyth, J.D., Dillman, D.A., Christian, L.M. et Stern, M.J. (2006). Comparing check-all and forced-choice question formats in Web surveys. *Public Opinion Quarterly,* 70(1), 66-77.

Statistique Canada (2016). Taux de réponse de la collecte du Recensement de la population de 2016. Accessible à l'adresse http://www12.statcan.gc.ca/census-recensement/2016/ref/Taux-reponse-fra.cfm. Consulté le 24 octobre 2016.

Statistics Japan (2015). Almost 20 million households responded online in the 2015 Population Census of Japan. Accessible à l'adresse http://www.stat.go.jp/english/info/news/20151019.htm. Consulté le 1er octobre 2016.

Stern, M., Sterrett, D. et Bilgen, I. (2016). The effects of grids on Web surveys completed with mobile devices. *Social Currents*, 3(3), 217-233.

Sullivan, M., Leong, C., Churchwell, C. et Dillman, D.A. (2015). Measurement and Cost Effects of Pushing Household Survey Respondents to the Web for Surveys of Electricity and Gas Customers in the United States. Article non-publié présenté à l'European Survey Research Association, Reykjavik, Islande, 16 juillet.

Tancreto, J. (2012). 2011 American Community Survey Internet Tests: Results from First Test in April 2011. #ACS12-RER-13-R2. 2012 American Community Survey Research and Evaluation Report Memorandum Series, 25 juin.

Tarnai, J., et Dillman, D.A. (1992). Questionnaire context as a source of response differences in mail and telephone surveys. Dans *Context Effects in Social and Psychological Research,* (Éds., N. Schwarz et S. Sudman), New York: Springer Verlag, Inc. 115-129.

The Economist (2017). Now we're talking: Voice technology is making computers less daunting and more accessible. Du 7 au 13 janvier, 422 (n^{o} 9022), 9.

Thomas, R., et Barlas, F. (2016). It's a Small Screen After All: Improving Measurement in an Ever-changing Online Survey World. GFK Webinar, 27 septembre.

Tourangeau, R. (2017). Mixing modes: Tradeoffs among coverage, nonresponse and measurement error. Dans *Total Survey Error in Practice,* (Éds., P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N. Clyde Tucker et B.T. West), New Jersey, Hoboken: John Wiley & Sons, Inc.

Tourangeau, R., Couper, M.P. et Conrad, F. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly,* 68(3), 368-393.

United Kingdom Cabinet Office (2016). Consultation Response: Community Life Survey: Development and implementation of online Survey Methodology for future survey years. Accessible à l'adresse https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/539111/community_life_survey_consultation_response_final.pdf.

United States Census Bureau (2014). American Community Survey Design and Methodology, Version 2.0. Accessible à l'adresse http://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html. Consulté le 15 octobre 2016.

Ware, C. (2004). Information Visualization: Perception for Design, 2nd Edition. Karlsruhe, West German: Morgan Kaufman.

Williams, D., Brick, J.M., Montaquila, J.M. et Han, D. (2014). Effects of screening questionnaires on response in a two-phase postal survey. *International Journal of Social Research Methodology,* 19(1), 51-67.

Techniques d'enquête, juin 2017
Vol. 43, nº 1, pp. 35-45
Statistique Canada, nº 12-001-X au catalogue

Méthode de perturbation multiniveau pour la protection des données tabulaires

Jean-Louis Tambay[1]

Résumé

La protection de la confidentialité des données qui figurent dans des tableaux de données quantitatives peut devenir extrêmement difficile lorsqu'on travaille avec des tableaux personnalisés. Une solution relativement simple consiste à perturber au préalable les microdonnées sous-jacentes, mais cela peut avoir un effet négatif excessif sur la précision des agrégats. Nous proposons plutôt une méthode perturbatrice qui vise à mieux concilier les besoins de protection et de précision des données en pareil cas. La méthode consiste à traiter par niveaux les données de chaque cellule en appliquant une perturbation minime, voire nulle, aux valeurs inférieures et une perturbation plus importante aux valeurs supérieures. La méthode vise avant tout à protéger les données personnelles, qui sont généralement moins asymétriques que les données des entreprises.

Mots-clés : Confidentialité; perturbation des données; données tabulaires.

1 Introduction

Des pressions sont exercées sur les organismes statistiques pour qu'ils fournissent davantage d'information, provenant de leurs propres données, aux utilisateurs externes. Bon nombre d'entre eux permettent aujourd'hui la création de tableaux personnalisés à l'aide de systèmes d'interrogation en ligne, mais les risques que des renseignements confidentiels soient divulgués s'accroissent avec la quantité de données diffusées. Pour résoudre ce problème, les organismes peuvent passer d'un extrême à l'autre, c'est-à-dire d'un souci de restreindre radicalement la quantité de renseignements diffusés à une volonté de modéliser des microdonnées synthétiques pour créer des produits. Les méthodes perturbatrices, qui ajoutent du bruit aux microdonnées ou aux résultats agrégés, se situent entre ces deux extrêmes. Nous proposons dans la présente une méthode de perturbation pour les données administratives quantitatives, comme les données sur l'impôt des particuliers, qui peut être utilisée aux fins de la création de tableaux personnalisés. La section 2 fournit des renseignements contextuels ainsi qu'une description des objectifs et des modes reconnus de protection des tableaux de données quantitatives. La section 3 décrit la méthode de perturbation multiniveau (MPM) proposée, ainsi que certaines de ses propriétés. La section 4 comporte une évaluation empirique, alors que la section 5 traite des questions encore en suspens.

2 Contexte

La stratégie proposée vise à protéger la confidentialité des tableaux de données quantitatives dans un cadre de production semi-contrôlée de tableaux personnalisés. Elle a été conçue avant tout pour des données administratives (s'apparentant à celles du recensement), et notamment pour les données sur l'impôt des particuliers. À Statistique Canada, la diffusion de telles données est assujettie à des règles de contrôle de la

1. Jean-Louis Tambay, Statistique Canada, Ottawa, Canada, K1A 0T6. Courriel : jean-louis.tambay@canada.ca.

divulgation, notamment la définition de tailles minimales de population pour les régions géographiques identifiables, l'application de règles relatives à la taille minimale des cellules et de règles de dominance pour supprimer des cellules sensibles (confidentielles), ou le recours à une suppression de cellules complémentaires (SCC) pour empêcher toute récupération de valeurs de cellules sensibles.

Alors que l'utilisation des données personnelles présente foncièrement moins de dangers que celle des données des entreprises, les données personnelles font plus fréquemment l'objet de tableaux personnalisés. Et si ces tableaux deviennent plus accessibles, il sera aussi de plus en plus difficile de procéder efficacement à des suppressions de cellules complémentaires. D'autres méthodes doivent donc être envisagées. La méthode que nous proposons consiste à appliquer indépendamment une technique perturbatrice à toute cellule non sensible de tout tableau. Seules les cellules sensibles sont supprimées, bien qu'on puisse envisager d'en diffuser quelques-unes une fois perturbées. La méthode vise à protéger les cellules sensibles des tableaux, ainsi qu'à prévenir la divulgation par recoupements découlant de tableaux multiples, surtout par la prise de différences sur des totaux imbriqués. Le but dans ce cas est de protéger deux totaux qui diffèrent par une unité.

Nous supposons l'existence d'un cadre semi-contrôlé où l'accès est quelque peu restreint, ou du moins jamais anonyme, et où donc il y a une surveillance et un contrôle quelconques des demandes. C'est une précaution qui s'impose, puisqu'en offrant sans restriction des tableaux à des pirates anonymes cherchant à exploiter toute vulnérabilité (en particulier, en multipliant les demandes pour obtenir des ensembles d'unités soigneusement choisis), on prête le flanc à une divulgation approximative de valeurs d'unités dans certaines conditions. Notre méthode est conçue pour des données s'apparentant à celles du recensement, qui sont plus à risque, mais elle pourrait sans aucun doute s'adapter à des données-échantillons au besoin. Notre stratégie convient mieux aux données personnelles, car elles sont moins susceptibles de dominance que les données des entreprises et les cellules quasi dominantes sont celles qui sont perturbées le plus. Mais sous réserve d'une certaine adaptation, les utilisateurs seraient à même de constater dans quelle mesure la stratégie pourrait répondre à leurs besoins pour d'autres types de données.

Dans la mesure du possible, nous aimerions employer cette stratégie pour remédier à d'autres problèmes de divulgation, notamment assurer la protection des rapports et d'autres genres de données. D'autres avantages seraient la capacité de traiter les zéros et les valeurs négatives, le maintien de la qualité des données, la préservation de l'additivité des tableaux, et des aspects opérationnels comme la simplicité de calcul et le recours à un minimum d'intervention manuelle.

Dans le présent exposé, nous appliquons une règle du pourcentage P pour reconnaître les totaux de cellules sensibles, une cellule étant sensible si la contribution globale des plus petites unités, à partir de la troisième en importance, est inférieure à tel pourcentage $P\,\%$ de la valeur de la plus grande unité (si $X - x_1 - x_2 < P\,\%\, x_1$, où X est le total de la cellule et où x_i est la contribution de sa i^{e} unité en importance). Nous supposons que les cellules non conformes à la règle de la taille minimale de cellule sont sensibles elles aussi.

Nous désirons préserver la qualité et la confidentialité des données quantitatives dans un cadre de production de tableaux personnalisés. Des techniques applicables à des tableaux de données quantitatives comme la suppression de cellules complémentaires (Cox et Sande 1979) et l'ajustement tabulaire contrôlé

(Cox et Dandekar 2004) ne donnent pas de très bons résultats dans un tel cadre. Il nous faut résoudre des problèmes d'optimisation pour dégager des solutions par tableau. Des problèmes commencent à se poser quand on a à protéger des tableaux vastes, complexes ou liés (couplés); on sera alors incapable d'en venir à une solution ou bien une démarche heuristique risquera de créer des incohérences de suppression ou de perturbation qu'exploiteraient des pirates. Il est bien plus facile de perturber directement les totaux de cellules, notamment par l'application d'un bruit aléatoire, mais on aura toujours à s'attacher aux microdonnées pour assurer une protection suffisante, tout en contrôlant l'effet sur la qualité. Sans des mesures complémentaires, des incohérences pourraient apparaître dans et entre les tableaux, et les pirates en profiteraient.

Une perturbation des microdonnées, c'est-à-dire au niveau des microdonnées, convient mieux à un cadre multitableaux. Les tableaux sont additifs et habituellement exempts de toute suppression, et les résultats sont cohérents entre tableaux. Si l'on permet des tableaux personnalisés, quelqu'un pourrait peut-être récupérer certaines valeurs perturbées, soit directement, soit par prise de différences. Le degré de bruit appliqué à chaque unité doit donc être assez élevé pour qu'on réalise le degré d'ambiguïté recherché, et c'est pourquoi le bruit accumulé risque d'être ample pour des agrégats donnés. Une méthode de perturbation des microdonnées conçue et employée au *U.S. Census Bureau* s'appelle la méthode EZS (Evans, Zayatz et Slanta 1998). Elle consiste à multiplier les différentes valeurs x_i par un poids $w_i = 1 + \varepsilon_i$, où ε_i représente des variables aléatoires indépendantes et identiquement distribuées (i.i.d.) à moyenne 0 et à variance σ_ε^2. Mentionnons deux distributions des ε_i d'intérêt, soit la distribution triangulaire divisée (voir la figure 2.1) et la distribution uniforme divisée (voir la figure 2.2) où les valeurs correspondantes de σ_ε^2 sont $(3a^2 + 2ab + b^2)/6$ et $(a^2 + ab + b^2)/3$, respectivement. Les ε_i (ou les w_i) sont attachés en permanence à leur unité i. Comme le même bruit est appliqué à toutes les variables, il n'y a aucune incidence sur les rapports. S'il est nécessaire de protéger les rapports, il devrait y avoir des valeurs de pondération w_i différentes selon les variables, ou des poids par unité pourraient être utilisés conjointement avec des poids par variable d'unité.

Figure 2.1 Distribution triangulaire divisée.

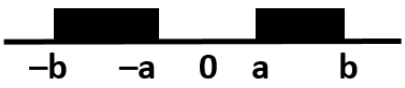

Figure 2.2 Distribution uniforme divisée.

Il existe des moyens d'atténuer l'effet accumulé de la perturbation des microdonnées sur la qualité. Massell et Funk (2007) proposent d'équilibrer les bruits aléatoires appliqués aux cellules d'un tableau primaire pour limiter leur incidence. Dans d'autres méthodes, on perturbe les microdonnées, mais pas toujours de la même manière et en créant donc certaines incohérences dans les résultats. Giessing (2011) propose de multiplier les valeurs d'unités x_i par $w_i = 1 \pm |\varepsilon_i|$, pour ε_i i.i.d. $\mathrm{N}(0, \sigma_0^2)$, sauf dans les cellules sensibles, où la valeur la plus grande serait multipliée par $w_i = 1 \pm (\mu_0 + |\varepsilon_i|)$. On choisit la valeur μ_0 pour assurer un degré approprié de protection des cellules sensibles, d'où la possibilité d'utiliser dans l'ensemble une valeur inférieure de σ_0^2. Il reste que, si σ_0^2 est trop bas, la méthode ne protège peut-être pas suffisamment contre la divulgation par prise de différences. L'*Australian Bureau of Statistics* a conçu la méthode des principales contributions (*Top Contributors Method* ou TCM) pour son application d'accès à distance TableBuilder; celle-ci consiste à perturber les principaux répondants dans chaque cellule d'une manière semi-cohérente, seule une partie du bruit étant appliquée uniformément (Thompson, Broadfoot et Elazar 2013). La méthode de perturbation multiniveau fait appel à certains de ces concepts, mais elle protège davantage contre la prise de différences, comme nous allons l'expliquer.

D'autres stratégies courantes comme l'arrondissement, l'échantillonnage (ou le sous-échantillonnage) et l'échange d'unités, entre régions voisines disons, se prêtent mieux à une protection des tableaux statistiques.

3 Méthode de perturbation multiniveau (MPM)

3.1 Description

Cette méthode perturbatrice porte sur les totaux et s'attaque aux possibilités de divulgation par prise de différences. Lorsqu'elle est utilisée dans des tableaux de données quantitatives, elle permet de limiter la suppression aux cellules sensibles. Trois idées fondamentales sont exploitées et les deux premières sont semblables à celle de la méthode TCM des principales contributions.

La première est l'attribution de nombres pseudo-aléatoires (NPA) de hachage aux unités pour la production de résultats cohérents de perturbation au besoin. On décourage ainsi la multiplication des demandes pour une meilleure estimation des totaux non perturbés. On se sert de la méthode EZS pour multiplier la valeur d'une unité i par un poids $w_i = 1 + \varepsilon_i$, avec $\varepsilon_i \sim (0, \sigma_\varepsilon^2)$ comme nous l'avons indiqué. Pour que des résultats cohérents puissent être obtenus, les ε_i sont tirés d'un NPA par unité en distribution uniforme sur l'intervalle $[0,1)$. On peut, par exemple, utiliser $h_i / 1000$, où les h_i sont tirés du numéro d'assurance sociale (SIN pour *Social Insurance Number*) (par exemple, $h_i = Mod(SIN_i \bullet P,\ 1000)$ pour P comme nombre premier élevé). Avec h_i, l'unité i est toujours perturbée de la même manière. Pour la perturber de la même manière seulement quand elle figure dans le même total de cellule, on produit le bruit $w_i' = 1 + \varepsilon_i'$ par unité de cellule à partir de $h_i' = Mod(h_i + h_{tot},\ 1000) / 1000$, où $h_{tot} = \sum_{i \in cell} h_i$. On emploie des guillemets pour désigner les bruits et les perturbations par unité de cellule. Toutes les valeurs de bruit sont calculées à partir de h_i ou h_i'.

La deuxième idée est celle d'une perturbation multiniveau des unités de chaque cellule. On se trouve à perturber les quatre unités les plus importantes d'une manière aléatoire mais *cohérente* à l'aide des poids de

perturbation w_i tirés de h_i. Les unités qui suivent en importance, 5 à 9 disons, sont perturbées à leur tour, mais d'une manière semi-cohérente. On emploie dans ce cas un mélange de poids w_i par unité et de poids w_i' par unité de cellule. Les plus petites unités ne sont pas perturbées et leurs valeurs sont protégées contre la prise de différences par la perturbation des unités 5 à 9 par unité de cellule : en effet, en ajoutant ou retranchant une unité dans une cellule, si petite soit-elle, on change les w_i' de ces unités. Le nombre d'unités par niveau est au choix, mais nous avons constaté que, s'il y en a quatre ou cinq, respectivement les résultats sont satisfaisants.

La troisième idée vise surtout la question de la prise de différences. La direction du bruit est inversée dans le cas des unités de rang pair (w_i tirés des $(-1)^{i+1}\varepsilon_i$), ce qui accroît la variance des différences quand une unité du haut est changée. Dans le cas des unités 5 à 9, un mélange aléatoire de w_i et de w_i' permet de diminuer le risque quand une petite unité est ajoutée ou retranchée. Enfin, on amplifie le bruit des trois unités du haut dans les cellules non sensibles plus dominantes. On peut ainsi abaisser le degré de bruit à appliquer généralement, ce qui atténue l'effet global de la perturbation sur la qualité des données.

Nous proposons comme application possible de la MPM de supprimer toutes les cellules sensibles et les petites cellules (par exemple, $n < 10$) et de perturber les autres. En raison de la protection permise par la perturbation, les cellules légèrement sensibles pourraient elles aussi être publiables. Pour les autres cellules d'un total $X = \sum_{i \in cell} x_i$, nous fixerions la valeur perturbée Z à

$$Z = X + K\varepsilon_1 x_1 - L\varepsilon_2 x_2 + M\varepsilon_3 x_3 - \varepsilon_4 x_4 - \sum_{i=5}^{9} \left\{(-1)^i \alpha_i \varepsilon_i - (1-\alpha_i)\varepsilon_i'\right\} x_i .$$

K, L et M permettent d'accroître le bruit de Z, au besoin (K, L et $M \geq 1$). Les α_i sont des variables aléatoires indépendantes des ε_i; par exemple, $\alpha_i \sim \text{Uniforme}(0,1)$ ou $\alpha_i = \text{Mod}(h_i, 8)/7$.

3.2 Quelques résultats

Prenons $\varepsilon_i, \varepsilon_i' \sim (0, \sigma_\varepsilon^2)$, $\alpha_i \sim \text{Uniforme}(0,1)$, i.i.d. et gardons K, L et M fixes (pour l'instant). Il s'ensuit :

$$E(Z) = X \text{ et } V(Z) = \left\{K^2 x_1^2 + L^2 x_2^2 + M^2 x_3^2 + x_4^2 + \tfrac{2}{3}\sum_{i=5}^{9} x_i^2\right\}\sigma_\varepsilon^2 .$$

Soit X_{-1}, X_{-2}, X_{-3} et Z_{-1}, Z_{-2}, Z_{-3} pour X et Z d'une cellule après le retrait des unités 1, 2 et 3. Si nous gardons les indices de la cellule de départ (c'est-à-dire que l'indice 2 est celui de l'unité deuxième pour X), nous obtenons :

$$\begin{aligned}
Z_{-1} &= X_{-1} + K\varepsilon_2 x_2 - L\varepsilon_3 x_3 + M\varepsilon_4 x_4 - \varepsilon_5 x_5 - \sum_{i=6}^{10} \left\{(-1)^i \alpha_i \varepsilon_i - (1-\alpha_i)\varepsilon_i'\right\} x_i , \\
Z_{-2} &= X_{-2} + K\varepsilon_1 x_1 - L\varepsilon_3 x_3 + M\varepsilon_4 x_4 - \varepsilon_5 x_5 - \sum_{i=6}^{10} \left\{(-1)^i \alpha_i \varepsilon_i - (1-\alpha_i)\varepsilon_i'\right\} x_i , \text{ et} \\
Z_{-3} &= X_{-3} + K\varepsilon_1 x_1 - L\varepsilon_2 x_2 + M\varepsilon_4 x_4 - \varepsilon_5 x_5 - \sum_{i=6}^{10} \left\{(-1)^i \alpha_i \varepsilon_i - (1-\alpha_i)\varepsilon_i'\right\} x_i .
\end{aligned}$$

Nous pouvons dégager Z_{-i} pour les autres unités de la même manière. Si nous estimons que les unités retirées sont $\hat{x}_i = Z - Z_{-i}$, on peut démontrer que, avec $G = 2\tfrac{2}{3}x_5^2 + 2\sum_{i=6}^{9} x_i^2 + \tfrac{2}{3}x_{10}^2$,

$$
\begin{aligned}
E\left(\hat{x}_i\right) &= x_i, \\
V\left(\hat{x}_1\right) &= \left\{K^2 x_1^2 + (K+L)^2 x_2^2 + (L+M)^2 x_3^2 + (M+1)^2 x_4^2 + G\right\}\sigma_\varepsilon^2, \\
V\left(\hat{x}_2\right) &= \left\{L^2 x_2^2 + (L+M)^2 x_3^2 + (M+1)^2 x_4^2 + G\right\}\sigma_\varepsilon^2, \\
V\left(\hat{x}_3\right) &= \left\{M^2 x_3^2 + (M+1)^2 x_4^2 + G\right\}\sigma_\varepsilon^2, \\
V\left(\hat{x}_4\right) &= \left\{x_4^2 + G\right\}\sigma_\varepsilon^2, \\
V\left(\hat{x}_5\right) &= \left\{\tfrac{2}{3}x_5^2 + 2x_6^2 + 2x_7^2 + 2x_8^2 + 2x_9^2 + \tfrac{2}{3}x_{10}^2\right\}\sigma_\varepsilon^2, \\
V\left(\hat{x}_6\right) &= \left\{\tfrac{2}{3}x_5^2 + \tfrac{2}{3}x_6^2 + 2x_7^2 + 2x_8^2 + 2x_9^2 + \tfrac{2}{3}x_{10}^2\right\}\sigma_\varepsilon^2, \\
V\left(\hat{x}_7\right) &= \left\{\tfrac{2}{3}x_5^2 + \tfrac{2}{3}x_6^2 + \tfrac{2}{3}x_7^2 + 2x_8^2 + 2x_9^2 + \tfrac{2}{3}x_{10}^2\right\}\sigma_\varepsilon^2, \\
V\left(\hat{x}_8\right) &= \left\{\tfrac{2}{3}x_5^2 + \tfrac{2}{3}x_6^2 + \tfrac{2}{3}x_7^2 + \tfrac{2}{3}x_8^2 + 2x_9^2 + \tfrac{2}{3}x_{10}^2\right\}\sigma_\varepsilon^2, \\
V\left(\hat{x}_9\right) &= \tfrac{2}{3}\left\{x_5^2 + x_6^2 + x_7^2 + x_8^2 + x_9^2 + x_{10}^2\right\}\sigma_\varepsilon^2, \quad \text{et} \\
V\left(\hat{x}_i\right) &= \tfrac{2}{3}\left\{x_5^2 + x_6^2 + x_7^2 + x_8^2 + x_9^2\right\}\sigma_\varepsilon^2, \quad \text{pour} \quad i > 9.
\end{aligned}
$$

Si nous supposons que K, L et M sont fixes, nous pouvons les établir en fonction d'une certaine exigence pour $\mathrm{V}\left(\hat{x}_i\right)$. Ainsi, nous pourrions vouloir $\mathrm{V}\left(\hat{x}_i\right) = x_i^2/30$ puisque, pour $z \sim \mathrm{N}(0,1)$, $\Pr\left(|z| > 0{,}44\right) = 0{,}66$, ce qui, pour $\hat{x}_i \sim \mathrm{N}\left(x_i, x_i^2/30\right)$, donne $\Pr\left\{|\hat{x}_i - x_i| \geq 8\,\% x_i\right\} = 66\,\%$.

Pour obtenir $\mathrm{V}\left(\hat{x}_i\right) = x_i^2/NN$, nous pouvons résoudre K, L et M (fixes) en ordre inverse, ce qui donne :

$$
\begin{aligned}
M &= \frac{\sqrt{\left(x_3^2 + x_4^2\right)\left(x_3^2/NN\sigma_\varepsilon^2 - G\right) - x_3^2 x_4^2} - x_4^2}{x_3^2 + x_4^2} \\
L &= \frac{\sqrt{\left(x_2^2 + x_3^2\right)\left(x_2^2/NN\sigma_\varepsilon^2 - G - x_4^2(M+1)^2\right) - M^2 x_2^2 x_3^2} - M x_3^2}{x_2^2 + x_3^2} \\
K &= \frac{\sqrt{\left(x_1^2 + x_2^2\right)\left(x_1^2/NN\sigma_\varepsilon^2 - G - x_3^2(L+M)^2 - x_4^2(M+1)^2\right) - L^2 x_1^2 x_2^2} - L x_2^2}{x_1^2 + x_2^2}
\end{aligned}
$$

Dans la pratique, L et M sont bornés au-dessous à 1 et au-dessus à une certaine valeur seuil inférieure à 2; K est borné au-dessous à 1 et peut être plafonné légèrement au-dessus de ce seuil. Ajoutons que les valeurs cibles de K, L et M dépendent de la situation dans chaque cellule. Pour l'illustrer simplement, nous avons posé ici qu'elles ne changeaient pas lorsque des observations étaient retranchées de la cellule.

Si nous appliquons le même bruit et inversons la direction dans le cas des unités de rang pair, nous tirons parti de la corrélation entre Z et Z_{-i} pour accroître la variance de $\hat{x}_i = Z - Z_{-i}$. Ainsi, la contribution de l'unité 2 à $V\left(\hat{x}_1\right)$ est $(K+L)^2 x_2^2 \sigma_\varepsilon^2$. Si nous avions appliqué des bruits indépendants (par unité de cellule) ε_i' au lieu de ε_i pour les unités 1 à 4, la contribution de l'unité 2 aurait seulement été $\left(K^2 + L^2\right) x_2^2 \sigma_\varepsilon^2$.

3.3 Comparaison avec les méthodes EZS et TCM

Avec la méthode EZS, le total de cellule perturbé est simplement $Z = X + \sum_{i \in cell} \varepsilon_i x_i$, d'où $V(Z) = \sum_{i \in cell} x_i^2 \sigma_\varepsilon^2$. Pour toute unité i, nous avons $E(\hat{x}_i) = x_i$ et $V(\hat{x}_i) = x_i^2 \sigma_\varepsilon^2$, ce qui est moins que la variance équivalente avec la MPM pour le même degré de bruit σ_ε^2 même quand nous établissons $K = L = M = 1$. Une exception possible pourrait être l'unité 5 si les unités qui suivent sont relativement des plus petites, ce qu'on peut voir en examinant $V(\hat{x}_5)$ plus haut.

La TCM applique trois facteurs multiplicatifs de perturbation aux unités les plus importantes, disons 4, de chaque cellule. Une composante quantitative M_i détermine la taille relative de la perturbation de l'unité au i^{e} rang. Les valeurs M_i sont fixes; on aurait normalement $M_1 > M_2 > M_3 > M_4$, par exemple [0,6; 0,4; 0,3; 0,2]. Un facteur aléatoire permanent $d_i = \pm 1$ fixe la direction du bruit pour chaque unité i. Un facteur pseudo-aléatoire $s_i > 0$ donne les valeurs de bruit par unité de cellule. On obtient ainsi $Z = X + \sum_{i=1}^{4} M_i d_i s_i x_i$. Par rapport à la MPM, la méthode peut être représentée sous une forme comparable avec $[M_1, M_2, M_3, M_4] = [K, L, M, 1]$, $d_i = \text{sign}(\varepsilon_i)$ et $s_i = |\varepsilon_i'|$. La façon de fixer les d_i s'écarte grandement de la méthode MPM qui diminue largement la protection assurée aux $\hat{x}_1$. Illustrons en considérant deux adaptations de ces méthodes qui livrent des variances identiques pour Z :

$$
\begin{aligned}
Z_{MPM} &= X + K\varepsilon_1 x_1 - L\varepsilon_2 x_2 + M\varepsilon_3 x_3 - \varepsilon_4 x_4, \quad \text{et} \\
Z_{TCM} &= X + K\,sign(\varepsilon_1)\,|\varepsilon_1'|\,x_1 + L\,sign(\varepsilon_2)\,|\varepsilon_2'|\,x_2 + M\,sign(\varepsilon_3)\,|\varepsilon_3'|\,x_3 + sign(\varepsilon_4)\,|\varepsilon_4'|\,x_4,
\end{aligned}
$$

où les conventions de notation que nous connaissons sont employées avec $K, L, M > 0$, ce qui donne :

$$
\begin{aligned}
V_{MPM}(\hat{x}_1) &= \left\{K^2 x_1^2 + (K+L)^2 x_2^2 + (L+M)^2 x_3^2 + (M+1)^2 x_4^2 + x_5^2\right\}\sigma_\varepsilon^2, \quad \text{et} \\
V_{TCM}(\hat{x}_1) &= K^2 x_1^2 \sigma_\varepsilon^2 + \left\{(K^2+L^2)x_2^2 + (L^2+M^2)x_3^2 + (M^2+1)x_4^2\right\}\sigma_{|\varepsilon|}^2 + x_5^2 \sigma_\varepsilon^2.
\end{aligned}
$$

Non seulement des facteurs comme $(K+L)^2$ sont plus grands que (K^2+L^2), mais la variance du bruit, σ_ε^2, est souvent remplacée par celle du bruit absolu, $\sigma_{|\varepsilon|}^2$, une valeur bien moindre. Dans le cas de la distribution triangulaire divisée, on va de $(3a^2 + 2ab + b^2)/6$ à $(b-a)^2/18$. Avec $b = 2a$, on tombe de $11a^2/6$ à $a^2/18$.

Ce n'est pas là une comparaison légitime entre les deux méthodes. Nous ne nous trouvons pas à utiliser la MPM réelle et les paramètres n'ont pas à être identiques, mais on peut voir la différence entre les deux pour les d_i.

4 Examen empirique

Nous avons appliqué les méthodes MPM et EZS aux données des particuliers d'un fichier fiscal. Deux variables ont été utilisées : $x =$ revenu (si > 0) et $y = x^2$ (pour accroître l'asymétrie). Nous avons produit des cellules comptant de 15 à 148 unités en combinant les groupes d'âge avec le code postal, le sexe et l'état matrimonial. Nous avons essayé différents degrés de bruit (ε_i) d'une distribution triangulaire divisée. Les résultats présentés sont ceux où $\sigma_\varepsilon^2 = 0{,}006$. À l'aide du cadre risque-utilité (Duncan, Keller-McNulty et Stokes 2001), nous avons regardé l'incidence des méthodes sur la précision des données et le risque.

Le tableau 4.1 montre l'effet de la MPM sur la qualité des totaux des cellules par tranche de taille de cellule. La méthode a été appliquée 500 fois à chaque cellule. Pour chaque tranche de taille, le tableau présente le nombre de cellules, le coefficient de variation (CV) moyen après perturbation et le pourcentage de fois que le total perturbé se situe dans une marge de 2 %, 5 %, 8 % et 12 % du total de départ. Aux fins de cette étude, nous avons présumé que les cellules non conformes à une règle du pourcentage P s'appliquant aux cellules sensibles où $P = 15$ seraient supprimées et exclues des résultats. Il y avait plus de ces cellules avec la variable y (comme on pourrait davantage l'observer avec les données des entreprises). Comme on pouvait s'y attendre, l'effet de la perturbation était supérieur pour les cellules plus petites et pour la variable y. Toutes les cellules perturbées à plus de 8 % étaient quasi sensibles et auraient été supprimées avec $P = 20$.

Tableau 4.1
Incidence de la méthode de perturbation multiniveau sur les totaux des cellules

	Variable = Revenu (x)						Variable = Revenu² (y)					
Taille	N^bre	CV	% de fois que la distance relative est ≤				N^bre	CV	% de fois que la distance relative est ≤			
de cell.	cell.	moy.	2 %	5 %	8 %	12 %	cell.	moy.	2 %	5 %	8 %	12 %
15 – 18	1 822	2,37	58,5	95,1	99,5	100,0	1 777	4,09	34,5	72,0	92,4	99,6
19 – 25	2 230	2,03	66,2	97,2	99,7	100,0	2 185	3,71	38,1	77,1	94,4	99,7
26 – 40	1 920	1,57	78,2	99,1	99,9	100,0	1 899	3,24	44,2	82,8	96,0	99,8
41 – 148	1 312	1,05	92,1	99,5	99,9	100,0	1 301	2,53	57,1	90,0	97,7	99,9
Ensemble	7 284	1,82	72,1	97,6	99,7	100,0	7 162	3,47	42,3	79,7	94,9	99,7

Nota : Une valeur de 100,0 est une valeur de plus de 99,95 arrondie à 100.

Le tableau 4.2 montre l'effet de l'application d'un bruit multiplicatif par la méthode EZS, pour le même σ_ε^2, aux totaux des cellules. Les résultats sont plutôt semblables pour le revenu (x) et ils sont sensiblement meilleurs pour y. Des résultats du même ordre ont été dégagés quand une valeur proche de 0,014 était utilisée pour σ_ε^2 (la méthode MPM était un peu meilleure avec x, et la méthode EZS, avec y).

Tableau 4.2
Incidence de l'application d'un bruit multiplicatif par la méthode EZS aux totaux des cellules

	Variable = Revenu (x)						Variable = Revenu² (y)					
Taille	N^bre	CV	% de fois que la distance relative est ≤				N^bre	CV	% de fois que la distance relative est ≤			
de cell.	cell.	moy.	2 %	5 %	8 %	12 %	cell.	moy.	2 %	5 %	8 %	12 %
15 – 18	1 822	2,33	58,7	97,1	100,0	100,0	1 777	3,19	41,2	86,4	99,8	100,0
19 – 25	2 230	2,08	64,5	98,5	100,0	100,0	2 185	2,93	45,2	90,0	99,9	100,0
26 – 40	1 920	1,74	73,9	99,6	100,0	100,0	1 899	2,59	51,4	93,8	99,9	100,0
41 – 148	1 312	1,30	86,9	99,9	99,9	100,0	1 301	2,09	63,4	97,1	100,0	100,0
Ensemble	7 824	1,91	69,6	98,7	99,9	100,0	7 162	2,76	49,2	91,4	99,9	100,0

Nota : Une valeur de 100,0 dans les colonnes 8 % est une valeur de plus de 99,95 arrondie à 100.

Nous avons ensuite examiné le degré de protection assuré aux unités les plus importantes de chaque cellule. Pour chacune des cellules, nous avons obtenu une estimation $\hat{x}_i$ pour l'unité x_i en prenant les différences sur les totaux perturbés des cellules avec et sans l'unité en question. Nous avons calculé les

différences relatives $d_i = 100\left|\hat{x}_i - x_i\right| / x_i$ et les avons intégrées à un score correspondant à $\sum_{cells} r_i$, où $r_i = 1$ si $d_i < 10$, $r_i = 0$ si $d_i > 15$ et $0 < r_i < 1$ dans les autres cas. Le tableau 4.3 présente les quartiles de d_i et les scores des variables x et y pour les douze unités les plus importantes de chaque cellule dans le cas de la méthode MPM et pour l'unité la plus importante dans le cas de la méthode EZS (laquelle assure le même degré de protection à toutes les unités).

Avec la MPM, les trois unités les plus importantes étaient généralement les plus protégées, comme on pouvait s'y attendre. La configuration est différente pour les variables x et y. Si on regarde les quartiles de d_i pour la variable x, le degré de protection diminue progressivement jusqu'à l'unité 10 et augmente par la suite. Comme les $\mathrm{V}(\hat{x}_i)$ sont les mêmes pour $i > 9$, les résultats devraient continuer à s'améliorer après la 10e unité en importance. Les scores racontent une même histoire. En ce qui concerne la variable y, la descente n'est pas aussi régulière et l'unité 5 est la moins protégée (l'unité 10 si on considère seulement le quartile 1). La protection la plus faible autour des unités 5 et 10 est prévue par les formules pour les $\mathrm{V}(\hat{x}_i)$, dont la configuration de base change autour de ces deux unités. L'unité 10 est la plus vulnérable en cas d'attaque ciblée répétée, une attaque consistant à tirer une estimation $\hat{x}_{10}$ des totaux pour les unités 1 à 10, et pour les unités 1 à 9, au moyen d'un certain ensemble d'unités plus petites (par exemple, tirer $\hat{x}_{10(i)}$ des totaux sans l'unité i et sans les unités i et 10 pour $i = 11, 12, 13\ldots$). Si l'on prend la moyenne des $\hat{x}_{10(i)}$, et s'il y en a suffisamment, on peut obtenir de bonnes estimations de x_{10}. De telles attaques exigent des demandes de tableaux soigneusement formulées, ce que pourrait décourager un cadre de production semi-contrôlée de tableaux.

Tableau 4.3
Protection des douze plus grandes unités avec la méthode MPM et de la plus grande avec la méthode EZS (quartiles de d_i)

	Variable = Revenu (x)					Variable = Revenu² (y)				
	Cellules	**Q1**	**Médiane**	**Q3**	**Score (%)**	**Cellules**	**Q1**	**Médiane**	**Q3**	**Score (%)**
Unité 1	7 962	7,9	15,7	26,6	3 196 (40)	7 823	7,6	14,4	23,2	3 365 (43)
Unité 2	7 962	8,6	17,5	29,3	2 895 (36)	7 782	7,2	15,0	25,2	3 311 (43)
Unité 3	7 962	8,1	16,9	28,7	3 021 (38)	7 782	6,6	14,1	24,2	3 522 (45)
Unité 4	7 962	7,2	15,5	26,2	3 314 (42)	7 799	6,1	13,3	22,5	3 726 (48)
Unité 5	7 962	6,4	13,9	23,8	3 647 (46)	7 808	5,5	11,9	20,5	4 052 (52)
Unité 6	7 962	6,4	13,9	23,3	3 614 (45)	7 811	6,0	12,6	21,6	3 885 (50)
Unité 7	7 962	6,2	13,3	22,4	3 765 (47)	7 814	6,0	12,6	22,2	3 868 (50)
Unité 8	7 962	6,3	13,4	22,3	3 731 (47)	7 818	6,5	13,8	23,7	3 581 (46)
Unité 9	7 962	5,1	11,5	19,9	4 267 (54)	7 818	5,7	13,0	24,2	3 750 (48)
Unité 10	7 962	3,3	10,7	20,9	4 373 (55)	7 818	4,4	13,5	27,4	3 704 (47)
Unité 11	7 962	3,8	11,8	22,4	4 121 (52)	7 818	4,8	15,7	32,1	3 422 (44)
Unité 12	7 962	3,8	12,2	24,7	4 031 (51)	7 820	5,8	17,9	37,9	3 110 (40)
U1/EZS	7 962	6,7	7,5	8,4	7 941 (100)	7 823	6,7	7,5	8,5	7 803 (100)

Par contraste, les résultats obtenus avec la méthode EZS montrent que le degré de protection assuré à l'unité 1 (et à toute autre au demeurant) est relativement constant et qu'il est généralement bien moindre qu'avec la méthode MPM. Le score obtenu avec la EZS est presque de 100 %, un résultat fort médiocre. Il reste que cette méthode a été conçue pour la protection des totaux, et non pour la prévention de la prise de différences. Si l'on devait se protéger contre la prise de différences, il faudrait fixer le degré de bruit bien

plus haut pour que le niveau de protection des valeurs soit comparable à celui qu'offre la méthode MPM. Mais avec la méthode EZS, les unités autour de l'unité 10 ne seraient pas plus vulnérables en cas d'attaque ciblée répétée.

Pour étudier les rôles respectifs de K, L et M, nous avons tiré des valeurs aléatoires d'une distribution uniforme, mais en créant une valeur aberrante dans chaque cellule et en fixant x_1 à la plus grande valeur ne rendant pas la cellule sensible; pour $P = 15$, on aurait alors $x_1 = \frac{100}{15}\sum_{i\geq 3} x_i$. Nous avons appliqué la MPM en établissant M à 1, et en calculant K et L comme nous l'avons suggéré précédemment ou en les fixant à 1. Pour les données que nous avons produites, la valeur calculée de L ne s'écartait jamais de 1. Le tableau 4.4 montre que le facteur K est utile, parce que, s'il est fixé à 1, le degré de protection pour la valeur aberrante n'est pas assez élevé quand $\sigma_\varepsilon^2 = 0{,}006$.

Tableau 4.4
Protection des valeurs aberrantes dans des populations artificielles pour 1 000 cellules (quartiles de d_1)

MPM normale (K ≥ 1)				MPM avec $K = L = M = 1$			
Q1	**Médiane**	**Q3**	**Score**	**Q1**	**Médiane**	**Q3**	**Score**
11,1	12,6	14,2	472	6,7	7,5	8,6	996

5 Traitement et problèmes à résoudre

Nous avons présenté une méthode perturbatrice de protection des tableaux de données quantitatives dans un cadre de production de tableaux personnalisés. La méthode n'est pas vorace en ressources : il faut seulement suivre les unités les plus importantes de chaque cellule avec leur nombre aléatoire permanent. Nous avons démontré que cette méthode permet de protéger les unités les plus grandes contre une attaque par prise de différences.

Comme la perturbation s'applique aux valeurs les plus élevées et que les cellules sensibles sont supprimées, on a moins besoin d'appliquer un bruit par variable pour protéger les rapports. On peut calculer les rapports à l'aide de valeurs perturbées (Z). De même, on peut calculer les moyennes en employant les valeurs Z et les fréquences perturbées (arrondies, par exemple). Une autre possibilité, selon ce que préfèrent les utilisateurs, est de calculer les moyennes en divisant Z par les vraies fréquences et dégage les totaux en multipliant les moyennes perturbées par les fréquences perturbées.

Les zéros ne sont pas traités, mais il y a suppression de X (et de Z) pour les cellules sensibles et petites. Si une cellule non sensible compte moins de cinq valeurs non nulles, on ne changera pas Z en ajoutant une unité de valeur nulle. Dans ce cas particulier, les utilisateurs pourraient être en mesure de dire si une unité ajoutée à la cellule était de valeur nulle. Si les valeurs des unités x_i peuvent être négatives, les valeurs absolues $|x_i|$ les plus élevées pourraient être traitées (perturbées) dans chaque cellule. On devrait adapter les règles de dominance aux valeurs négatives (voir par exemple Tambay et Fillion 2013).

Des problèmes de divulgation par recoupements avec des données liées comme des totaux non perturbés et des tableaux de répartition continuent à se poser. Si l'organisme devait diffuser des totaux non perturbés, un pirate pourrait tenter une attaque par prise de différences en prenant un total non perturbé comme point

de départ. Il serait préférable de s'en tenir à un minimum de résultats non perturbés, en les limitant par exemple aux publications officielles de données. Les tableaux de répartition (du revenu total par tranche de revenu, par exemple) pourraient également poser des problèmes de divulgation par recoupements à cause de l'information véhiculée par les tranches. Une solution pourrait être de limiter strictement les tranches à employer dans de tels tableaux.

L'additivité des tableaux n'est pas maintenue, et les cellules supprimées viennent compliquer le recours à l'estimation par ratissage croisé pour la rétablir. Une solution possible passerait par l'imputation des valeurs des cellules supprimées, l'estimation par ratissage croisé de ces cellules, puis leur suppression. Nous pourrions au départ imputer les valeurs des cellules supprimées isolées dans une rangée ou une colonne en nous reportant à d'autres valeurs de cellule (valeur plancher fixée à 0 au besoin) et répéter l'opération en cas de nouvelle suppression isolée dans une rangée ou une colonne. D'autres méthodes pourraient servir à l'imputation des valeurs du reste des cellules supprimées.

Bibliographie

Cox, L.H., et Dandekar, R.A. (2004). A new disclosure limitation method for tabular data that preserves data accuracy and ease of use. *Proceedings of the 2002 FCSM Statistical Policy Seminar,* Document de travail 35 sur la Statistical Policy, Federal Committee on Statistical Methodology, Washington, DC.

Cox, L.H., et Sande, G. (1979). Techniques for preserving statistical confidentiality. *Proceedings of the 42nd Session of the International Statistical Institute,* Manille, Philippines.

Duncan, G., Keller-McNulty, S. et Stokes, S. (2001). *Disclosure Risk vs. Data Utility: The r-u Confidentiality Map*. Rapport technique LA-UR-01-6428, Los Alamos National Laboratory, Statistical Sciences group, Los Alamos, Nouveau-Mexique.

Evans, T., Zayatz, L. et Slanta, J. (1998). Using noise for disclosure limitation of establishment tabular data. *Journal of Official Statistics,* 14, 537-551.

Giessing, S. (2011). Post-tabular stochastic noise to protect skewed business data. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality,* Tarragone, Espagne, 26 au 28 octobre 2011.

Massell, P., et Funk, J. (2007). Recent developments in the use of noise for protecting magnitude data tables: Balancing to improve data quality and rounding that preserves protection. *Proceedings of the Research Conference of the Federal Committee on Statistical Methodology,* Arlington, Virginie.

Tambay, J.-L., et Fillion, J.-M. (2013). Strategies for processing tabular data using the G-Confid cell suppression software. *Proceedings of the Survey Research Methods Section,* American Statistical Association Joint Statistical Meetings, Montréal, 3 au 8 août 2013.

Thompson, G., Broadfoot, S. et Elazar, D. (2013). Methodology for the automatic confidentialisation of statistical outputs from remote servers at the Australian Bureau of Statistics. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality,* Ottawa, 28 au 30 octobre 2013.

Techniques d'enquête, juin 2017
Vol. 43, nº 1, pp. 47-75
Statistique Canada, nº 12-001-X au catalogue

La modélisation espace-état appliquée aux séries chronologiques de l'Enquête sur la population active des Pays-Bas : sélection de modèles et estimation de l'erreur quadratique moyenne

Oksana Bollineni-Balabay, Jan van den Brakel et Franz Palm[1]

Résumé

La modélisation de séries chronologiques structurelle est une puissante technique de réduction des variances pour les estimations sur petits domaines (EPD) reposant sur des enquêtes répétées. Le bureau central de la statistique des Pays-Bas utilise un modèle de séries chronologiques structurel pour la production des chiffres mensuels de l'Enquête sur la population active (EPA) des Pays-Bas. Cependant, ce type de modèle renferme des hyperparamètres inconnus qui doivent être estimés avant que le filtre de Kalman ne puisse être appliqué pour estimer les variables d'état du modèle. Le présent article décrit une simulation visant à étudier les propriétés des estimateurs des hyperparamètres de tels modèles. La simulation des distributions de ces estimateurs selon différentes spécifications de modèle viennent compléter les diagnostics types pour les modèles espace-état. Une autre grande question est celle de l'incertitude entourant les hyperparamètres du modèle. Pour tenir compte de cette incertitude dans les estimations d'erreurs quadratiques moyennes (EQM) de l'EPA, différents modes d'estimation sont pris en compte dans une simulation. En plus de comparer les biais EQM, cet article examine les variances et les EQM des estimateurs EQM envisagés.

Mots-clés : Bootstrap; hyperparamètre; modèles espace-état; EQM réelle; chômage.

1 Introduction

Les chiffres de la population active que produisent les organismes nationaux de statistique (ONS) sont généralement tirés d'enquêtes sur la population active. On constate un intérêt grandissant pour la production de ces indicateurs à intervalles mensuels (EUROSTAT 2015). Toutefois, la taille des échantillons est généralement trop faible, même à l'échelon national, pour pouvoir se fier aux estimateurs fondés sur le plan des théories classiques de l'échantillonnage, pour produire des chiffres mensuels suffisamment précis de la population active (Särndal, Swensson et Wretman 1992; Cochran 1977). Dans ces situations, il est cependant possible d'utiliser les techniques d'estimation sur petits domaines (EPD) pour améliorer la taille utile des échantillons des domaines en question, en empruntant les renseignements de périodes antérieurs ou d'autres domaines (voir Rao et Molina 2015; Pfeffermann 2013). Les enquêtes répétées, notamment, se prêtent à l'amélioration dans le cadre des modèles de séries chronologiques structurels (SCS) ou multiniveau.

Les modèles SCS, tout comme les modèles multiniveau, comportent normalement des hyperparamètres inconnus qui doivent être estimés. Si l'incertitude qui les accompagne (c'est ce que nous appellerons l'incertitude des hyperparamètres) n'est pas prise en compte, les erreurs quadratiques moyennes (EQM) estimées des variables explicatives de domaine seront entachées d'un biais négatif. Dans le cadre de la modélisation multiniveau, la prise en compte de cette incertitude est une pratique à la fois nécessaire et

1. Oksana Bollineni-Balabay, Statistics Netherlands, Division de la méthodologie et de la qualité, C.P. 4481, 6401CZ Heerlen, Pays-Bas. Courriel : oksana-bl@yandex.ru, obay@cbs.nl; Jan van den Brakel, Statistics Netherlands et School of Business and Economics de l'Université de Maastricht, C.P. 616, 6200 MD Maastricht, Pays-Bas; Franz Palm, Université de Maastricht, Pays-Bas.

courante; elle se fait habituellement grâce à l'utilisation de la méthode du meilleur prédicteur linéaire sans biais empirique (MPLSBE) ou d'un modèle bayésien hiérarchique (voir Rao et Molina 2015, chapitres 6, 7 et 10). Les modèles SCS ne sont pas utilisés aussi courrамment que les modèles multiniveau dans les estimations sur petits domaines. Le filtre de Kalman, habituellement appliqué en ajustement aux modèles SCS, ne tient pas compte de l'incertitude des hyperparamètres et produit donc des estimations EQM à biais négatif. Les applications qui démontrent les avantages considérables des modèles SCS par rapport au modèles types fondés sur le plan traitent les hyperparamètres estimés d'un modèle comme étant connus (voir, par exemple, Bollineni-Balabay, van den Brakel et Palm 2016a; Krieg et van den Brakel 2012; Pfeffermann et Rubin-Bleuer 1993; Tiller 1992).

Au bureau central de la statistique des Pays-Bas (Statistics Netherlands), un modèle SCS à plusieurs variables proposé par Pfeffermann (1991) est utilisé pour produire les chiffres mensuels officiels de population active aux fins de l'Enquête sur la population active (EPA) des Pays-Bas. Comme dans bien d'autres pays, l'EPA est fondé sur un plan de sondage avec renouvellement de panel et ses échantillons sont trop petits pour produire ces chiffres mensuels. Le modèle SCS appliqué aux estimations fondées sur le plan de sondage utilise les données d'échantillonnage de périodes antérieures et tient compte du biais de renouvellement de l'échantillon (BRE) et de l'autocorrélation des erreurs d'enquête. C'est ainsi qu'on obtient des estimations mensuelles suffisamment précises de la population active en chômage (voir van den Brakel et Krieg 2015). Les modèles SCS sont également utilisés pour la production des statistiques officielles du *US Bureau of Labor Statistics* (Tiller 1992). Plusieurs ONS dans le monde commencent à manifester de l'intérêt à l'égard de cette technique, notamment en Australie (Zhang et Honchar 2016) en Israël et au Royaume-Uni (ONS 2015).

Nous présentons ici une étude élargie par simulation de Monte-Carlo, où le modèle de l'EPA sert de processus de génération de données. Cette simulation nous éclaire sur le processus de sélection de modèle, avant la mise en œuvre, aux fins de la production des statistiques officielles. D'abord, l'évaluation des distributions des estimateurs des hyperparamètres selon différentes spécifications de modèles fait ressortir l'importance de conserver certains hyperparamètres dans le modèle. Les diagnostics types pour les modèles espace-état ne fournissent que des renseignements limités sur des hyperparamètres non pertinents. S'il y a surspécification, non seulement la distribution des estimations des hyperparamètres redondants risque de grandement s'éloigner de la normalité, mais les estimations des autres hyperparamètres pourraient aussi s'en trouver perturbées. Disons donc que, même si le diagnostic est satisfaisant, il serait encore avisé de simuler le modèle et d'examiner la distribution de l'estimateur de maximum de vraisemblance (MV) de ses hyperparamètres.

Un autre but de la simulation est d'évaluer dans quelle mesure l'incertitude entourant les estimations des hyperparamètres influe sur l'estimation des EQM dans les modèles SCS. L'absence de prise en compte de cette incertitude dans l'estimation EQM est acceptable seulement si les séries chronologiques disponibles sont suffisamment longues. Ce qu'on appréciera comme période suffisamment longue variera selon les applications. Le plus souvent, les séries chronologiques ininterrompues dont disposent les ONS sont relativement courtes, surtout à cause du remaniement des enquêtes. Les études spécialisées proposent plusieurs moyens de tenir compte de l'incertitude des hyperparamètres dans un modèle SCS, qu'il s'agisse de l'approximation asymptotique, du bootstrap ou d'un traitement bayésien complet (pour ce dernier cas,

voir Durbin et Koopman 2012, chapitre 13). Nous considérerons notamment dans notre exposé l'approximation asymptotique conçue par Hamilton (1986) et le bootstrap, paramétrique ou non, conçu par Pfeffermann et Tiller (2005) et Rodriguez et Ruiz (2012). Appliquées au modèle de l'EPA, ces méthodes visent à dégager la meilleure méthode d'estimation EQM dans cette application de la vie réelle. Nous montrerons aussi comment le problème de l'incertitude des hyperparamètres s'atténue au gré d'une progression de 48 à 200 mois des séries chronologiques de l'EPA.

Notre contribution sera quadruple. Premièrement, nous démontrerons comment la simulation de Monte-Carlo peut servir à contrôler la surspécification d'un modèle (hyperparamètres redondants). Deuxièmement, nous ferons voir le meilleur des modes proposés d'estimation EQM dans l'EPA et livrerons une évaluation plus réaliste de la réduction des variances dans le modèle SCS par opposition à la modélisation type fondée sur le plan. Troisièmement, notre étude de Monte-Carlo viendra infirmer ce que disent Rodriguez et Ruiz (2012) de la supériorité de leur méthode sur la méthode bootstrap de Pfeffermann et Tiller (2005) dans un modèle plus complexe. Quatrièmement, nous jetterons un éclairage, en dehors de la comparaison des biais EQM, sur la variance et les EQM des estimateurs EQM. Autant que nous sachions, la variabilité des méthodes bootstrap mentionnées n'a pas encore été étudiée.

Voici comment se structure notre propos. À la section 2, nous décrirons l'EPA et le modèle actuellement utilisé par Statistics Netherlands. À la section 3, nous passerons en revue les modes énumérés d'estimation EQM. À la section 4, nous détaillerons le cadre de simulation propre à l'EPA. Enfin, nous décrirons nos résultats à la section 5 et livrerons des observations en conclusion à la section 6.

2 Enquête sur la population active des Pays-Bas

2.1 Plan de sondage

L'Enquête sur la population active (EPA) des Pays-Bas repose sur un plan de sondage avec renouvellement de panel depuis octobre 1999. Chaque mois, on prélève un échantillon d'adresses selon un plan d'échantillonnage stratifié à deux degrés. Les strates correspondent géographiquement à des régions. Les municipalités sont les unités primaires d'échantillonnage et les adresses, les unités secondaires. Tous les ménages résidant à une adresse sont compris dans l'échantillon. Nous considérerons ici les données d'observation de l'EPA entre janvier 2001 et juin 2010, période où on a recueilli les données de la première vague par des interviews sur place assistées par ordinateur (IPAO) et par les soins d'intervieweurs visitant à domicile les ménages échantillonnés. Après un maximum de six tentatives, l'intervieweur dépose une lettre pour le répondant, lui demandant d'appeler pour prendre rendez-vous. Quand un membre d'un ménage ne peut être contacté, on permet une interview par substitution auprès des membres du même ménage. Les répondants sont interviewés à nouveau à quatre reprises, à des intervalles trimestriels. Au cours de ces quatre vagues subséquentes, les données sont recueillies par interview téléphonique assistée par ordinateur (ITAO) et les personnes répondent à un questionnaire condensé permettant d'établir tout changement de leur situation sur le marché du travail. Les interviews par substitution sont permises. Les numéros de téléphone cellulaire et les numéros confidentiels de lignes terrestres sont recueillis dès la première vague pour prévenir

toute érosion du panel. Au début de l'application du plan de sondage avec renouvellement de panel pour l'EPA, la taille brute d'échantillon était d'environ 6 200 adresses par mois en moyenne et, dans environ 65 % des cas, il s'agissait de ménages qui répondaient entièrement. Les taux de réponse des vagues qui suivent sont d'environ 90 % du taux de la vague qui précède.

L'estimateur par la régression généralisée (ERG) (Särndal et coll. 1992) est appliqué pour estimer la population active en chômage totale. Cet estimateur tient compte de la complexité du plan d'échantillonnage et exploite l'information auxiliaire disponible dans les registres pour corriger, du moins en partie, toute non-réponse sélective. Soit Y_t^j l'ERG du nombre total de chômeurs dans le mois t pour la j^{e} vague de répondants. On obtient cinq estimations semblables par mois, chacune étant directement fondée sur l'échantillon ayant accédé à l'enquête dans le mois $t-l,\ l=\{0,3,6,9,12\}$. L'estimateur ERG de ce total de population se définit ainsi :

$$Y_t^j = \sum_{k\in s} w_{k,t}\left(\sum_{i=1}^{n_{k,t}} y_{i,k,t}\right), \tag{2.1}$$

où $y_{i,k,t}$ représente les observations de l'échantillon avec 1 si la i^{e} personne dans le k^{e} ménage est en chômage et avec zéro dans les autres cas; $n_{k,t}$ est le nombre de personnes de 15 ans et plus dans le k^{e} ménage; enfin, les $w_{k,t}$ sont les poids de régression du ménage k au moment t. La méthode de Lemaître et Dufour (1987) sert à l'obtention de poids égaux pour toutes les personnes appartenant à un même ménage :

$$w_{k,t} = \frac{1}{\pi_{k,t}}\left[1+\left(\mathbf{X}_t - \sum_{k\in s}\frac{\mathbf{x}_{k,t}}{\pi_{k,t}}\right)\left(\sum_{k\in s}\frac{\mathbf{x}_{k,t}\mathbf{x}_{k,t}'}{\pi_{k,t}g_{k,t}}\right)^{-1}\frac{\mathbf{x}_{k,t}}{g_{k,t}}\right], \tag{2.2}$$

où $\pi_{k,t}$ est la probabilité d'inclusion du ménage k au moment t, $g_{k,t}$ la taille du ménage k au moment t et $\mathbf{x}_{k,t} = \sum_{i=1}^{n_{k,t}} \mathbf{x}_{i,k,t}$, $\mathbf{x}_{i,k,t}$ étant un vecteur de J dimensions avec l'information auxiliaire de modèle de pondération sur la i^{e} personne dans le k^{e} ménage au moment t. Le vecteur $\mathbf{X}_t$ contient les totaux de population des variables auxiliaires. Le modèle de pondération est défini par les variables suivantes (le nombre de catégories figure entre parenthèses) : âge(5)sexe + région(44) + sexe(2) × âge(21) + âge(5) × état matrimonial(2) + ethnicité(8), où × désigne l'interaction des variables et où âge(5)sexe est une variable en huit classes avec l'âge en cinq catégories, dont les deuxième, troisième et quatrième se détaillent pour les deux sexes.

La variance de l'estimateur ERG Y_t^j est ainsi approchée :

$$\widehat{\mathrm{Var}}\left(Y_t^j\right) = \sum_{h=1}^{H}\frac{n_{h,t}}{n_{h,t}-1}\left(\sum_{k=1}^{n_{h,t}}\left(w_{k,t}\hat{e}_{k,t}\right)^2 - \frac{1}{n_{h,t}}\left(\sum_{k=1}^{n_{h,t}} w_{k,t}\hat{e}_{k,t}\right)^2\right), \quad j=\{1,2,3,4,5\}, \tag{2.3}$$

où les résidus ERG sont $\hat{e}_{k,t} = \sum_{i=1}^{n_{k,t}}\left(y_{i,k,t} - \mathbf{x}_{i,k,t}'\hat{\boldsymbol{\beta}}_t\right)$; $n_{h,t}$ est le nombre de ménages dans la strate h (H étant le nombre total de strates). Le vecteur $\hat{\boldsymbol{\beta}}_t$ est un estimateur du type Horvitz-Thompson du coefficient de régression qui vient de la régression de la variable cible sur les variables auxiliaires de l'échantillon.

2.2 Le modèle SCS de l'EPA

Il y a deux raisons pour lesquelles Statistics Netherlands a décidé de passer à un modèle de production fondé sur les séries chronologiques, en juin 2010. La première était que l'échantillon de l'EPA était de trop petite taille pour produire des estimations mensuelles. Puisque l'échantillon dans la première vague était constitué d'environ 4 000 ménages, en moyenne, les estimations ERG de la population active en chômage présentaient un coefficient de variation d'environ 4 % à l'échelon national, ce qui était jugé trop instable pour la publication des statistiques officielles. Il faut aussi dire que les chiffres mensuels du chômage doivent être diffusés pour six domaines en fonction d'une classification sexe-âge. Les estimations fondées sur le plan pour ces domaines présentent des coefficients de variation bien plus élevés. Une autre difficulté avec l'EPA est ce que l'on appelle le biais de renouvellement de l'échantillon (BRE), c'est-à-dire les différences systématiques entre les estimations issues des différentes vagues (voir, par exemple, Bailar 1975, ou Pfeffermann 1991). Parmi les explications courantes de ce biais figurent l'érosion de l'échantillon, les effets d'échantillon longitudinal et les différences entre les questionnaires et les modes propres aux diverses vagues successives. Dans le cas de l'EPA, on présume que les estimations de la première vague sont les plus fiables et que celles des vagues subséquentes sous-estiment systématiquement les effectifs de chômeurs. Pour un examen plus détaillé, voir van den Brakel et Krieg (2009).

Les deux problèmes sont résolus avec le modèle SCS qui utilise en entrée cinq séries d'estimations ERG pour les cinq vagues considérées. Dans cette modélisation, on décompose une série observée en plusieurs composantes inobservées (tendance et composante saisonnière, par exemple). On peut employer le filtre de Kalman, en combinaison facultative avec un algorithme de lissage, pour extraire ces composantes de la série chronologique observée. C'est ainsi qu'on sépare les estimations des composantes définissant le signal du chômage de la variance inexpliquée du paramètre de population, ainsi que de la variance d'échantillonnage. Cela donne généralement des estimations ponctuelles moins instables et des erreurs-types bien moindres que celles qui caractérisent les estimations ERG. En modélisant les différences systématiques entre les cinq séries en entrée, le modèle tient aussi compte du biais de renouvellement du panel.

Dans chaque mois t, un vecteur à cinq dimensions $\mathbf{Y}_t = \left(Y_t^1\ Y_t^2\ Y_t^3\ Y_t^4\ Y_t^5\right)'$ est observé. Il contient les estimations ERG de nombre total de chômeurs pour les cinq vagues considérées. En se fondant sur Pfeffermann (1991), van den Brakel et Krieg (2009) ont conçu le modèle suivant pour les estimations ERG $\mathbf{Y}_t$:

$$\mathbf{Y}_t = \mathbf{1}_5 \xi_t + \boldsymbol{\lambda}_t + \mathbf{e}_t, \tag{2.4}$$

où $\mathbf{1}_5$ est un vecteur colonne à cinq dimensions de uns, où ξ_t est le paramètre réel de population (scalaire) qui est inconnu, où $\boldsymbol{\lambda}_t$ est un vecteur contenant des variables d'état pour le BRE et enfin où $\mathbf{e}_t$ est un vecteur des erreurs d'enquête en corrélation avec les erreurs correspondantes des vagues antérieures (nous présentons cette structure plus loin). Pour le paramètre réel de population, nous posons que $\xi_t = L_t + \gamma_t + \varepsilon_t$, soit la somme d'une tendance stochastique L_t, d'une composante saisonnière stochastique γ_t, et d'une composante irrégulière $\varepsilon_t \overset{\text{iid}}{\sim} N\left(0, \sigma_\varepsilon^2\right)$.

Dans le cas de la tendance stochastique L_t, nous posons ce qu'on appelle le modèle de lissage de la tendance :

$$\begin{aligned} L_t &= L_{t-1} + R_{t-1}, \\ R_t &= R_{t-1} + \eta_{R,t}, \end{aligned}$$

où L_t et R_t correspondent au niveau et à la pente du paramètre réel de population; le terme de perturbation de la pente présente la distribution suivante : $\eta_{R,t} \overset{\text{iid}}{\sim} N\left(0, \sigma_R^2\right)$.

Pour la composante saisonnière γ_t, nous posons le modèle trigonométrique :

$$\gamma_t = \sum_{l=1}^{6} \gamma_{t,l},$$

où chacune de ces six harmoniques suit le processus suivant :

$$\begin{aligned} \gamma_{t,l} &= \cos\left(h_l\right)\gamma_{t-1,l} + \sin\left(h_l\right)\gamma^*_{t-1,l} + \omega_{t,l}, \\ \gamma^*_{t,l} &= -\sin\left(h_l\right)\gamma_{t-1,l} + \cos\left(h_l\right)\gamma^*_{t-1,l} + \omega^*_{t,l}, \end{aligned}$$

$h_l = \frac{\pi l}{6}$ étant la l^{e} fréquence saisonnière, $l = \{1, \ldots 6\}$. Nous posons que les termes stochastiques $\omega_{t,l}$ et $\omega^*_{t,l}$ à espérance nulle sont normalement et indépendamment distribués et présentent la même variance dans et entre tous les harmoniques :

$$\begin{aligned} &\text{Cov}\left(\omega_{t,l}, \omega_{t',l'}\right) = \text{Cov}\left(\omega^*_{t,l}, \omega^*_{t',l'}\right) = \begin{cases} \sigma^2_\omega & \text{si } l = l' \text{ et } t = t', \\ 0 & \text{si } l \neq l' \text{ ou } t \neq t', \end{cases} \\ &\text{Cov}\left(\omega_{t,l}, \omega^*_{t,l}\right) = 0 \text{ pour tous les } l \text{ et } t. \end{aligned}$$

La deuxième composante en (2.4) est le biais de renouvellement (BRE). Nous posons que la première vague est sans biais, ainsi que l'expliquent van den Brakel et Krieg (2009). Les BRE des vagues qui suivent sont fonction du temps et se modélisent comme des processus à marche aléatoire. On justifie le tout en disant que les procédures de terrain subissent de fréquents changements et que, par ailleurs, les taux de réponse évoluent progressivement dans le temps, ce qui rend le BRE tributaire du temps, comme l'illustrent van den Brakel et Krieg (2015) (voir la figure 4.3). Le vecteur BRE des cinq vagues peut s'écrire ainsi : $\boldsymbol{\lambda}_t = \left(0 \ \lambda_t^2 \ \lambda_t^3 \ \lambda_t^4 \ \lambda_t^5\right)'$, avec :

$$\lambda_t^j = \lambda_{t-1}^j + \eta_{\lambda,t}^j, \quad j = \{2, 3, 4, 5\}.$$

Nous posons que les perturbations BRE ne sont pas corrélées entre les vagues et que leur distribution est normale, c'est-à-dire $\eta_{\lambda,t}^j \overset{\text{iid}}{\sim} \left(0, \sigma_\lambda^2\right)$, avec égalité des variances dans les quatre vagues.

La dernière composante en (2.4) est celle des erreurs d'enquête pour les cinq estimations ERG, c'est-à-dire $\mathbf{e}_t = \left(e_t^1 \ e_t^2 \ e_t^3 \ e_t^4 \ e_t^5\right)'$. Pour tenir compte de l'hétérogénéité des erreurs d'échantillonnage causée par les variations temporelles de taille d'échantillon, nous modélisons ces erreurs en proportion des erreurs-types fondées sur le plan, d'après le modèle d'erreur de mesure proposé par Binder et Dick (1990), c'est-à-dire $e_t^j = \tilde{e}_t^j z_t^j$, où $z_t^j = \sqrt{\widehat{\text{Var}}\left(Y_t^j\right)}$ et $\tilde{e}_t^j$ sont des erreurs d'échantillonnage réduites ou normalisées en

fonction d'un processus stationnaire que nous définirons plus loin. Ici, les $\widehat{\mathrm{Var}}(Y_t^j)$ sont les estimations des variances, fondées sur le plan, qui sont tirées des microdonnées en (2.3). Ils sont traités comme variances d'échantillonnage connues a priori dans le modèle SCS.

Comme l'échantillon de la première vague n'est pas en chevauchement avec les échantillons observés par le passé, les $\tilde{e}_t^t$ peuvent se modéliser comme du bruit blanc avec $E(\tilde{e}_t^1) = 0$ et $\mathrm{Var}(\tilde{e}_t^1) = \sigma_{v_1}^2$. La variance des erreurs d'échantillonnage e_t^t sera égale à la variance des estimations ERG si l'estimation de maximum de vraisemblance des $\sigma_{v_1}^2$ est à peu près égale à l'unité.

Dans les vagues qui suivent, les erreurs d'enquête sont en corrélation avec les erreurs d'enquête des vagues antérieures. Nous estimons le coefficient d'autocorrélation à partir des données d'enquête par la méthode que proposent Pfeffermann, Feder et Signorelli (1998). La structure d'autocorrélation est mise en modélisation autorégressive AR(1) et le coefficient d'autocorrélation s'obtient par les équations de Yule-Walker (van den Brakel et Krieg 2009):

$$\tilde{e}_t^j = \rho \tilde{e}_{t-3}^{j-1} + \nu_t^j,\ \nu_t^j \overset{\text{iid}}{\sim} N\left(0, \sigma_{v_j}^2\right), \quad j = \{2, 3, 4, 5\}.$$

Nous posons que le coefficient d'autocorrélation du premier ordre est commun aux quatre vagues. Son estimation fait fonction d'information a priori dans le modèle. Comme $\tilde{e}_t^j$ représente un processus AR(1), $\mathrm{Var}(\tilde{e}_t^j) = \sigma_{v_j}^2 \,/ (1 - \rho^2)$. La variance de l'erreur d'échantillonnage e_t^j correspond approximativement à $\widehat{\mathrm{Var}}(Y_t^j)$ si l'estimation de maximum de vraisemblance des $\sigma_{v_j}^2$ est à peu près égale à $(1 - \rho^2)$. Nous posons cinq hyperparamètres différents $\sigma_{v_j}^2$, $j = \{1, 2, 3, 4, 5\}$, pour les erreurs d'échantillonnage comme composantes des cinq vagues.

Nous regroupons les variances de perturbation avec le paramètre d'autocorrélation ρ dans un vecteur d'hyperparamètres appelé $\boldsymbol{\theta} = \left(\sigma_R^2\ \sigma_\omega^2\ \sigma_\varepsilon^2\ \sigma_\lambda^2\ \sigma_{v_1}^2\ \sigma_{v_2}^2\ \sigma_{v_3}^2\ \sigma_{v_4}^2\ \sigma_{v_5}^2\ \rho\right)'$, le vecteur contenant seulement les variances de perturbation est $\boldsymbol{\theta}_\sigma = \left(\sigma_R^2\ \sigma_\omega^2\ \sigma_\varepsilon^2\ \sigma_\lambda^2\ \sigma_{v_1}^2\ \sigma_{v_2}^2\ \sigma_{v_3}^2\ \sigma_{v_4}^2\ \sigma_{v_5}^2\right)'$. Pour éviter les estimations négatives, nous estimons à l'échelle logarithmique les hyperparamètres des variances de perturbation dans $\boldsymbol{\theta}_\sigma$. Nous employons la méthode du quasi-maximum de vraisemblance (voir, par exemple, Harvey 1989) où on traite les estimations $\hat{\rho}$ comme étant connues. Dans cette étude, l'analyse numérique se fait avec OxMetrics 5 (Doornik 2007) en combinaison avec le progiciel *SsfPack 3.0* (Koopman, Shephard et Doornik 2008).

3 Modes d'estimation EQM

D'ordinaire, on ajuste les modèles structurels linéaires de séries chronologiques ayant des composantes inobservées en appliquant le filtre de Kalman à l'espace-état une fois formé à partir de ces composantes. On peut voir dans Bollineni-Balabay, van den Brakel et Palm (2016b) quelle est la représentation en espace-état du modèle SCS pour l'EPA. Le vecteur d'état $\boldsymbol{\alpha}_t$ contient les variables d'état définies à la section précédente, c'est-à-dire la tendance, la pente, les harmoniques saisonnières, le BRE, le bruit blanc de population et les erreurs d'enquête. Nous initialisons toutes les variables d'état non stationnaires en prenant une distribution antérieure diffuse (à moyenne nulle et à très grande variance). Les cinq composantes des

erreurs d'enquête $\tilde{e}_t^j,\ j = \{1, 2, 3, 4, 5\}$ et le bruit blanc de population ε_t sont des variables d'état stationnaires initialisées avec des zéros. Nous tenons la variance initiale des erreurs d'échantillonnage de la première vague pour égale à l'unité et nous considérons que la variance des autres vagues correspond à $(1-\rho^2)$. On pourrait même prendre une petite valeur pour la variance initiale de ε_t.

On extrait habituellement des estimations filtrées du vecteur d'état $\boldsymbol{\alpha}_t$ et de sa matrice des covariances $\mathbf{P}_{t|t}$ à l'aide du filtre de Kalman (voir Harvey 1989). Ainsi, $\mathbf{P}_{t|t}$ contient les EQM extraites par le filtre conditionnellement à l'information obtenue jusqu'au moment t inclusivement :

$$\mathbf{P}_{t|t} = \mathrm{E}_t\left[\left(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t\right)\left(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t\right)'\right], \tag{3.1}$$

où nous posons que $\boldsymbol{\theta}$ est la valeur réelle des hyperparamètres et où l'espérance se prend sur la codistribution du vecteur d'état et des valeurs Y au moment t. Dans la pratique, le vecteur réel des hyperparamètres est remplacé par son estimation $\hat{\boldsymbol{\theta}}$ dans les récursions par filtre de Kalman. Dans ce cas, l'EQM en (3.1) n'est plus l'EQM réelle. On la qualifie de « naïve », puisqu'elle ne tient pas compte de l'incertitude autour des estimations $\hat{\boldsymbol{\theta}}$. L'EQM réelle devient ainsi :

$$\mathbf{EQM}_{t|t} = \mathrm{E}_t\left[\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t\right)\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t\right)'\right],$$

ce qui représente une valeur supérieure à la valeur EQM en (3.1) et peut se décomposer comme la somme de l'incertitude du filtre et de l'incertitude des paramètres dans une condition de normalité des termes d'erreur :

$$\mathbf{EQM}_{t|t} = \mathrm{E}_t\left[\left(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t\right)\left(\hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_t\right)'\right] + \mathrm{E}_t\left[\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta})\right)\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \hat{\boldsymbol{\alpha}}_{t|t}(\boldsymbol{\theta})\right)'\right]. \tag{3.2}$$

Le premier terme, l'incertitude du filtre, est ce qui est estimé par les estimations EQM $\mathbf{P}_{t|t}$ par le filtre de Kalman. Il faut aller plus loin pour estimer le deuxième terme, l'incertitude des paramètres. Les études spécialisées consacrées à l'estimation EQM proposent deux grandes méthodes, à savoir l'approximation asymptotique et le bootstrap. Le bootstrap peut être paramétrique ou non paramétrique. Quelques observations s'imposent ici au sujet de ces méthodes dans le contexte du modèle SCS appliqué à l'EPA.

Dans le cas du bootstrap paramétrique, les perturbations d'état, $\boldsymbol{\eta}_t$, disons, sont tirées de coestimations de densité normale conditionnelle à plusieurs variables $\boldsymbol{\eta}_t \overset{\text{iid}}{\sim} \mathrm{MN}\left(\mathbf{0}, \hat{\boldsymbol{\Omega}}\right)$, $\hat{\boldsymbol{\Omega}}$ étant évalué à l'estimation $\hat{\boldsymbol{\theta}}$ des hyperparamètres. Ces perturbations servent dans les récursions d'état par filtre de Kalman à produire les variables d'état. Par ailleurs, le bootstrap non paramétrique a pour avantage de ne dépendre d'aucune hypothèse particulière au sujet de cette codistribution. Si dans le bootstrap paramétrique les perturbations d'état viennent de l'estimation de leur distribution, dans le bootstrap non paramétrique il y a rééchantillonnage avec remise dans un nouvel ensemble normalisé en fonction des estimations initiales des hyperparamètres. Les nouveaux ensembles normalisés qui sont rééchantillonnés servent en outre à produire des séries bootstrap $\{\mathbf{Y}_1^b, \ldots, \mathbf{Y}_T^b\}$ par ce qu'on appelle la forme d'innovation du filtre de Kalman (voir les détails dans Harvey 1989, ou Bollineni-Balabay et coll. 2016b). Dans le modèle de l'EPA, les 13 premiers

points temporels d'un nouvel ensemble normalisé ne font pas l'objet d'un rééchantillonnage et ils constituent ce qu'on appelle l'échantillon diffus (c'est le temps dont on a besoin pour construire une distribution appropriée pour les variables d'état non stationnaires; voir dans Koopman (1997) l'initialisation de telles variables).

Si un modèle SCS compte des composantes non stationnaires comme dans le modèle de l'EPA, les séries produites divergeront probablement de l'ensemble de données au départ de l'application du bootstrap, c'est-à-dire de $\{\mathbf{Y}_1,\ldots,\mathbf{Y}_T\}$. Il nous faut donc recourir à une procédure spéciale pour que les échantillons bootstrap soient mis en correspondance avec la configuration de l'ensemble de données initial, ce qu'on peut faire à l'aide d'un algorithme de lissage par simulation qui a été conçu par Durbin et Koopman (2002). On trouvera les détails techniques sur cette application dans Koopman et coll. (2008), chapitre 8.4.2. On n'a pas à prévoir de corrections pour les erreurs d'enquête issues comme nous l'avons décrit des récursions inconditionnelles d'état par le bootstrap paramétrique ou non paramétrique, puisqu'il s'agit d'un bruit (en autocorrélation).

Dans les sections qui suivent, nous présenterons brièvement la méthode asymptotique, ainsi que les applications bootstrap récentes de Rodriguez et Ruiz (2012) (bootstrap RR) et de Pfeffermann et Tiller (2005) (bootstrap PT).

3.1 Application bootstrap de Rodriguez et Ruiz

Rodriguez et Ruiz (2012) ont conçu leur méthode bootstrap d'estimation EQM conditionnelle aux données, ce qui veut dire qu'on applique en plus les hyperparamètres bootstrap à l'ensemble de données initial pour obtenir des estimations bootstrap des variables d'état. Il peut s'agir d'un bootstrap paramétrique ou non avec les étapes suivantes :

1. On estime le modèle et obtient les estimations $\hat{\boldsymbol{\theta}}$ des hyperparamètres.
2. On produit un échantillon bootstrap $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$ à l'aide de $\hat{\boldsymbol{\theta}}$ par bootstrap paramétrique ou non (voir l'introduction de cette section). Si le modèle est non stationnaire, on se doit de corriger l'échantillon bootstrap par simulation de lissage.
3. On se sert de l'ensemble bootstrap $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$ pour obtenir tant les estimations paramétriques d'autocorrélation des erreurs d'enquête $\hat{\rho}^b$ que les estimations bootstrap de maximum de vraisemblance $\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^b$. On applique ensuite le filtre de Kalman à la série initiale $\{\mathbf{Y}_1,\ldots,\mathbf{Y}_T\}$ et aux $\hat{\boldsymbol{\theta}}^b$, nouvellement estimés, ce qui donne $\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$ et $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$.
4. On reprend B fois les étapes 2 et 3, puis procède à l'estimation EQM de la manière suivante :

$$\widehat{\mathbf{EQM}}_{t|t}^{\mathrm{RR}} = \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)+\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\overline{\boldsymbol{\alpha}}_{t|t}\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\overline{\boldsymbol{\alpha}}_{t|t}\right]', \tag{3.3}$$

où $\overline{\boldsymbol{\alpha}}_{t|t} = \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$.

L'équation (3.3) est applicable aux estimations EQM par bootstrap paramétrique et non paramétrique (nous emploierons dans ce cas les abréviations $\mathrm{EQM}^{\mathrm{RR1}}$ et $\mathrm{EQM}^{\mathrm{RR2}}$ dans la suite du texte).

3.2 Application bootstrap de Pfeffermann et Tiller

La méthode bootstrap conçue par Pfeffermann et Tiller (2005) est un bootstrap inconditionnel, c'est-à-dire que variables d'état bootstrap sont dérivées de l'ensemble de données bootstrap $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$, et non de l'ensemble de données initial $\{\mathbf{Y}_1,\ldots,\mathbf{Y}_T\}$ comme dans Rodriguez et Ruiz (2012). Pfeffermann et Tiller (2005) ont démontré que leur méthode approche l'EQM réelle jusqu'à un ordre de $O(1/T^2)$ (Pfeffermann et Tiller (2005), annexe C) :

$$\widehat{\mathbf{EQM}}_{t|t}^{\mathrm{PT}} = 2\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right) + \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)\right]'. \tag{3.4}$$

L'équation (3.4) est applicable aux estimateurs EQM par bootstrap paramétrique ou non (nous emploierons dans ce cas les abréviations $\mathrm{EQM}^{\mathrm{PT1}}$ et $\mathrm{EQM}^{\mathrm{PT2}}$ dans la suite du texte). Le calcul EQM en (3.4) exige deux exécutions du filtre de Kalman pour chaque série bootstrap. À la première exécution, $\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right)$ est estimé à partir de l'ensemble bootstrap $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$ et des paramètres bootstrap $\hat{\boldsymbol{\theta}}^b$. Dans cette exécution, on peut aussi obtenir $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$ par $\hat{\boldsymbol{\theta}}^b$, puisque la matrice $\mathbf{P}_{t|t}$ ne dépend pas des données. Il faut appliquer le filtre de Kalman une deuxième fois pour produire les estimations d'état $\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)$ en fonction de $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$ et des estimations $\hat{\boldsymbol{\theta}}$ tirées de l'ensemble initial. La procédure se résume ainsi :

1. Estimer le modèle à l'aide de l'ensemble de données initial et obtenir les estimations $\hat{\boldsymbol{\theta}}$ du vecteur des hyperparamètres. Garder les estimations EQM « naïves » $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right)$ pour une utilisation future en (3.4).
2. Utiliser le bootstrap paramétrique ou non pour produire un échantillon bootstrap $\{\mathbf{Y}_1^b,\ldots,\mathbf{Y}_T^b\}$. Apporter la correction par simulation de lissage si le modèle est non stationnaire.
3. Établir les estimations bootstrap $\hat{\boldsymbol{\theta}}^b$ des hyperparamètres à partir de l'ensemble bootstrap nouvellement produit. Appliquer le filtre de Kalman une première fois pour obtenir $\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right)$ et $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$, et une autre fois pour dégager $\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)$, comme décrit en (3.4).
4. Répéter B fois les étapes 2 et 3, puis procéder à l'estimation EQM en (3.4).

Pfeffermann et Tiller (2005) signalent que, dans le cas du bootstrap paramétrique, il est possible d'éviter le deuxième filtre de Kalman, parce que le vecteur d'état réel est produit (et donc connu) pour chaque série bootstrap. On peut donc remplacer les estimations d'état $\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)$ en (3.4) par le vecteur réel $\boldsymbol{\alpha}_t^b$ pour obtenir l'estimateur EQM suivant :

$$\widehat{\mathbf{EQM}}_{t|t}^{\mathrm{PT1}} = \mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right) - \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right) + \frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \boldsymbol{\alpha}_t^b\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \boldsymbol{\alpha}_t^b\right]'. \tag{3.5}$$

Il y a un seul $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right)$ du côté droit de (3.5), puisque le nouveau terme $\mathrm{E}_B\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \boldsymbol{\alpha}_t^b\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \boldsymbol{\alpha}_t^b\right]'$, qui correspond au dernier terme du côté droit de (3.5), peut lui-même se décomposer comme en (3.2) en une mesure de l'incertitude des paramètres $\mathrm{E}_B\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}^b\right) - \hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right)\right]'$ et de l'incertitude du filtre $\mathbf{P}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right) = \mathrm{E}\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t^b\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t^b\right]'$, $\hat{\boldsymbol{\theta}}$ étant le vecteur réel des paramètres par lequel on produit les variables d'état bootstrap $\boldsymbol{\alpha}_t^b$. Toutefois, on aura peut-être à prévoir beaucoup plus d'itérations bootstrap pour le terme moyen bootstrap $\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t^b\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}^b\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\alpha}_t^b\right]'$ remplaçant $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right)$ si on

veut qu'il y ait convergence. Ajoutons que cette méthode simplifiée peut créer plus de biais si l'hypothèse de normalité n'est pas respectée au sujet des termes d'erreur du modèle. Dans ce cas, la décomposition du terme $\mathrm{E}_B\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\boldsymbol{\alpha}^b_t\right]\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\boldsymbol{\alpha}^b_t\right]'$ comme en (3.2) laissera aussi un terme croisé non nul : $\mathrm{E}\left\{\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)-\boldsymbol{\alpha}^b_t\right]\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)\right]\right\}$. Dans cette application, les moyennes bootstrap à terme croisé non nul se sont révélées négligeables, mais la moyenne bootstrap $\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)-\boldsymbol{\alpha}^b_t\right]\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)-\boldsymbol{\alpha}^b_t\right]'$ s'éloignait largement (dans les deux sens) du terme qu'elle était censée remplacer, ce qu'expliquerait le fait que l'EQM réelle par filtre de Kalman en (3.1) puisse être tirée de séries en simulation si, dans sa distribution, le vecteur d'état est suffisamment dispersé. Quand on met des modèles non stationnaires en bootstrap, les séries bootstrap suivent forcément la configuration de la série initiale sous-jacente, comme nous l'avons mentionné dans la description de l'algorithme de lissage par simulation. Il se peut donc que le terme $\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)-\boldsymbol{\alpha}^b_t\right]\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)-\boldsymbol{\alpha}^b_t\right]'$, qui remplace $\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}\right)$ en (3.5), n'en soit pas suffisamment proche. C'est pourquoi le bootstrap paramétrique (PT1) ou non (PT2) dans cette application dépend de l'estimateur en (3.4).

Disons quelques mots du rôle de la simulation de lissage de Durbin et Koopman (2002) dont nous avons fait mention à la fin de l'introduction à la présente section. Nous avons proposé de l'employer à l'étape de la production des séries boostrap, sans quoi la distribution bootstrap des hyperparamètres tirée de séries non corrigées pour un modèle non stationnaire pourrait être fort différente de ce qu'elle devrait être pour une réalisation particulière des données dont nous disposons. Dans le cas de l'EPA du moins, les distributions bootstrap des hyperparamètres étaient bien plus diffuses sans la simulation de lissage qu'avec celle-ci. De plus, les distributions bootstrap des hyperparamètres qui viennent de séries non corrigées dans l'EPA sont centrées sur des valeurs bien supérieures aux valeurs des hyperparamètres qui ont servi à produire les séries. Le résultat est une moyenne bootstrap extrêmement élevée $\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)$ (par rapport à $\mathbf{P}_{t|t}(\hat{\boldsymbol{\theta}})$) et, par la suite, des estimations EQM même inférieures aux estimations naïves. Il faut aussi dire que le terme $\frac{1}{\mathrm{B}}\sum_{b=1}^{B}\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)\right]\left[\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}^b\right)-\hat{\boldsymbol{\alpha}}^b_{t|t}\left(\hat{\boldsymbol{\theta}}\right)\right]'$ devient très instable dans le temps et prend des proportions excessives quand il n'y a pas de simulation de lissage, ce qui ne compense pas le biais négatif en (3.4) sans la simulation de lissage.

3.3 Approximation asymptotique

Hamilton (1986) a conçu une approximation asymptotique (AA) de l'EQM réelle à l'équation (3.2). Cette approximation peut s'exprimer comme une espérance sur la codistribution asymptotique des hyperparamètres $\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)$, celle-ci étant conditionnelle à l'ensemble de données initial $\mathbf{Y}\equiv\{\mathbf{Y}_1,\ldots,\mathbf{Y}_T\}$. Dans la présente application, la partie du vecteur des hyperparamètres qui est estimée par la méthode du maximum de vraisemblance $\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}\right)$ dépend de la valeur estimée du paramètre autorégressif $\hat{\rho}$. Ainsi, la codistribution asymptotique de l'estimateur des hyperparamètres est de la forme suivante : $\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)=\pi\left(\hat{\rho}\middle|\mathbf{Y}\right)\pi\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}\middle|\hat{\rho},\mathbf{Y}\right)$. L'EQM est ainsi approchée :

$$\mathbf{EQM}_{t|t}=\mathrm{E}_{\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)}\left[\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}},\mathbf{Y}\right)\right]+\mathrm{E}_{\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)}\left[\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}},\mathbf{Y}\right)-\hat{\boldsymbol{\alpha}}_{t|t}\left(\mathbf{Y}\right)\right)\left(\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}},\mathbf{Y}\right)-\hat{\boldsymbol{\alpha}}_{t|t}\left(\mathbf{Y}\right)\right)'\right], \qquad (3.6)$$

où $\mathrm{E}_{\pi(\hat{\boldsymbol{\theta}}|\mathbf{Y})}$ est une espérance prise sur la codistribution asymptotique de l'estimateur des hyperparamètres $\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)$, et où les $\hat{\boldsymbol{\alpha}}_{t|t}(\mathbf{Y})$ sont les estimations du vecteur d'état quand les hyperparamètres ne sont pas connus $\mathrm{E}_{\pi(\hat{\boldsymbol{\theta}}|\mathbf{Y})}\left[\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}},\mathbf{Y}\right)\right]$.

Dans ce cas, nous choisissons la distribution $N(\hat{\rho},\mathrm{Var}(\hat{\rho}))$ comme la distribution asymptotique $\pi(\hat{\rho}|\mathbf{Y})$ des $\hat{\rho}$, d'où sont tirées les réalisations aléatoires $\hat{\rho}$. En général, la distribution d'échantillonnage du coefficient de corrélation revêt une forme complexe, mais elle peut fort bien être approchée par une distribution normale; tel était le cas dans cette application (la distribution normale était un très bon ajustement de la distribution en simulation et de la distribution bootstrap de $\hat{\rho}$). Si on prend l'équation (3) dans Bartlett (1946) et qu'on considère que le coefficient autorégressif dans un processus AR(1) est égal à la corrélation pour le décalage 1, l'estimateur de variance de $\hat{\rho}$ devient $\mathrm{Var}(\hat{\rho}) \approx (1-\hat{\rho}^2)/T$. Dans le cas de l'EPA où $\hat{\rho} = 0{,}208$, cela veut dire que $\widehat{\mathrm{Var}}(\hat{\rho}) \approx 0{,}96(1/T)$. Comme l'erreur-type des $\hat{\rho}$ sert à tirer des réalisations de la distribution asymptotique et que l'extraction de la racine carrée est une fonction concave, l'écart-type de l'échantillon serait une sous-estimation. En tirant donc $\hat{\rho}$ réalisations au moyen de $1/\sqrt{T}$ comme écart-type de la distribution asymptotique, on ferait un choix raisonnable.

On obtient de la manière suivante un échantillon de B réalisations de la distribution asymptotique des hyperparamètres. Après avoir tiré une valeur, $\hat{\rho}^a$ disons, de $\pi(\hat{\rho}|\mathbf{Y})$, nous réestimons les autres hyperparamètres de l'ensemble de données initial pour obtenir $\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^{\mathrm{MV}}\left|\hat{\rho}^a,\mathbf{Y}\right.$ et la matrice d'information $\hat{\mathbf{I}}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^{\mathrm{MV}}\left|\hat{\rho}^a,\mathbf{Y}\right.\right)$. Finalement, nous tirons une réalisation $\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^a$ de la distribution $\mathrm{MN}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^{\mathrm{MV}},\hat{\mathbf{I}}^{-1}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^{\mathrm{MV}}\left|\hat{\rho}^a,\mathbf{Y}\right.\right)\right)$. Nous appliquons à nouveau le filtre de Kalman avec les réalisations $\hat{\rho}^a$ et $\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^a$ pour obtenir les estimations d'état $\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^a,\mathbf{Y}\right)$ et leurs EQM $\hat{\mathbf{P}}_{t|t}\left(\hat{\boldsymbol{\theta}}^a\right)$. La procédure se répète jusqu'à ce que B itérations $\hat{\boldsymbol{\theta}}^a$ aient été effectuées, après quoi nous dégageons (3.6) en prenant la moyenne des quantités nécessaires sur B itérations. Si tous les hyperparamètres du modèle sont estimés par la méthode du maximum de vraisemblance, B itérations peuvent se faire directement à partir de $\mathrm{MN}\left(\hat{\boldsymbol{\theta}}^{\mathrm{MV}},\hat{\mathbf{I}}^{-1}\left(\hat{\boldsymbol{\theta}}^{\mathrm{MV}}\right)\right)$.

On peut approcher le premier terme en (3.6) par la valeur moyenne de la variance $\mathbf{P}_{t|t}$ par filtre de Kalman sur B réalisations du vecteur des hyperparamètres. Le deuxième terme peut être approché par la variance des estimations du vecteur d'état sur ces mêmes B itérations. Une approximation asymptotique des EQM pourrait se dégager de la manière suivante :

$$\widehat{\mathbf{EQM}}_{t|t}^{\mathrm{AA}} = \frac{1}{\mathrm{B}}\sum_{a=1}^{B}\mathbf{P}_{t|t}\left(\hat{\boldsymbol{\theta}}^a\right) + \frac{1}{\mathrm{B}}\sum_{a=1}^{B}\left[\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^a,\mathbf{Y}\right)-\overline{\hat{\boldsymbol{\alpha}}}_{t|t}\right]\left[\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^a,\mathbf{Y}\right)-\overline{\hat{\boldsymbol{\alpha}}}_{t|t}\right]', \tag{3.7}$$

où $\hat{\boldsymbol{\theta}}^a$ est le résultat du a^{e} tirage à partir de la distribution asymptotique $\pi\left(\hat{\boldsymbol{\theta}}\middle|\mathbf{Y}\right)$. Comme le propose Hamilton (1986), la moyenne d'échantillon $\overline{\hat{\boldsymbol{\alpha}}}_{t|t} = \frac{1}{\mathrm{B}}\sum_{a=1}^{B}\hat{\boldsymbol{\alpha}}_{t|t}\left(\hat{\boldsymbol{\theta}}^a,\mathbf{Y}\right)$ peut remplacer $\hat{\boldsymbol{\alpha}}_{t|t}(\mathbf{Y})$ en (3.6). Cet auteur ajoute qu'une telle décomposition de l'incertitude du total en une incertitude du filtre et une incertitude des paramètres ressemble à la décomposition bien connue $\mathrm{var}(X) = E[\mathrm{var}(X|Y)] + \mathrm{var}[E(X|Y)]$. Manifestement, cet estimateur EQM repose entièrement sur l'hypothèse d'une normalité asymptotique de l'estimateur du vecteur des hyperparamètres. De plus, cette application produit habituellement des biais significatifs si les séries ne sont pas d'une longueur suffisante, auquel cas la distribution asymptotique normale qui est posée ne pourrait approcher la distribution finie (ordinairement asymétrique) des estimations de maximum de vraisemblance.

Un autre problème est susceptible de se poser avec le traitement asymptotique si on estime que les hyperparamètres sont proches de zéro, ce qui peut advenir des estimations du modèle au départ ou pendant l'application de la procédure même à cause de certaines réalisations extrêmes de $\hat{\rho}$. Dans ce cas, la variance asymptotique de ces hyperparamètres sera très élevée, ce qui viendra gonfler les estimations EQM du signal et de ses composantes inobservées. Il pourrait en résulter un défaut d'inversion de la matrice d'information pour le vecteur des hyperparamètres.

4 L'EPA et son cadre précis de simulation

Nous allons examiner le rendement des cinq méthodes d'estimation EQM par rapport à des séries de la longueur initiale de cette enquête (114 points mensuels de 2001(1) à 2010(6)) et à des séries soit plus courtes de 48 et 80 mois soit plus longues de 200 mois. Pour chacune de ces durées, nous montons une expérience de Monte-Carlo où des séries multiples (1 000) font l'objet d'une simulation en fonction du modèle EPA du nombre de chômeurs. Nous estimons les EQM de chacune de ces séries en prenant $B = 300$ séries bootstrap. Dans le cas de l'approximation asymptotique toutefois, il nous a fallu prévoir au moins $B = 500$ itérations. Nous avons jugé que ce nombre était suffisant pour qu'il y ait convergence des EQM approchées. Nous comparons les EQM issues des cinq méthodes et mises en moyenne sur 1 000 simulations aux moyennes EQM produites par un filtre de Kalman « naïf ». Dans ce cas, au moins 10 000 simulations sont nécessaires pour que les estimations EQM convergent sur une certaine moyenne.

Voici comment nous obtenons paramétriquement les séries artificielles $\mathbf{Y}_t^s$ mentionnées pour des simulations $s = 1,\ldots, 1\,000$ (ou 10 000) : d'abord, nous établissons les estimations $\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}$ de maximum de vraisemblance des hyperparamètres par ajustement du modèle SCS aux séries initiales; ensuite, nous tirons aléatoirement des perturbations d'état (on se rappellera que les erreurs d'enquête sont aussi modélisées comme variables d'état) de leur codistribution normale $N\left(\mathbf{0}, \boldsymbol{\Omega}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}\right)\right)$, et produisons les séries par récursion de filtre de Kalman. Comme le système est non stationnaire, les séries produites $\mathbf{Y}_t^s$ peuvent donner des nombres de chômeurs négatifs ou excessivement élevés. Pour éviter tout nombre démesuré de séries aux valeurs négatives, nous appliquons la récursion des variables d'état à partir des estimations lissées des états à un des points les plus hauts des séries observées. De plus, nous écartons les 30 premiers points temporels pour empêcher que les séries ne commencent au même point temporel. Dans l'hypothèse que le chômage aux Pays-Bas ne sera pas de plus de 15 % de toute la population active, nous limitons l'ensemble de données en simulation aux séries dont les valeurs vont de zéro à 1 million de chômeurs (il s'agit d'environ 15 % de la population active des Pays-Bas en 2010), les autres séries étant éliminées. Si nous gardons nos séries artificielles sous la borne supérieure, c'est aussi pour ne pas extrapoler en dehors de la plage initiale des données dans la simulation des erreurs-types z_t^j fondées sur le plan.

Toute série d'estimations ponctuelles ERG en simulation exige sa propre série d'estimations des erreurs-types fondées sur le plan en simulation z_t^j. Les estimations initiales connues des erreurs-types fondées sur le plan $\sqrt{\widehat{\mathrm{Var}}\left(Y_t^j\right)}$ ne conviendraient pas à cette simulation, parce que la variance de l'erreur d'échantillonnage est proportionnelle à l'estimation ponctuelle correspondante. La fonction de variance suivante nous a permis de produire des variances fondées sur le plan pour la série simulée d'estimations ponctuelles (voir les détails à l'annexe B dans Bollineni-Balabay et coll. 2016b) :

$$\ln\left[\widehat{\mathrm{Var}}\left(Y_t^1\right)\right] = \ln\left[\left(z_t^1\right)^2\right] = c + \beta_1 \ln\left(l_t^1\right) + \varepsilon_t^1,\ \varepsilon_t^t \sim N\left(0, \left(\sigma_\varepsilon^1\right)^2\right);$$
$$\ln\left[\widehat{\mathrm{Var}}\left(Y_t^j\right)\right] = \ln\left[\left(z_t^j\right)^2\right] = \psi_j \ln\left[\left(z_{t-3}^{j-1}\right)^2\right] + \beta_j \ln\left(l_t^j\right) + \varepsilon_t^j,\ \varepsilon_t^j \sim N\left(0, \left(\sigma_\varepsilon^j\right)^2\right),\ j = \{2, 3, 4, 5\}, \quad (4.1)$$

où $l_t^j,\ j = \{1, 2, 3, 4, 5\}$ est le signal d'une vague comme somme de la tendance, de la composante saisonnière et du BRE. Les coefficients de régression en (4.1) sont invariants dans le temps et s'obtiennent par régression de $\ln\left(z_t^j\right)^2$ sur $\ln\left(l_t^j\right)$ et $\ln\left(\left(z_{t-3}^{j-1}\right)^2\right)$ de la série initiale de l'EPA. Les exposants servent à désigner la vague à laquelle se rattachent les coefficients. Nous présentons les estimations des coefficients au tableau 4.1 avec la mesure R^2 corrigée de la qualité d'ajustement.

Tableau 4.1
Estimations de régression du processus des erreurs-types fondées sur le plan de sondage

	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
$\hat{c}$	12,219	-	-	-	-
$\hat{\beta}_j$	0,630	0,468	0,354	0,414	0,413
$\hat{\psi}_j$	-	0,717	0,786	0,749	0,751
$\hat{\sigma}_\varepsilon^j$	0,202	0,204	0,228	0,225	0,267
R^2_{adj}	0,351	0,373	0,386	0,477	0,342

La simulation se déroule de la manière suivante : pour chaque durée de série considérée et chaque simulation s, nous employons cinq signaux simulés $l_{t,s}^j,\ j = \{1, 2, 3, 4, 5\}$, pour produire cinq ensembles d'erreurs-types $z_{t,s}^j$ fondées sur le plan selon le processus défini en (4.1) et avec les coefficients de régression du tableau 4.1. Dès qu'un ensemble de données artificiel est produit, une estimation $\hat{\rho}_s$ est obtenue, après quoi le reste des hyperparamètres est estimé par la méthode du quasi-maximum de vraisemblance. À noter que le même ensemble d'erreurs-types $\mathbf{z}_{t,s}$ fondées sur le plan sert à produire toutes les séries bootstrap dans une simulation particulière.

Pour dégager les EQM réelles, nous mettons le modèle EPA en simulation un grand nombre de fois $(M = 50\,000)$, opération où chacune des itérations est assujettie aux mêmes limites que plus haut (entre zéro et un million de chômeurs). Nous calculons l'EQM réelle en prenant les valeurs réelles de vecteur d'état $\boldsymbol{\alpha}_{m,t}$ qui sont connues pour toute simulation m :

$$\mathbf{EQM}_t^{\text{Réel}} = \frac{1}{\mathrm{M}} \sum_{m=1}^{M} \left[\left(\hat{\boldsymbol{\alpha}}_{m,t}\left(\hat{\boldsymbol{\theta}}_m\right) - \boldsymbol{\alpha}_{m,t} \right) \left(\hat{\boldsymbol{\alpha}}_{m,t}\left(\hat{\boldsymbol{\theta}}_m\right) - \boldsymbol{\alpha}_{m,t} \right)' \right]. \quad (4.2)$$

L'EQM réelle du signal se calcule de la même manière à l'aide des valeurs de signal de la vague $\mathbf{l}_{m,t}$.

5 Résultats

5.1 Autres spécifications de modélisation pour l'EPA

On choisit et évalue habituellement les modèles SCS en employant des tests formels de diagnostic de normalité, d'homoscédasticité et d'indépendance des innovations normalisées. Une paramétrisation parcimonieuse est fondée sur des tests de rapport de vraisemblance logarithmique ou des critères

d'information (d'Akaike, de Bayes, etc.). Toutefois, les résultats de ces tests et critères dépendent des estimations ponctuelles particulières des hyperparamètres plutôt que de leurs distributions entières. Les distributions en simulation de Monte-Carlo (décrite à la section 4) des estimateurs des hyperparamètres nous éclairent davantage sur l'adéquation de la modélisation SCS. Les distributions en simulation nous livrent des indices sur l'éventuelle surspécification d'un modèle, en ce sens que certaines variables d'état pourraient être modélisées comme invariantes dans le temps.

Dans notre étude, nous considérons quatre modèles qui diffèrent pour le nombre d'hyperparamètres à estimer par la méthode du maximum de vraisemblance. Le modèle le plus complet, le modèle 1, est actuellement utilisé par Statistics Netherlands, mais après retrait de la composante de bruit blanc ε_t du paramètre réel de population ξ_t. On a constaté que cette composante avait une variance excessivement élevée et représentait une estimation perturbée d'autres hyperparamètres marginalement significatifs (variances de perturbation du BRE et de la composante saisonnière) dans le cas de l'EPA. En retranchant la composante irrégulière ε_t du modèle, on atténue l'instabilité des deux hyperparamètres précités. Cette formulation implique que le paramètre de population ξ_t n'accuse pas d'irrégularités impossibles à appréhender par la structure stochastique de la tendance et de la composante saisonnière. L'adoption de cette hypothèse peut être favorisée par une rigidité relative du marché du travail. L'évolution des niveaux de chômage est normalement progressive et doit donc être largement intégrée aux mouvements de la tendance stochastique. Les trois autres modèles sont des cas d'espèce du modèle 1, tous avec la composante irrégulière ε_t en moins (voir tableau 5.1).

Tableau 5.1
Hyperparamètres estimés dans les quatre versions du modèle EPA; les variances de perturbation sont estimées à l'échelle logarithmique

Modèles	Description	Paramètres estimés
M1	Modèle complet	$\rho, \sigma^2_{\eta_R}, \sigma^2_{\omega}, \sigma^2_{\eta_\lambda}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3}, \sigma^2_{v_4}, \sigma^2_{v_5}$
M2	Modèle saisonnier indépendant du temps	$\rho, \sigma^2_{\eta_R}, \sigma^2_{\eta_\lambda}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3}, \sigma^2_{v_4}, \sigma^2_{v_5}$
M3	Modèle BRE indépendant du temps	$\rho, \sigma^2_{\eta_R}, \sigma^2_{\omega}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3}, \sigma^2_{v_4}, \sigma^2_{v_5}$
M4	Modèle saisonnier et BRE indépendant du temps	$\rho, \sigma^2_{\eta_R}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3}, \sigma^2_{v_4}, \sigma^2_{v_5}$

Les distributions simulées des estimateurs des hyperparamètres dans le modèle 1 montrent que les hyperparamètres de variance pour la composante saisonnière et, en particulier, pour le BRE sont souvent estimés comme étant proches de zéro. Cela cause une bimodalité dans la distribution de ces estimations de variance avec une masse significative concentrée près de zéro. De plus, une tentative d'estimation de $\ln\left(\hat{\sigma}^2_{\omega}\right)$ ainsi que de $\ln\left(\hat{\sigma}^2_{\eta_\lambda}\right)$, comme dans le modèle 1, cause une distorsion dans la distribution des estimateurs de maximum de vraisemblance des autres hyperparamètres, laquelle devrait être normale. Ainsi, la normalité dans $\ln\left(\hat{\sigma}^2_{v_3}\right), \ln\left(\hat{\sigma}^2_{v_4}\right)$ et $\ln\left(\hat{\sigma}^2_{v_5}\right)$ est gravement compromise avec des valeurs aberrantes extrêmes et/ou un énorme coefficient d'applatissement (voir la figure A.1 en annexe où l'axe des x est étiré à cause des valeurs aberrantes), alors que les variances correspondantes sont moins susceptibles de présenter des valeurs extrêmes, étant censées fluctuer autour de l'unité. Si on rend la composante saisonnière invariante dans le temps comme dans le modèle 2, on ne change guère la situation des hyperparamètres de la tendance et du BRE. On pourrait même y voir un traitement moins qu'optimal, car les valeurs aberrantes

sont plus extrêmes et le coefficient d'applatissement est excessif dans la distribution des cinq hyperparamètres des erreurs d'enquête (figure A.2). Par contraste, nous avons pu constater (voir les figures A.3 et A.4) que, dans les deux modèles où la composante BRE est fixe dans le temps (modèles 3 et 4), toutes les estimations des hyperparamètres correspondant aux erreurs d'enquête étaient en distribution normale. Dans le modèle 3, les distributions demeurent asymétriques pour la pente et la composante saisonnière (asymétrie de -0,88 et -0,72 et applatissement de 5,56 et 4,61 respectivement). En fixant à zéro l'hyperparamètre saisonnier dans le modèle 4, l'amélioration est seulement marginale et la distribution de $\ln\left(\hat{\sigma}^2_{\eta_R}\right)$ présente un coefficient négatif d'asymétrie (-0,81) et un coefficient excessif d'applatissement (1,76).

Ces données de simulation semblent indiquer que, dans la modélisation des séries EPA, la préférence pourrait aller au modèle 3 plus parcimonieux, où la seule variance de perturbation BRE est fixée à zéro, mais comme le BRE même dépend du nombre de chômeurs, Statistics Netherlands conserve la variance de cet hyperparamètre à des fins de production afin de garder une souplesse suffisante devant l'évolution progressive du processus sous-jacent.

On peut recourir au test du rapport de vraisemblance pour vérifier si les hyperparamètres de la composante saisonnière et du BRE sont significativement différents de zéro, les modèles 2 à 4 étant imbriqués dans le modèle 1. La variable à tester comporte des valeurs très basses pour les trois autres modèles (0; 0,18 et 0,18 encore pour les modèles 2, 3 et 4, l'absence de différences entre les modèles 2 et 1 et entre les modèles 3 et 4 étant attribuable à la très faible valeur de l'hyperparamètre de la composante saisonnière). Ainsi, ces tests n'indiquent pas que les modèles plus parcimonieux présentent des résultats inférieurs à ceux du modèle 1. Une autre façon d'évaluer l'adéquation des quatre modèles est de les comparer sous l'angle de leur valeur prévisionnelle par la racine carrée des différences quadratiques moyennes (RDQM) entre les estimations ERG et les prédictions des signaux à un pas avant. On peut le faire pour chaque vague séparément : $\text{RDQM}^j = 1/(T-d)\sum_{t=d}^{T}\left(\hat{I}^j_{t|t-1} - Y_t{}^j\right)^2$, d étant égal à 20, 30 et 60 mois. Les résultats figurant en annexe (tableau B.1) montrent cependant qu'il n'y a guère de différence de rendement des quatre modèles dans leur application à la série initiale. Les modèles plus parcimonieux font voir une légère augmentation de la RDQM.

Les reformulations de modèle ne semblent pas influer sur la distribution de l'estimateur du paramètre autorégressif ρ des erreurs d'enquête sur les 1 000 séries simulées : on approche d'assez près la distribution normale et les valeurs vont de 0 à 0,4 quand $T = 114$, ce qui s'accorde avec l'approximation de sa distribution asymptotique à la sous-section 3.3. L'intervalle des valeurs est un peu plus étendu pour les séries temporelles plus courtes et plus étroites quand $T = 200$. Nous exécutons séparément pour les quatre modèles la procédure de simulation décrite dans la section précédente et l'analyse des méthodes bootstrap.

5.2 Estimation EQM

L'objet de notre étude par simulation est l'estimation EQM de la tendance et du signal de population, ce dernier étant la somme de la tendance et de la composante saisonnière. Nous évaluons le rendement du filtre de Kalman et des cinq méthodes d'estimation EQM à la section 3 en considérant le biais relatif et les EQM des estimateurs EQM. D'abord, nous prenons la moyenne des estimations EQM filtrées en (3.3), (3.4) et

(3.7) sur les 1 000 simulations (la moyenne est indiquée par la barre sur $\overline{\text{EQM}}_{t|t}$), alors que, dans le cas des estimations EQM par filtre de Kalman, nous l'établissons sur 10 000 simulations, comme nous l'avons mentionné au début de la section 4. Ces estimations EQM filtrées et mises en moyenne pour le modèle 3 (sauf pour la méthode AA; voir l'explication plus loin) sont décrites aux figures 5.1 à 5.4 pour $T = 48$, $T = 80$, $T = 114$ et $T = 200$ respectivement. Nous sautons les $d = 30$ premiers points temporels de l'échantillon (d devrait dépasser le nombre de points temporels nécessaires au début de la série pour éliminer l'effet d'une initialisation diffuse par le filtre). À noter que l'analyse est fondée sur des estimations filtrées plutôt que lissées, car ce sont les premières qui reproduisent le mieux le processus de production des chiffres officiels. Les EQM des quatre figures sont en configuration décroissante, comme on pouvait s'y attendre, parce que des estimations filtrées augmentent en précision si on dispose de plus d'information dans le temps pour estimer les variables d'état. Une exception à la règle, ce sont les EQM réelles de la figure 5.2. Une explication possible est que, dans cette application, les EQM des signaux sont proportionnelles aux signaux mêmes par les erreurs-types fondées sur le plan et que les EQM réelles reposent sur un autre ensemble (bien plus étendu) de séries simulées (50 000 pour les EQM réelles et 1 000 pour les EQM estimées). On remarquera que les traits de la figure 5.1 paraissent bien plus lisses, puisqu'ils s'étendent sur moins de points temporels. Ajoutons que, dans les figures 5.2 et 5.3, la configuration semble plus irrégulière, l'échelle de l'axe des y étant plus fine si on compare ces figures aux figures 5.1 et 5.4.

Nous calculons le biais relatif en pourcentage comme $\text{BR}_t^f = 100\%\left(\overline{\text{EQM}}_{t|t}^f \Big/ \text{EQM}_{t|t}^{\text{Réel}} - 1\right)$, où f correspond à une méthode d'estimation particulière et où $\text{EQM}_{t|t}^{\text{Réel}}$ est défini en (4.2). Les biais EQM relatifs en pourcentage et en moyenne dans le temps (après retrait des $d = 30$ premiers points temporels) pour le signal, la tendance et la composante saisonnière sont présentés aux tableaux 5.2, 5.3, 5.4 et 5.5.

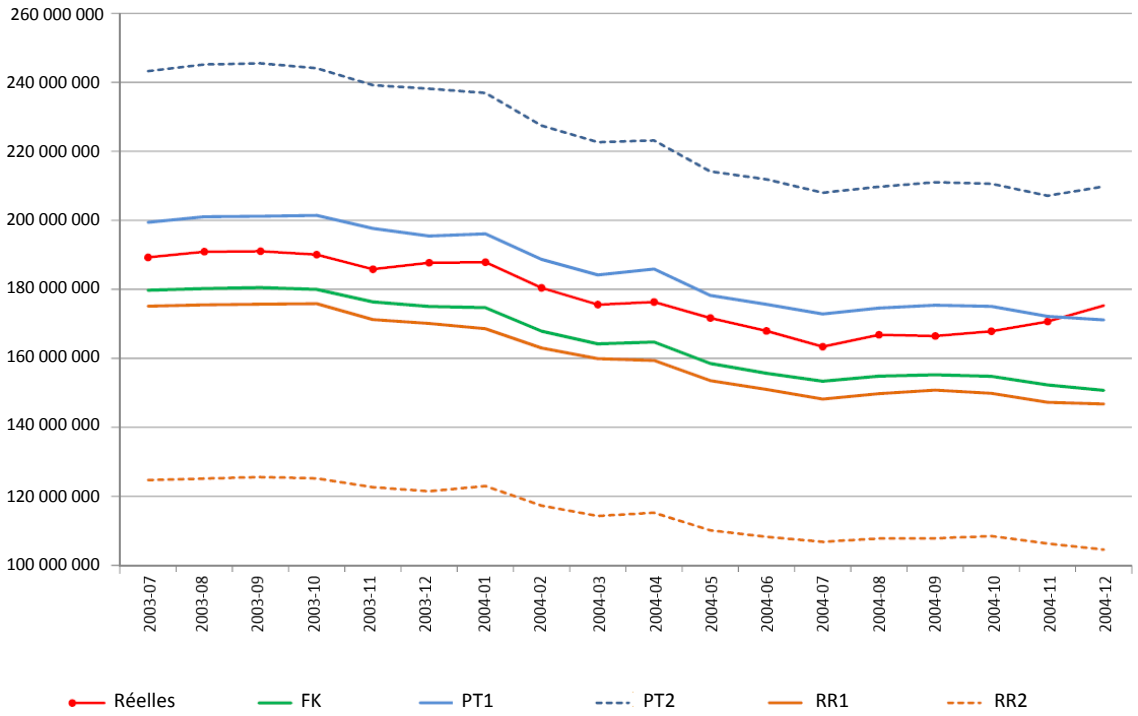

Figure 5.1 EQM réelles et EQM estimées moyennes pour le paramètre réel de population filtré (tendance et composante saisonnière) dans le modèle 3, $T = 48$ mois.

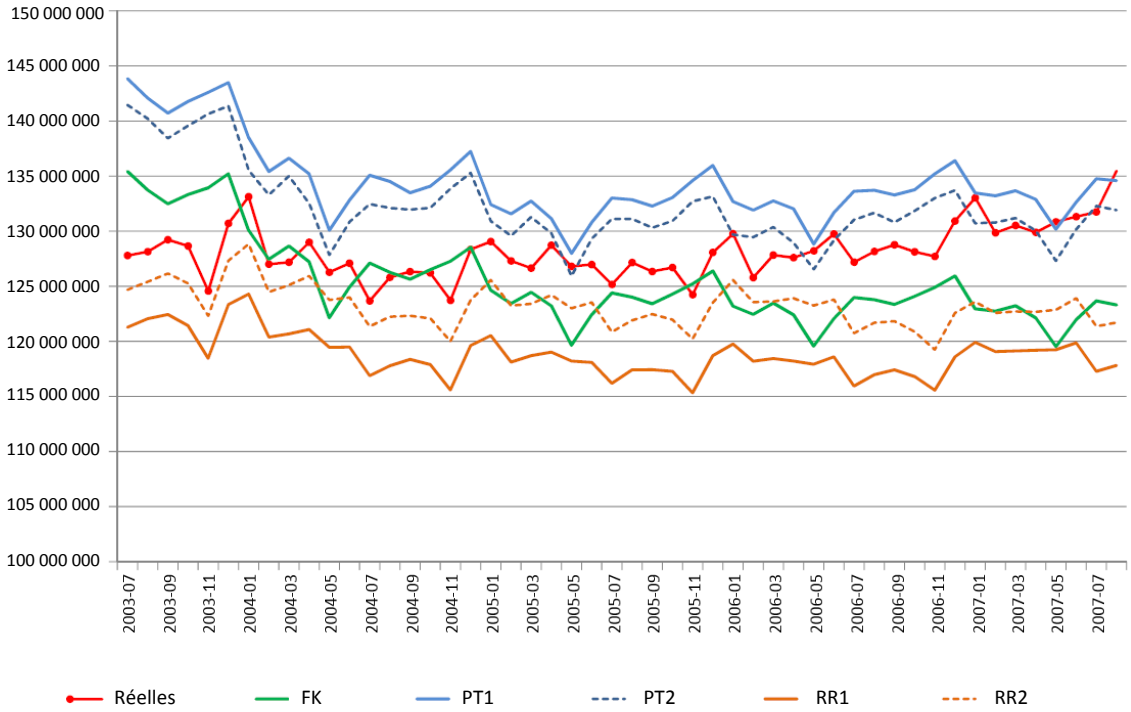

Figure 5.2 EQM réelles et EQM estimées moyennes pour le paramètre réel de population filtré (tendance et composante saisonnière) dans le modèle 3, $T = 80$ mois.

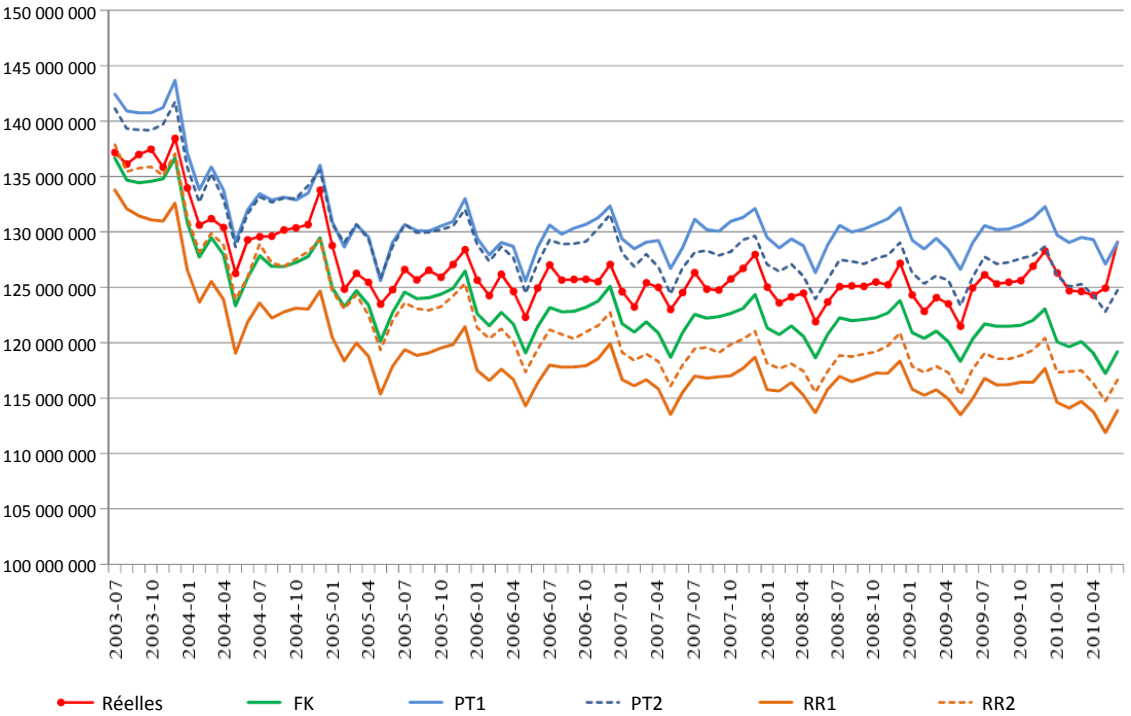

Figure 5.3 EQM réelles et EQM estimées moyennes pour le paramètre réel de population filtré (tendance et composante saisonnière) dans le modèle 3, $T = 114$ mois.

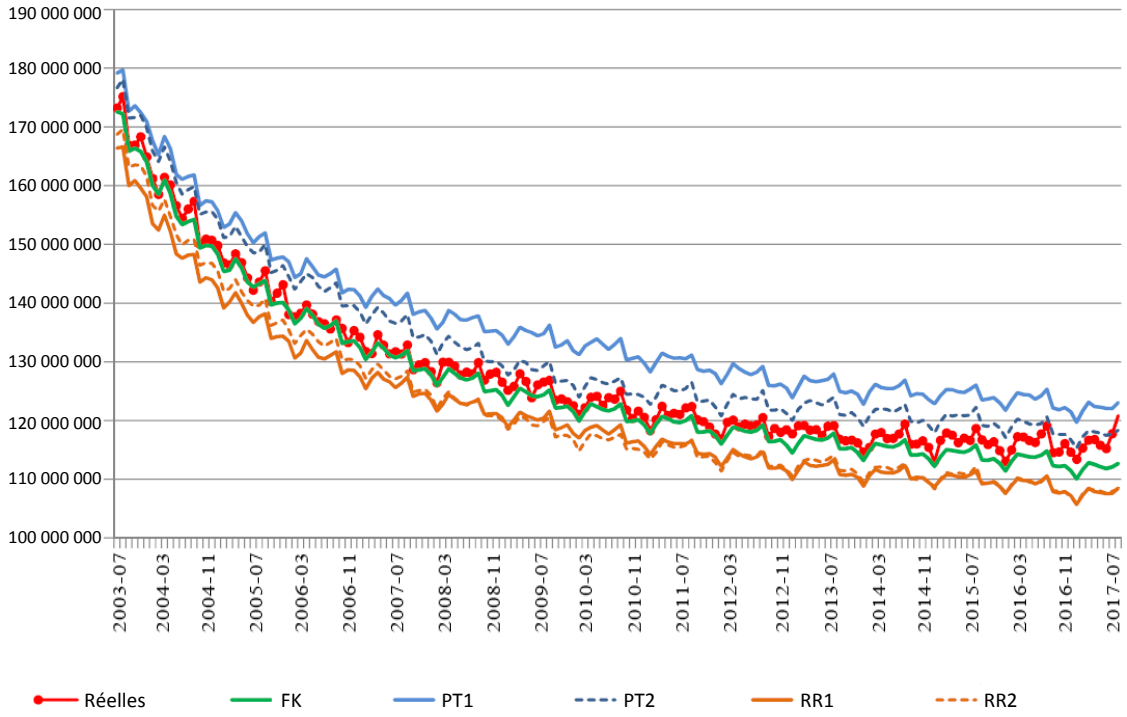

Figure 5.4 EQM réelles et EQM estimées moyennes pour le paramètre réel de population filtré (tendance et composante saisonnière) dans le modèle 3, $T = 200$ mois.

Tableau 5.2
Biais moyen en pourcentage des estimateurs EQM dans le modèle de l'EPA, $t = \{31,\ldots,T\}, T = 48$

Modèles	Signal*				Tendance				Composante saisonnière			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
FK	S.O.	S.O.	-7,1	-7,6	S.O.	S.O.	-6,5	-6,6	S.O.	S.O.	-6,7	-7,0
PT1	S.O.	S.O.	*4,4*	*1,4*	S.O.	S.O.	8,7	6,4	S.O.	S.O.	4,9	2,4
PT2	S.O.	S.O.	26,2	-4,4	S.O.	S.O.	22,4	-3,1	S.O.	S.O.	25,6	-4,6
RR1	S.O.	S.O.	-9,8	-10,8	S.O.	S.O.	-13,9	-13,8	S.O.	S.O.	-9,5	-10,1
RR2	S.O.	S.O.	-35,3	-5,6	S.O.	S.O.	-29,9	-3,2	S.O.	S.O.	-29,7	-5,1

* Le signal est la somme de la tendance et de la composante saisonnière.

Tableau 5.3
Biais moyen en pourcentage des estimateurs EQM dans le modèle de l'EPA, $t = \{31,\ldots,T\}$, $T = 80$

Modèles	Signal*				Tendance				Composante saisonnière			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
FK	-3,0	-3,2	-2,1	-2,2	-3,5	-3,8	-2,5	-2,5	8,8	2,5	2,9	2,4
AA	S.O.	S.O.	S.O.	14,9	S.O.	S.O.	S.O.	15,0	S.O.	S.O.	S.O.	14,9
PT1	8,6	6,7	4,9	6,2	10,6	8,9	7,1	8,4	20,8	10,7	10,3	11,1
PT2	*4,8*	*3,7*	*1,4*	*2,1*	4,8	4,9	2,1	2,3	17,3	8,2	6,9	7,1
RR1	-7,2	-9,0	-7,3	-7,2	-9,6	-11,2	-9,6	-9,5	-3,8	-9,0	-6,7	-6,6
RR2	6,7	-3,5	-3,9	-4,2	5,3	-4,1	-4,6	-5,4	18,6	-4,7	-4,1	-4,3

* Le signal est la somme de la tendance et de la composante saisonnière.

Tableau 5.4
Biais moyen en pourcentage des estimateurs EQM dans le modèle de l'EPA, $t = \{31,\ldots,T\}$, $T = 114$

Modèles	Signal*				Tendance				Composante saisonnière			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
FK	-2,1	-2,6	-2,4	-2,2	-2,3	-2,7	-2,4	-2,3	2,5	-3,2	-3,1	-2,6
AA	S.O.	S.O.	S.O.	5,2	S.O.	S.O.	S.O.	4,1	S.O.	S.O.	S.O.	12,5
PT1	8,1	5,7	3,3	5,5	10,0	7,9	5,2	7,6	4,9	1,4	1,4	0,3
PT2	*2,2*	*3,2*	*1,9*	*1,5*	3,3	4,3	3,1	2,8	1,2	-2,0	1,0	0,6
RR1	-8,3	-7,8	-6,4	-6,5	-10,7	-9,9	-8,7	-8,9	-3,1	-7,2	-5,5	-5,6
RR2	-1,1	-6,0	-3,9	-3,5	-3,0	-7,6	-5,5	-5,0	7,3	-5,9	-3,2	-3,0

* Le signal est la somme de la tendance et de la composante saisonnière.

Tableau 5.5
Biais moyen en pourcentage des estimateurs EQM dans le modèle de l'EPA, $t = \{31,\ldots,T\}$, $T = 200$

Modèles	Signal*				Tendance				Composante saisonnière			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
FK	-1,3	-1,6	-1,3	-1,3	-1,7	-1,8	-1,6	-1,6	3,8	-1,7	-1,6	-1,6
AA	S.O.	S.O.	S.O.	5,9	S.O.	S.O.	S.O.	5,6	S.O.	S.O.	S.O.	5,6
PT1	*6,3*	6,2	6,3	5,5	7,5	7,7	7,8	7,1	10,8	2,6	3,0	3,0
PT2	6,8	*4,0*	*3,0*	*2,3*	7,6	4,9	4,2	3,6	12,5	2,1	1,3	0,6
RR1	-8,0	-8,0	-4,9	-5,9	-10,0	-9,9	-6,8	-7,1	-1,1	-5,3	-3,8	-3,9
RR2	-5,1	-5,6	-4,5	-5,0	-7,0	-7,4	-6,0	-6,4	3,6	-3,1	-3,3	-3,9

* Le signal est la somme de la tendance et de la composante saisonnière.

Tableau 5.6
Variance estimée moyenne des EQM des estimateurs EQM pour le nombre de chômeurs dans le modèle de l'EPA (division par 10^{15}), $t = \{31,\ldots,T\}$, $T = 48$

Modèles	Signal*				Tendance				Composante saisonnière			
	M3		M4		M3		M4		M3		M4	
	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}
PT1	3,39	3,46	3,64	3,66	3,61	3,83	3,67	3,81	0,59	0,61	0,64	0,65
PT2	5,03	7,26	3,03	3,10	4,02	5,27	2,56	2,61	1,00	1,50	0,52	0,54
RR1	2,51	2,83	2,68	3,06	2,03	2,51	2,13	2,62	0,44	0,51	0,48	0,55
RR2	1,59	5,93	2,74	2,85	1,52	3,97	2,50	2,56	0,55	1,28	0,50	0,52

* Le signal est la somme de la tendance et de la composante saisonnière.

Tableau 5.7
Variance estimée moyenne des EQM des estimateurs EQM pour le nombre de chômeurs dans le modèle de l'EPA (division par 10^{15}), $t = \{31,\ldots,T\}$, $T = 80$

Modèles	Signal*				Tendance				Composante saisonnière			
	M3		M4		M3		M4		M3		M4	
	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}	Var_{EQM}	EQM_{EQM}
PT1	2,24	2,29	2,43	2,52	1,82	1,91	1,97	2,09	0,27	0,30	0,27	0,31
PT2	2,20	2,23	2,14	2,18	1,71	1,74	1,66	1,69	0,27	0,28	0,27	0,29
RR1	1.86	1,95	1,74	1,82	1,42	1,56	1,33	1,46	0,22	0,23	0,22	0,23
RR2	1,98	2,01	1,94	1,97	1,57	1,60	1,49	1,54	0,23	0,23	0,23	0,23

* Le signal est la somme de la tendance et de la composante saisonnière.

Voici les principales conclusions de notre étude par simulation :

1. Pour $T = 48$ et en moyenne dans le temps (à partir de $t = 31$), le biais relatif de l'EQM du signal après application du filtre de Kalman est d'environ -7 %. Ce biais tend à décroître à mesure que s'allonge la série. Le biais de filtre de Kalman (FK) est des plus modestes quand $T = 200$ et la situation est telle qu'aucune des méthodes d'estimation n'offre d'amélioration par rapport aux estimations EQM par filtre de Kalman. Nous pourrions toujours appliquer la meilleure méthode d'estimation avec des biais positifs pour dégager une plage de valeurs contenant l'EQM réelle.

2. Nous avons pu voir que la méthode AA (approximation asymptotique) est inapplicable aux modèles comportant des hyperparamètres marginalement significatifs. Quand on estime que certains des hyperparamètres sont proches de zéro, la matrice $\mathbf{I}^{-1}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{\sigma}}^{\mathrm{MV}} \middle| \rho^{a}\right)$ est numériquement singulière, d'où un échec de la procédure, ou quasi singulière. Dans ce dernier cas, la variance asymptotique devient excessivement élevée et perd donc toute fiabilité. Cela étant dit, la méthode AA serait uniquement envisageable pour le modèle 4. Comme on pouvait s'y attendre, la méthode donne de piètres résultats avec de courtes séries et laisse des biais positifs d'environ 15 %. Le rendement pour $T = 114$ et $T = 200$ est comparable à celui de la méthode bootstrap PT1, mais demeure significativement inférieur à celui de la méthode PT2.

3. Comme on peut immédiatement l'observer, l'emploi du bootstrap RR crée un biais négatif contrairement au bootstrap PT qui engendre un biais positif. À l'encontre de l'affirmation faite par Rodriguez et Ruiz (2012) que leur méthode offre de meilleures propriétés d'échantillon fini que la méthode de Pfeffermann et Tiller (2005), nous pouvons voir dans le cas de l'EPA que les estimations EQM par le bootstrap RR paramétrique ou non créent des biais négatifs plus importants que les estimations EQM par filtre de Kalman à l'échelle des modèles et des longueurs de séries (sauf pour RR2 dans le modèle 4 quand $T = 48$ et dans le modèle 1 quand $T = 80$ et $T = 114$). Alors que Pfeffermann et Tiller (2005) démontrent que leur méthode bootstrap présente des propriétés asymptotiques satisfaisantes, Rodriguez et Ruiz (2012) illustrent la supériorité de leur méthode dans de petits échantillons avec un modèle simple (à marche aléatoire et à bruit). La présente étude par simulation révèle que le bootstrap RR pourrait mal se comporter dans des applications plus complexes. Les méthodes PT n'ont jamais créé de biais négatifs pour l'EPA, ce qui en établit la « prudence » (sauf pour le bootstrap PT2 dans le modèle 4 quand $T = 48$ où le biais négatif demeure inférieur à celui de l'application du filtre de Kalman). Un autre résultat frappant pour $T = 48$ est que le biais positif du bootstrap PT2 et le biais négatif du bootstrap RR prennent des valeurs très élevées dans le modèle 3. Il reste que, avec une série si courte et autant de composantes non stationnaires comme dans le modèle de l'EPA, il est difficile de tirer des estimations fiables des méthodes bootstrap non paramétriques, puisque la période d'initialisation (avec son échantillon diffus) nécessaire à la production non paramétrique d'une série prend plus du quart de sa durée (13 mois sur 48).

4. Pour les séries de longueur $T = 114$ et $T = 80$, les biais positifs engendrés par la méthode PT2 dépassent légèrement les biais FK en valeur absolue dans les modèles comportant des hyperparamètres non significatifs (modèles 1 et 2). Dans les modèles plus stables (modèles 3 et 4), les biais positifs sont inférieurs aux biais négatifs FK en valeur absolue. Pour $T = 48$, nous présentons les résultats bootstrap seulement pour les modèles 3 et 4 (nous ne tenons pas compte des modèles 1 et 2 qui tendent à la surspécification à cause de problèmes numériques). Comme on pouvait s'y attendre, les biais sont plus importants pour une

telle durée des séries : les biais négatifs FK et RR s'accroissent en valeur absolue, tout comme les biais positifs PT, sauf pour le résultat PT2 précité dans le modèle 4.

L'EQM du signal dans le modèle 3, que nous pourrions considérer comme un meilleur choix pour la production des chiffres officiels de l'EPA, est estimée au mieux par la méthode PT2 avec des biais relatifs de 1,4 % et 1,9 % respectivement pour $T = 80$ et $T = 114$. Le bootstrap PT2 serait aussi la meilleure méthode pour $T = 200$, mais comme nous l'avons fait observer, les biais négatifs FK sont déjà des plus modestes pour des séries de cette longueur. Dans le cas de séries très courtes comme $T = 48$, le bootstrap PT1 paramétrique serait le meilleur.

5. Pour les méthodes PT et RR à la fois (sauf pour RR2 dans le modèle 4 avec $T = 48$), les valeurs absolues des biais relatifs sont moindres dans le cas des méthodes non paramétriques par rapport aux méthodes paramétriques. La supériorité du bootstrap non paramétrique peut s'expliquer par une distorsion de la normalité de la distribution des erreurs dans les modèles. Ainsi, notre préférence devrait aller aux bootstraps non paramétriques sauf pour des séries chronologiques très courtes.

6. Il n'y a pas que le biais des estimateurs EQM, puisque leur variabilité nous éclaire grandement aussi sur leur fiabilité. Autant que nous sachions, cet aspect n'a pas encore été exposé dans les études statistiques. Les tableaux 5.6 et 5.7 présentent les variances et les EQM des quatre estimateurs EQM bootstrap pour le signal, la tendance et la composante saisonnière dans le cas des longueurs de série les plus intéressantes, à savoir $T = 48$ et $T = 80$ (nous ne tenons pas compte des modèles 1 et 2, ni de l'approximation asymptotique en raison des problèmes numériques déjà évoqués). Les EQM des deux estimateurs EQM PT sont plus élevées que celles des deux estimateurs EQM RR tant pour le modèle 3 que pour le modèle 4. Si ces derniers semblent d'un rendement supérieur, comme en témoigneraient leurs EQM moindres, c'est que leurs variances sont plus petites. Toutefois, les biais sont parfois assez élevés pour porter les EQM de ces estimateurs EQM presque au niveau des EQM des estimateurs PT. Plus important encore, les biais des estimateurs EQM RR sont le plus souvent négatifs et dépassent fréquemment ceux des estimateurs par filtre de Kalman. Ce phénomène rend les bootstraps RR difficilement applicables dans le cas qui nous occupe.

Outre les résultats de simulation déjà mentionnés, il est également intéressant de voir si les modèles de séries chronologiques structurels (SCS) continuent d'offrir des estimations plus précises que les estimations de variance fondées sur le plan, même après correction de l'incertitude des hyperparamètres. C'est pourquoi nous mettons en comparaison les racines des EQM (REQM) obtenues avec les différentes procédures d'estimation EQM pour la série initiale $(T = 114)$, d'une part, et les erreurs-types (ET) de l'estimateur ERG. De telles différences moyennes des erreurs-types (DMET) dans le modèle m des séries chronologiques $(m = \{1,2,3,4\})$ se définissent ainsi : $\mathrm{DMET}_m^f = 100\,\%/(T-d)\sum_{t=d}^{T}\left[\mathrm{REQM}^f\left(\hat{l}_{t|t}^m\right) - \mathrm{ET}(Y_t)\right]/\mathrm{ET}(Y_t)$. Elles sont présentées au tableau 5.8, $\hat{l}_{t|t}^m$ étant l'estimation filtrée du paramètre réel de population défini comme la tendance et la composante saisonnière dans le modèle m. Nous décrivons les résultats pour le filtre de Kalman (FK) quand nous négligeons l'incertitude des hyperparamètres, ainsi que dans les cas où les cinq méthodes d'estimation EQM sont appliquées dans une prise en compte de cette même incertitude. Nous comparons aussi les REQM réelles en (4.2) aux erreurs-types ERG (« Réel » en ligne au tableau 5.8). À noter que le BRE et, en particulier, les estimations saisonnières des hyperparamètres par l'ensemble de données initial de l'EPA sont plutôt petits. Il n'y a donc pas de différences dignes de mention entre les

estimations ponctuelles du signal dans les quatre modèles. La méthode AA, la moins sûre, produit des erreurs-types surestimées (par rapport à la diminution de 18 % à 20 % pour les REQM réelles) à cause des matrices d'information quasi singulières des estimations de maximum de vraisemblance des hyperparamètres. Vu ce phénomène, on devrait se sentir plus en confiance dans l'utilisation des estimateurs PT. Bien que notre étude par simulation indique que le bootstrap PT2 est normalement d'un meilleur rendement que le bootstrap paramétrique PT1, pour cette série en particulier les ET dégagées par le bootstrap PT1 sont les plus proches des REMQ réelles avec une diminution d'environ 20 % des erreurs-types de l'estimation ERG. Ainsi, la modélisation permet une baisse significative de la variance comparativement à une approche plus classique fondée sur le plan, et ce, même après avoir pris en compte l'incertitude des hyperparamètres.

Tableau 5.8
Différences moyennes en pourcentage des erreurs-types (DMET) entre les estimateurs par la régression généralisée et les estimateurs de modélisation pour la série initiale de l'EPA, $d = 30$; augmentation en pourcentage des ET par filtre de Kalman après application de la correction EQM (entre parenthèses)

	Modèle 1	Modèle 2	Modèle 3	Modèle 4
FK	-24,1	-24,1	-24,5	-24,5
Valeur réelle	-20,0 (5,56)	-20,1 (5,5)	-20,6 (5,4)	-20,7 (5,3)
AA	-18,8 (6,9)	-19,0 (6,7)	-19,1 (7,1)	-19,5 (6,6)
PT1	-20,1 (5,2)	-20,1 (5,2)	-21,1 (4,6)	-21,2 (4,4)
PT2	-22,9 (1,6)	-21,2 (3,8)	-22,2 (3,1)	-22,5 (2,6)
RR1	-26,5 (-3,2)	-26,6 (-3,4)	-26,5 (-2,7)	-26,5 (-2,7)
RR2	-24,0 (-0,1)	-25,4 (-1,8)	-25,6 (-1,4)	-25,7 (-1,6)

6 Observations en conclusion

Les organismes nationaux de statistique s'intéressent de plus en plus à l'utilisation de modèles de séries chronologiques structurels (SCS) pour la production des chiffres mensuels de la population active. Aux Pays-Bas, un tel modèle est appliqué depuis 2010. Le modèle SCS représente une sorte d'estimation sur petits domaines (EPD) où l'information tirée d'échantillons de périodes antérieures permet d'obtenir des estimations plus précises, et de tenir compte du plan de sondage avec renouvellement de panel, lequel est souvent employé dans les enquêtes sur la population active.

Si l'on ne tient pas compte de l'incertitude des hyperparamètres dans les EQM des estimations fondées sur des modèles SCS, on se trouve à sous-estimer les EQM des estimations de domaines. Le biais qui se crée lorsqu'on écarte ainsi l'incertitude des hyperparamètres peut être important, plus particulièrement quand les séries sont courtes, ce qui est souvent le cas dans les organismes nationaux de statistique. La plupart des applications des procédures EPD dans les études spécialisées reposent sur des modèles multiniveaux, pratique courante lorsqu'il s'agit de tenir compte de l'incertitude des hyperparamètres. Les études consacrées au modèles SCS dans le contexte des estimations sur petits domaines sont plutôt limitées et la plupart des applications ne tiennent pas compte de cette incertitude dans les estimations EQM. L'importance du biais dans les EQM obtenues dépend de la structure du modèle et de la longueur de la série. Le présent article décrit une simulation de Monte-Carlo appliquée au modèle SCS qu'utilise Statistics Netherlands pour estimer le chômage mensuel. Cette simulation a un double but. D'abord, elle établit la

quantité de biais dans les EQM de l'EPA quand on néglige l'incertitude des hyperparamètres. De plus, nous comparons notre simulation à plusieurs méthodes d'estimation EQM disponibles dans la documentation spécialisée pour le cadre de modèles SCS et établissons ainsi la meilleure méthode pour l'EPA des Pays-Bas. En deuxième lieu, nous jugeons que la simulation des distributions des estimateurs des hyperparamètres permet de mieux comprendre la dynamique des composantes inobservées de le modèle SCS et donc de vérifier la nécessité de modéliser les composantes comme variant dans le temps. Dans le cas de l'EPA, la simulation fait voir l'intérêt éventuel d'adopter une version plus restreinte du modèle où le biais de renouvellement de l'échantillon serait invariant dans le temps et où le bruit blanc de population serait négligé. Pour cette double raison, nous recommandons d'effectuer une simulation comme celle que nous décrivons dans le processus de mise en œuvre du modèle servant à la production des statistiques officielles.

La comparaison des méthodes d'estimation EQM jette en outre un nouvel éclairage sur leurs propriétés. L'approximation asymptotique est inapplicable aux cas où les hyperparamètres sont proches de zéro, parce que la matrice d'information des estimations des hyperparamètres devient (presque) singulière. Les bootstraps non paramétriques, parce qu'ils dépendent moins d'hypothèses de normalité, sont d'un meilleur rendement que les bootstraps paramétriques selon Pfeffermann et Tailler (2005) et Rodriguez et Ruiz (2012) à la fois sauf si les séries sont très courtes. Notre constatation première est que les bootstraps PT présentent des biais positifs et sont invariablement d'un rendement supérieur à celui des bootstraps RR dont les biais sont généralement négatifs et plus importants (en valeur absolue) que dans l'application du filtre de Kalman. Elle contredit Rodriguez et Ruiz (2012) qui affirment la supériorité de leur méthode lorsque les séries chronologiques sont courtes. On peut penser que leurs résultats sont purement heuristiques, étant fondés sur un modèle simple (marche aléatoire et bruit), alors que Pfeffermann et Tiller (2005) démontrent que leur méthode bootstrap produit des estimations EQM avec un biais d'un bon ordre.

Les variances des estimateurs EQM PT sont plus élevées que celles des estimateurs RR correspondants. Les différences entre ces deux types d'estimateurs varient de modestes à modérées (les EQM des seconds sont inférieures de 28 % à 8 % aux EQM des premiers selon le modèle et la longueur de la série). Aspect plus important encore, la tendance des estimateurs RR à engendrer des biais négatifs parfois supérieurs à ceux de l'application du filtre de Kalman rend inapplicables ces méthodes bootstrap. Ainsi, on devrait généralement envisager de recourir aux méthodes PT pour d'autres données d'enquête, quoique leur rendement le cède occasionnellement à celui des méthodes RR.

Dans le cas des séries chronologiques très courtes, les bootstraps non paramétriques ne seraient pas un choix possible pour un modèle qui aurait la complexité que nous présentons. Il reste que le bootstrap paramétrique PT corrige les EQM aux biais négatifs jusqu'à dégager un léger biais positif (de 1,4 % à 4,4 % selon le modèle). Pour la présente durée de série de 114 mois, il est possible d'abaisser de -2,4 % à 1,9 % le biais EQM négatif grâce à la méthode non paramétrique de Pfeffermann et Tiller (2005) dans le modèle où le BRE est invariant dans le temps. Les racines des EQM réelles par filtre de Kalman sont inférieures d'environ 20 % aux erreurs-types des estimations ERG dans les quatre modèles appliqués aux données de l'EPA. En général, les biais des estimations EQM par filtre de Kalman sont relativement modestes dans l'application de l'EPA, aussi parraîtrait-il suffisant de s'en remettre à ces estimations naïves pour la publication des chiffres officiels.

Remerciements

Nous remercions Statistics Netherlands d'avoir financé cette étude. Nous remercions également le rédacteur adjoint et les examinateurs anonymes d'avoir lu attentivement notre manuscrit et formulé de précieuses observations. Les points de vue exprimés dans la présente sont ceux des auteurs et ne reflètent pas nécessairement les politiques de Statistics Netherlands.

Annexes

A. Densités simulées des hyperparamètres dans les quatre versions du modèle de l'EPA

Nous présentons en annexe les fonctions de densité des hyperparamètres obtenues par simulation quand les quatre versions du modèle de l'EPA (voir le tableau 5.1) servent de processus de génération de données. L'axe des x présente les hyperparamètres de variance à l'échelle logarithmique et l'axe des y porte les valeurs de fréquence. L'axe des x peut être étiré à cause des valeurs aberrantes.

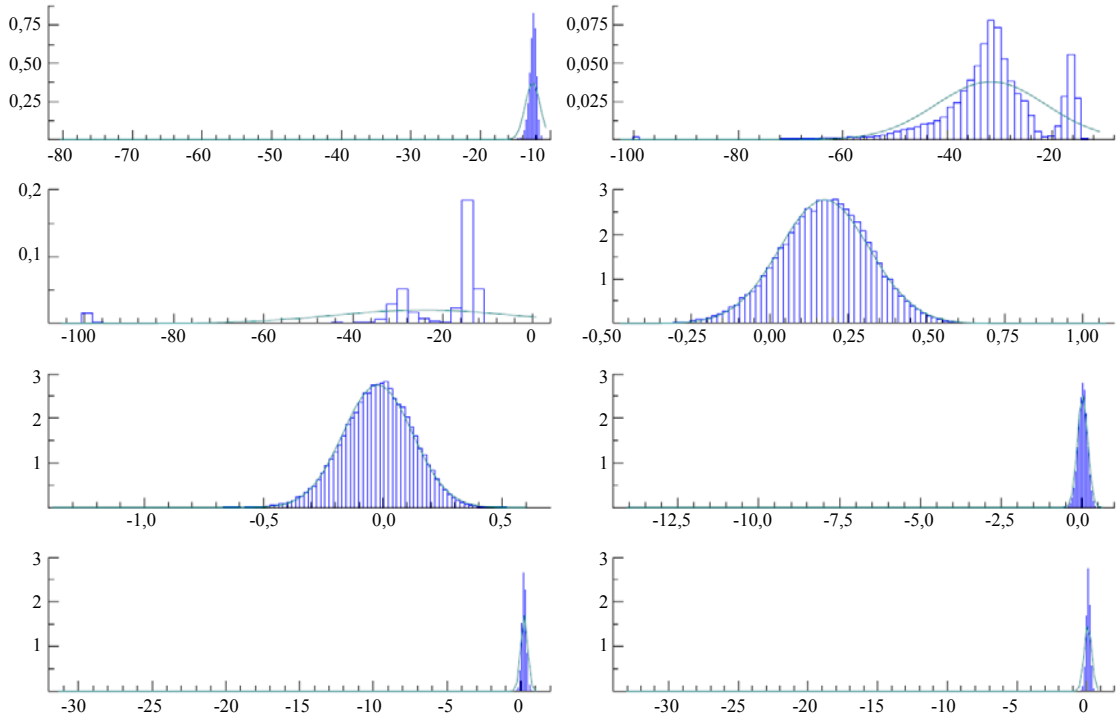

Figure A.1 Distribution des hyperparamètres sous le modèle complet de l'EPA (modèle 1), de gauche à droite sur l'axe des x : $\ln(\hat{\sigma}^2_R)$, $\ln(\hat{\sigma}^2_\gamma)$, $\ln(\hat{\sigma}^2_\lambda)$, $\ln(\hat{\sigma}^2_{v_t^t})$, $\ln(\hat{\sigma}^2_{v_t^{t-3}})$, $\ln(\hat{\sigma}^2_{v_t^{t-6}})$, $\ln(\hat{\sigma}^2_{v_t^{t-9}})$, $\ln(\hat{\sigma}^2_{v_t^{t-12}})$; densité normale avec les mêmes moyenne et variance superposée; 50 000 simulations, $T = 114$.

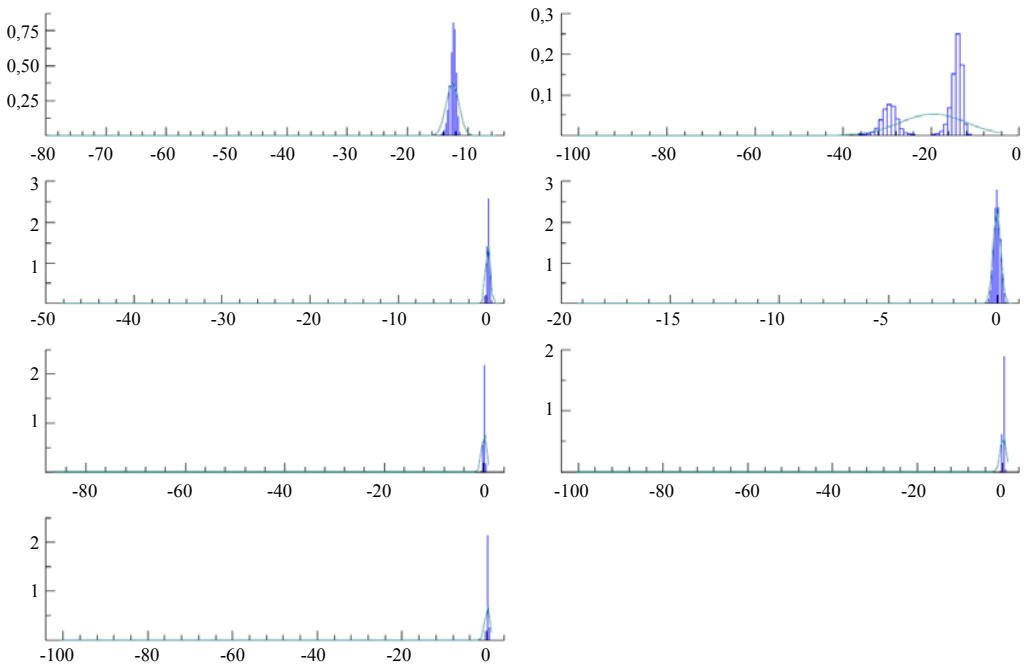

Figure A.2 Distribution des hyperparamètres sous le modèle 2, de gauche à droite sur l'axe des x : $\ln(\hat{\sigma}^2_R)$, $\ln(\hat{\sigma}^2_\lambda)$, $\ln(\hat{\sigma}^2_{v_t^t})$, $\ln(\hat{\sigma}^2_{v_t^{t-3}})$, $\ln(\hat{\sigma}^2_{v_t^{t-6}})$, $\ln(\hat{\sigma}^2_{v_t^{t-9}})$, $\ln(\hat{\sigma}^2_{v_t^{t-12}})$; densité normale avec les mêmes moyenne et variance superposée; 50 000 simulations, $T = 114$.

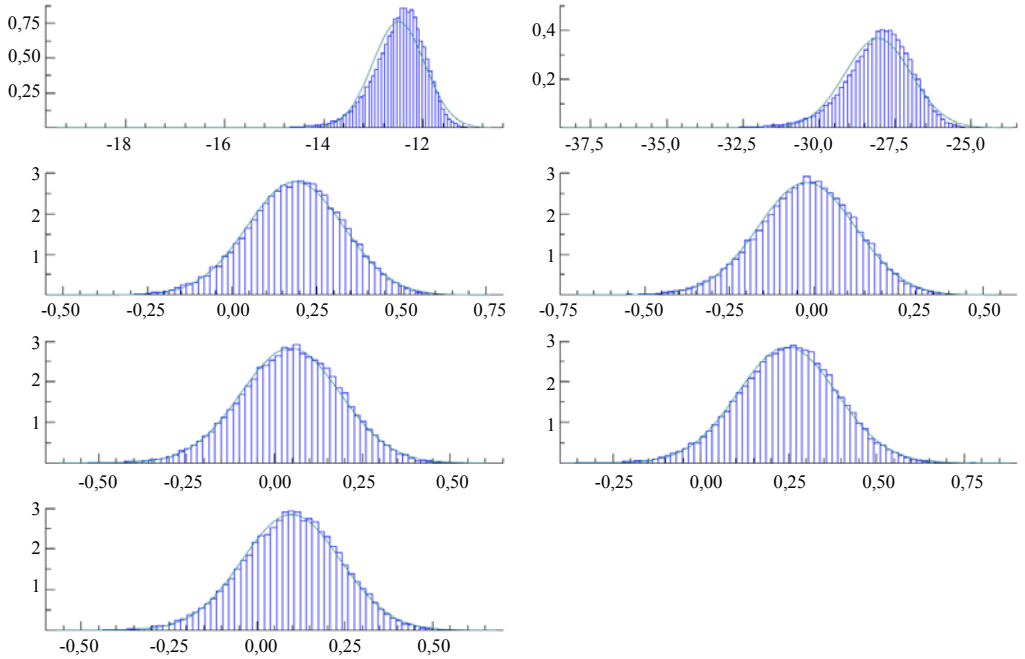

Figure A.3 Distribution des hyperparamètres sous le modèle 3, de gauche à droite sur l'axe des x : $\ln(\hat{\sigma}^2_R)$, $\ln(\hat{\sigma}^2_\gamma)$, $\ln(\hat{\sigma}^2_{v_t^t})$, $\ln(\hat{\sigma}^2_{v_t^{t-3}})$, $\ln(\hat{\sigma}^2_{v_t^{t-6}})$, $\ln(\hat{\sigma}^2_{v_t^{t-9}})$, $\ln(\hat{\sigma}^2_{v_t^{t-12}})$; densité normale avec les mêmes moyenne et variance superposée; 50 000 simulations, $T = 114$.

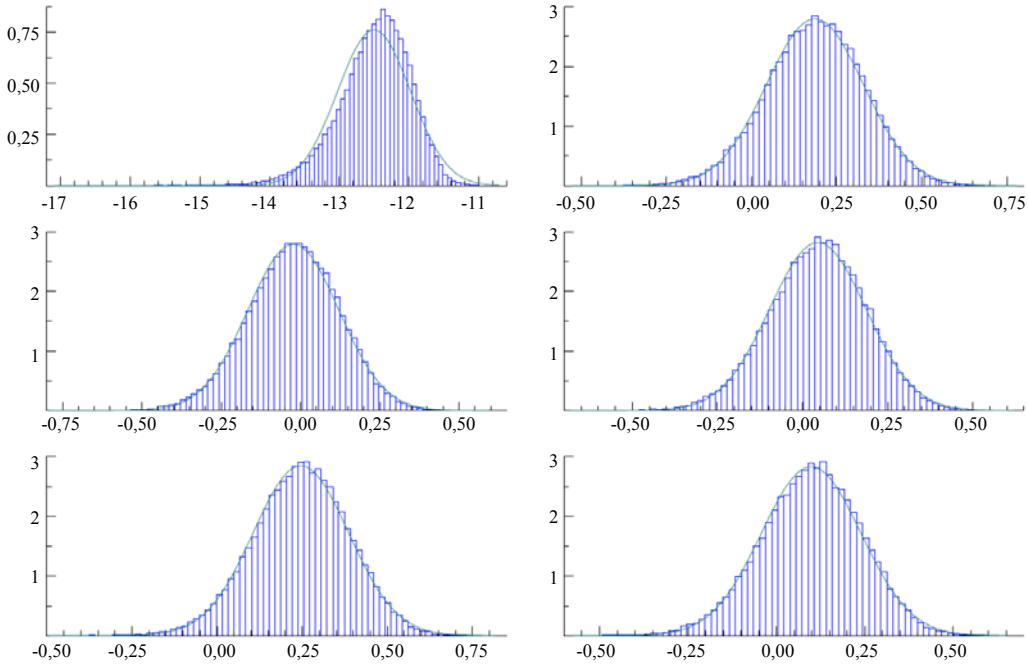

Figure A.4 Distribution des hyperparamètres sous le modèle 4, de gauche à droite sur l'axe des x : $\ln(\hat{\sigma}^2_R)$, $\ln(\hat{\sigma}^2_{v_t})$, $\ln(\hat{\sigma}^2_{v_t^{t-3}})$, $\ln(\hat{\sigma}^2_{v_t^{t-6}})$, $\ln(\hat{\sigma}^2_{v_t^{t-9}})$, $\ln(\hat{\sigma}^2_{v_t^{t-12}})$; densité normale avec les mêmes moyenne et variance superposée; 50 000 simulations, $T = 114$.

B. Rendement prévisionnel des quatre modèles de l'EPA

Tableau B.1
Racine des écarts quadratiques moyens des estimations par la régression généralisée du nombre de chômeurs par prédiction « un pas avant » et par vague

Vague	Modèle 1			Modèle 2			Modèle 3			Modèle 4		
	$d = 20$	$d = 30$	$d = 60$	$d = 20$	$d = 30$	$d = 60$	$d = 20$	$d = 30$	$d = 60$	$d = 20$	$d = 30$	$d = 60$
1	34 370	33 582	34 641	34 370	33 582	34 641	34 518	33 754	34 881	34 525	33 757	34 885
2	30 130	29 770	29 410	30 130	29 770	29 410	30 138	29 780	29 418	30 144	29 779	29 409
3	35 792	32 631	34 654	35 792	32 631	34 654	35 714	32 535	34 499	35 716	32 532	34 499
4	39 647	38 556	36 797	39 647	38 556	36 797	39 753	38 640	36 891	39 743	38 633	36 889
5	38 271	37 622	36 341	38 271	37 622	36 341	38 183	37 528	36 225	38 177	37 523	36 226

Bibliographie

Bailar, B. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association,* 70, 23-30.

Bartlett, M.S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society,* 8, 27-41.

Binder, D.A., et Dick, J.P. (1990). Méthode pour l'analyse des modèles ARMMI. *Techniques d'enquête,* 16, 2, 251-265. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1990002/article/14533-fra.pdf.

Bollineni-Balabay, O., van den Brakel, J. et Palm, F. (2016a). Multivariate state space approach to variance reduction in series with level and variance breaks due to survey redesigns. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 179, 377-402.

Bollineni-Balabay, O., van den Brakel, J. et Palm, F. (2016b). State space time series modelling of the Dutch Labour Force Survey: Model selection and MSE estimation, - Extended version. Document de travail, Statistics Netherlands, Heerlen. https://www.cbs.nl/en-gb/background/2016/41/state-space-time-series.

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

Doornik, J. (2007). *An Object-Oriented Matrix Programming Language Ox 5*. Timberlake Consultants Press, Londres.

Durbin, J., et Koopman, S.J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika,* 89, 603-615.

Durbin, J., et Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*. Numéro 38. Oxford University Press.

EUROSTAT (2015). Task force on monthly unemployment - revised report. Working group labour market statistics.

Hamilton, J. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics,* 33, 387-397.

Harvey, A. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.

Koopman, S.J. (1997). Exact initial kalman filtering and smoothing for nonstationary time series models. *Journal of the American Statistical Association,* 92, 1630-1638.

Koopman, S.J., Shephard, N. et Doornik, J. (2008). *SsfPack 3.0: Statistical Algorithms for Models in State Space Form*. Timberlake Consultants Press, Londres.

Krieg, S., et van den Brakel, J. (2012). Estimation of the monthly unemployment rate for six domains through structural time series modelling with cointegrated trends. *Computational Statistics & Data Analysis,* 56, 2918-2933.

Lemaître, G., et Dufour, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête,* 13, 2, 211-220. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1987002/article/14607-fra.pdf.

ONS (2015). A state space model for LFS estimates: Agreeing the target and dealing with wave specific bias. Rapport de la 29e réunion du Comité consultatif de la méthodologie des services statistiques du gouvernement. http://www.ons.gov.uk/ons/guide-method/method-quality/advisory-committee/previous-meeting-papers-and-minutes/mac-29-papers.pdf.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics,* 9, 163-175.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science,* 28, 40-68.

Pfeffermann, D., et Rubin-Bleuer, S. (1993). Modélisation conjointe robuste de séries de données sur l'activité pour de petites régions. *Techniques d'enquête,* 19, 2, 159-174. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1993002/article/14458-fra.pdf.

Pfeffermann, D., et Tiller, R. (2005). Bootstrap approximation to prediction MSE for state-space models with estimated parameters. *Journal of Time Series Analysis,* 26, 893-916.

Pfeffermann, D., Feder, M. et Signorelli, D. (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business and Economic Statistics,* 16, 339-348.

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

Rodriguez, A., et Ruiz, E. (2012). Bootstrap prediction mean squared errors of unobserved states based on the Kalman filter with estimated parameters. *Computational Statistics and Data Analysis,* 56, 62-74.

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

Tiller, R. (1992). Time series modelling of sample survey data from the US current population survey. *Journal of Official Statistics,* 8, 149-166.

van den Brakel, J., et Krieg, S. (2009). Estimation du taux de chômage mensuel par modélisation structurelle de séries chronologiques dans un plan de sondage avec renouvellement de panel. *Techniques d'enquête,* 35, 2, 193-207. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2009002/article/11040-fra.pdf.

van den Brakel, J., et Krieg, S. (2015). Remédier aux petites tailles d'échantillon, au biais de groupe de renouvellement et aux discontinuités dans les plans de sondage avec renouvellement de panel. *Techniques d'enquête,* 41, 2, 281-312. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2015002/article/14231-fra.pdf.

Zhang, M., et Honchar, O. (2016). Predicting survey estimates by state space models using multiple data sources. Article pour le Comité consultatif de la méthodologie de l'*Australian Bureau of Statistics*.

Techniques d'enquête, juin 2017
Vol. 43, nº 1, pp. 77-102
Statistique Canada, nº 12-001-X au catalogue

Inférence bayésienne prédictive sur une proportion sous un modèle double pour petits domaines avec corrélations hétérogènes

Danhyang Lee, Balgobin Nandram et Dalho Kim[1]

Résumé

Nous utilisons une méthode bayésienne pour inférer sur une proportion dans une population finie quand des données binaires sont recueillies selon un plan d'échantillonnage double sur des petits domaines. Le plan d'échantillonnage double correspond à un plan d'échantillonnage en grappes à deux degrés dans chaque domaine. Un modèle bayésien hiérarchique établi antérieurement suppose que, pour chaque domaine, les réponses binaires de premier degré suivent des lois de Bernoulli indépendantes et que les probabilités suivent des lois bêta paramétrisées par une moyenne et un coefficient de corrélation. La moyenne varie selon le domaine, tandis que la corrélation est la même dans tous les domaines. En vue d'accroître la flexibilité de ce modèle, nous l'avons étendu afin de permettre aux corrélations de varier. Les moyennes et les corrélations suivent des lois bêta indépendantes. Nous donnons à l'ancien modèle le nom de modèle homogène et au nouveau, celui de modèle hétérogène. Tous les hyperparamètres possèdent des distributions a priori non informatives appropriées. Une complication supplémentaire tient au fait que certains paramètres sont faiblement identifiés, ce qui rend difficile l'utilisation d'un échantillonneur de Gibbs classique pour les calculs. Donc, nous avons imposé des contraintes unimodales sur les distributions bêta a priori et utilisé un échantillonneur de Gibbs par blocs pour effectuer les calculs. Nous avons comparé les modèles hétérogène et homogène au moyen d'un exemple et d'une étude en simulation. Comme il fallait s'y attendre, le modèle double avec corrélations hétérogènes est celui qui est privilégié.

Mots-clés : Échantillonneur de Gibbs par blocs; modèle bayésien hiérarchique; corrélations intragrappe et intergrappes; qualité de l'ajustement; unimodalité; faiblement identifiable.

1 Introduction

Nous supposons qu'il existe plusieurs petits domaines, que chaque domaine est formé de plusieurs grappes et que chaque grappe contient un certain nombre d'unités (individus). Un échantillon aléatoire de grappes est tiré de chaque domaine et un échantillon aléatoire d'unités est tiré de chaque grappe échantillonnée. Il s'agit d'un plan d'échantillonnage double; voir Rao et Molina (2015). En cas d'échantillonnage en grappes, les unités à l'intérieur d'une grappe sont généralement positives et cette corrélation peut avoir une grande incidence sur l'inférence. Nous examinons cette situation pour les réponses binaires; voir Nandram (2015) qui a défini une corrélation intragrappe (entre deux unités dans la même grappe) et une corrélation intergrappes (entre deux unités dans deux grappes différentes du même domaine). Nous étendons le modèle de Nandram (2015), qui suppose que la corrélation demeure constante sur tous les domaines, pour traiter le cas où les corrélations peuvent être différentes. Nous nous intéressons à la proportion de la population finie pour chaque domaine, et comme Nandram (2015), nous utilisons un modèle bayésien hiérarchique à cette fin.

Lorsque les données présentent les corrélations susmentionnées, la corrélation intragrappe pose un problème statistique qui aboutit à une taille effective d'échantillon plus petite et, par conséquent, à une plus

1. Danhyang Lee, Department of Statistics, Iowa State University, Ames, Iowa 50011, États-Unis. Courriel : danhyang@iastate.edu; Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609, États-Unis. Courriel : balnan@wpi.edu; Dalho Kim, Department of Statistics, Kyungpook National University, 80 Daehakro, BukGu, Daegu 702-701, Corée. Courriel : dalkim@knu.ac.kr.

forte variabilité des estimations. Donc, quand il existe un effet de grappe, les analyses fondées sur l'hypothèse d'indépendance des unités donneront généralement des valeurs p plus petites (c'est-à-dire rejet quand cela ne devrait pas être le cas). Rao et Scott (1981, 1984) ont étudié ce problème et présenté de simples corrections de la statistique du khi carré classique pour le test d'indépendance dans des tableaux de contingence à double entrée sous un plan d'échantillonnage complexe (par exemple, échantillonnage en grappes à deux degrés).

Nandram et Sedransk (1993) ont présenté un modèle bayésien hiérarchique sous échantillonnage en grappes à deux degrés. Il s'agit du plan que nous appliquons dans chaque domaine dans le cas d'un plan d'échantillonnage double avec réponses binaires. En tant qu'analogue discret du modèle pour l'échantillonnage en grappes à deux degrés avec des données normales (Scott et Smith 1969), ce modèle fait une inférence au sujet de la proportion globale dans la population finie. Ce modèle a également été étendu par Nandram (1998) à des données multinomiales, ce qui peut être considéré comme un analogue bayésien du modèle multinomial-Dirichlet pour l'échantillonnage en grappes (Brier 1980).

En ce qui concerne la modélisation double, un nombre restreint d'études portent sur les variables de réponse continues et presque aucune ne s'applique aux données discrètes (binaires). La plupart des analyses relatives à la modélisation double sont fondées sur le cadre bayésien empirique. Fuller et Battese (1973) ont présenté des modèles de régression à erreurs emboîtées simples et doubles. Ghosh et Lahiri (1988) ont étudié l'échantillonnage à plusieurs degrés sous linéarité a posteriori en utilisant des méthodes bayésiennes ainsi que bayésiennes empiriques. Sous l'échantillonnage en grappes à deux et à trois degrés, l'estimation des modèles de régression avec structure d'erreurs emboîtées et variances d'erreur inégales a été examinée plus en profondeur par Stukel et Rao (1997). Des modèles pour petits domaines sous modèles de régression à erreurs emboîtées doubles ont également été étudiés par Stukel et Rao (1999); voir Rao et Molina (2015) pour une synthèse. Nandram (2015) a proposé un modèle bayésien hiérarchique pour les données binaires issues d'un plan d'échantillonnage double.

Nandram (2015) a montré qu'il est important de tenir compte du plan d'échantillonnage dans chaque domaine. En particulier, à l'instar de Rao et Scott (1981, 1984), il a établi que, si un modèle ne traduit pas le plan d'échantillonnage en grappes à deux degrés dans chaque petit domaine, le résultat sera trop optimiste. Autrement dit, la variabilité sera trop faible. Il s'avère aussi que les estimations ponctuelles pourraient être différentes si l'on omet de tenir compte de l'échantillonnage en grappes à deux degrés. Il a aussi remarqué qu'il existe d'autres situations où l'on pourrait observer le résultat opposé. Ainsi, sous un plan stratifié plutôt qu'un plan d'échantillonnage en grappes à deux degrés, la précision augmentera dans chaque domaine (c'est-à-dire pour chaque domaine, l'effet de plan sera inférieur à un). Consulter Nandram, Bhatta, Sedransk et Bhadra (2013) pour une analyse bayésienne de ce problème.

Afin d'accroître la flexibilité et la généralité du modèle bayésien hiérarchique double de Nandram (2015), nous généralisons ce dernier afin d'y intégrer des corrélations intragrappe inégales. Notre idée est d'étendre le modèle de Nandram (2015) en considérant une couche supplémentaire pour permettre à la corrélation intragrappe de varier d'un domaine à l'autre dans le plan d'échantillonnage double et de comparer le modèle double avec corrélation homogène (constante sur tous les domaines) et celui avec corrélations hétérogènes (variables d'un domaine à l'autre). Comme ceux du modèle homogène, les

paramètres du modèle hétérogène sont identifiés faiblement. L'utilisation d'un échantillonneur Monte Carlo par chaîne de Markov pour ajuster un tel modèle peut donner lieu à une dépendance de grande portée, et il sera difficile de surveiller la convergence d'un échantillonneur de Gibbs. Nandram (2015) a montré comment contourner la difficulté que créent ces paramètres faiblement identifiés en utilisant des tirages aléatoires. Molina, Nandram et Rao (2014), ainsi que Toto et Nandram (2010) discutent de tirages aléatoires similaires, en ayant évité entièrement l'ajustement de modèles Monte Carlo par chaîne de Markov. Malheureusement, le recours à des tirages aléatoires pour ajuster le modèle hétérogène n'est pas simple; nous sommes forcés d'utiliser l'échantillonneur de Gibbs.

Nous utilisons l'échantillonneur de Gibbs par blocs pour ajuster notre modèle double pour petits domaines. Deux difficultés se posent. Premièrement, les densités a posteriori conditionnelles des paramètres de corrélation peuvent être multimodales. Deuxièmement, certains paramètres peuvent être reliés de façon complexe. Durant l'utilisation d'un échantillonneur Monte Carlo par chaîne de Markov, ces situations risquent toutes deux aboutir à une dépendance de grande portée dans les itérations. Donc, pour essayer de contourner ces difficultés, nous avons appliqué une contrainte d'unimodalité aux densités a priori des paramètres de domaine et nous avons utilisé l'échantillonneur de Gibbs par blocs pour effectuer le tirage simultané de groupes de paramètres. Les deux stratégies accroissent la complexité, mais donnent des échantillonneurs nettement mieux ajustés.

En résumé, nous étendons le modèle de Nandram (2015) afin de tenir compte des corrélations hétérogènes. Le modèle avec corrélations hétérogènes est souhaitable, parce que si l'on suppose que la corrélation ne varie pas avec le domaine alors qu'elle le fait en réalité, les résultats pourraient être inexacts. Manifestement, cette extension de Nandram (2015) est une importante contribution. Toutefois, nous rencontrons trois difficultés.

1. Les corrélations hétérogènes introduisent des paramètres faiblement identifiables dans le modèle.
2. Contrairement à Nandram (2015), des méthodes Monte Carlo par chaîne de Markov sont nécessaires pour ajuster le modèle.
3. Une contrainte unimodale utile est imposée sur les hyperparamètres pour faciliter l'obtention d'un mélange approprié.

Nous présentons une construction novatrice d'un échantillonneur de Gibbs « à grille » (pour *griddy*) par blocs pour ajuster le modèle avec corrélations hétérogènes. Notre modèle est soumis à des tests approfondis, allant au-delà de Nandram (2015).

Dans le présent article, nous considérons l'inférence bayésienne prédictive sur les proportions d'un certain nombre de petits domaines dans une population finie quand on applique un plan d'échantillonnage en grappes dans chaque domaine. Dans nos principales contributions, nous utilisons un modèle bayésien hiérarchique contenant des corrélations intragrappe inégales pour faire une inférence a posteriori sur la proportion de chaque domaine dans la population finie. À la section 2, nous décrivons en détail le modèle hétérogène. En particulier, en guise de motivation et de mise à jour, nous commençons par passer brièvement en revue le modèle homogène de Nandram (2015). Nous montrons que certains paramètres peuvent être identifiés faiblement. Nous décrivons aussi les calculs pour tirer un échantillon aléatoire de la distribution

a posteriori en utilisant l'échantillonneur de Gibbs par blocs. À la section 3, afin de comparer les modèles avec corrélations homogènes et avec corrélations hétérogènes, nous présentons un exemple s'appuyant sur la *Third International Mathematics and Science Study* (TIMSS), ainsi qu'une petite étude en simulation. Enfin, à la section 4, nous exposons nos conclusions et les futures orientations de la recherche. Les annexes A et B fournissent les preuves et des renseignements complémentaires.

2 Modèles doubles bayésiens pour petits domaines et calculs

Nous considérons une population finie de ℓ domaines et de M_i grappes dans le i^{e} domaine, et nous supposons qu'il existe N_{ij} individus dans la j^{e} grappe dans le i^{e} domaine. Les réponses binaires sont y_{ijk} pour $i=1,\ldots,\ell,\ j=1,\ldots,M_i,\ k=1,\ldots,N_{ij}$. Nous supposons qu'un échantillon aléatoire simple de m_i grappes est tiré du i^{e} petit domaine et qu'un échantillon aléatoire simple de n_{ij} individus est tiré des m_i grappes échantillonnées provenant du i^{e} domaine. Ici, nous supposons que les poids de sondage sont les mêmes dans toutes les grappes dans chaque domaine. Soit $n_i=\sum_{j=1}^{m_i} n_{ij},\ s_{ij}=\sum_{k=1}^{n_{ij}} y_{ijk}$ et $s_i=\sum_{j=1}^{m_i} s_{ij}$.

Notre cible est la proportion du i^{e} domaine dans la population finie, qui est donnée par

$$P_i=\frac{\sum_{j=1}^{M_i}\sum_{k=1}^{N_{ij}} y_{ijk}}{N_i},\quad i=1,\ldots,\ell,$$

où $N_i=\sum_{j=1}^{M_i} N_{ij}$. Soit $T_{ij}^{(1)}=\sum_{k=n_{ij}+1}^{N_{ij}} y_{ijk}$ les totaux des unités non échantillonnées des grappes échantillonnées $(j=1,\ldots,m_i)$, et $T_{ij}^{(2)}=\sum_{k=1}^{N_{ij}} y_{ijk}$, les totaux des grappes non échantillonnées $(j=m_i+1,\ldots,M_i)$. En posant que $n_i=\sum_{j=1}^{m_i} n_{ij},\ \hat{p}_i=\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}} y_{ijk}\big/ n_i$, nous pouvons exprimer notre cible, P_i, sous la forme

$$P_i=\frac{n_i\hat{p}_i+\sum_{j=1}^{m_i} T_{ij}^{(1)}+\sum_{j=m_i+1}^{M_i} T_{ij}^{(2)}}{N_i},\quad i=1,\ldots,\ell. \tag{2.1}$$

Pour faire une inférence au sujet de P_i, nous ajustons des modèles bayésiens hiérarchiques aux données. En utilisant la représentation bêta-binomiale, ces modèles s'adaptent à la structure du plan double. Nous décrivons deux modèles, l'un avec une corrélation homogène et l'autre avec des corrélations hétérogènes, ce qui représente notre principale contribution à l'extension du modèle de Nandram (2015). À la section 2.1, nous examinons le modèle bayésien hiérarchique avec corrélation homogène de Nandram (2015) et nous montrons comment le rendre comparable à notre modèle bayésien hiérarchique avec corrélations hétérogènes que nous décrivons à la section 2.2. À la section 2.3, nous décrivons l'échantillonneur de Gibbs par blocs utilisé pour ajuster notre modèle avec corrélations hétérogènes.

2.1 Une revue du modèle double avec corrélation homogène

Nandram (2015) a décrit le modèle double pour petits domaines avec corrélation homogène. Ici, nous examinons brièvement les principales hypothèses qui le sous-tendent, à savoir

$$y_{ijk}\,\big|\,p_{ij}\overset{\text{ind}}{\sim}\text{Bernoulli}\left(p_{ij}\right), \tag{2.2}$$

$$\mu_i \,|\, \theta, \gamma \overset{\text{iid}}{\sim} \text{Bêta}\left[\theta\frac{1-\gamma}{\gamma},\ (1-\theta)\frac{1-\gamma}{\gamma}\right], \tag{2.3}$$

$$\rho, \theta, \gamma \overset{\text{iid}}{\sim} \text{Uniforme}(0,1), \tag{2.4}$$

où ρ et γ représentent les corrélations intragrappe et intergrappes, respectivement. L'hypothèse est que $0 < \theta, \rho, \gamma < 1$ strictement. Notons que, dans un même domaine, la corrélation intragrappe ρ, c'est-à-dire la corrélation entre deux unités dans une même grappe, est $\text{cor}\left(y_{ijk}, y_{ijk'} \,\middle|\, \mu_i, \gamma, \rho\right) = \rho,\ k \neq k'$. Semblablement, dans un même domaine, la corrélation intergrappes γ, c'est-à-dire la corrélation entre deux unités dans deux grappes différentes, est $\text{cor}\left(y_{ijk}, y_{ij'k'} \,\middle|\, \theta, \gamma, \rho\right) = \gamma,\ j \neq j',\ k \neq k'$. Ici, c'est ρ qui fait la distinction entre les modèles simple et double, et quand ρ tend vers zéro, le modèle double devient le modèle simple, Nandram (2015).

Pour ajuster le modèle spécifié par (2.2) à (2.4), Nandram (2015) a recouru à l'échantillonnage aléatoire et à la quadrature gaussienne pour exécuter des intégrations numériques unidimensionnelles. Il a également utilisé l'échantillonnage de Gibbs pour la comparaison et constaté de légères différences. Cependant, notre généralisation aux corrélations hétérogènes (nombre accru de paramètres) aboutit à des paramètres faiblement identifiés supplémentaires et l'ajustement du modèle devient plus difficile. Donc, nous intégrons des contraintes d'unimodalité sur les distributions a priori des paramètres de domaine, ce qui permet d'analyser des données éparses. Pour faire des comparaisons entre les deux modèles, l'un avec des corrélations homogènes et l'autre avec des corrélations hétérogènes, nous imposons aussi des contraintes d'unimodalité dans le modèle spécifié par (2.2) à (2.4). Nos résultats sous ce modèle homogène légèrement modifié sont semblables à ceux de Nandram (2015).

Les méthodes exposées dans le présent article permettent d'imposer l'unimodalité sur certaines distributions pour faciliter l'estimation des paramètres faiblement identifiés. Les conditions d'unimodalité sont suffisamment flexibles pour éviter de contraindre excessivement les modèles. Pour une procédure bayésienne non paramétrique complète, consulter Damien, Laud et Smith (1997). Donc, tout au long de nos calculs, nous appliquons la contrainte d'unimodalité aux hyperparamètres de $\mu_i\ (i = 1, \ldots, \ell)$,

$$\frac{\gamma}{1-\gamma} < \theta < \frac{1-2\gamma}{1-\gamma},\ 0 < \gamma < \frac{1}{3}. \tag{2.5}$$

Nous imposons aussi des contraintes d'unimodalité similaires à la section 2.2 pour le modèle avec corrélations hétérogènes. D'où, nous donnons au modèle spécifié par (2.2) à (2.5) le nom de modèle CHO (pour corrélation homogène).

Pour ajuster le modèle, Nandram (2015) utilise la règle de multiplication en obtenant p_{ij} après le tirage d'échantillons aléatoires de ($\boldsymbol{\mu}, \rho, \theta,$ et γ) à partir de leur densité a posteriori conjointe, où $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_\ell)'$. La densité a posteriori conditionnelle des p_{ij} est donnée par

$$p_{ij} \,|\, s_{ij}, \mu_i, \rho \overset{\text{ind}}{\sim} \text{Bêta}\left\{s_{ij} + \mu_i\frac{1-\rho}{\rho}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho}{\rho}\right\},$$

et, en posant que $s_{ij} = \sum_{k=1}^{n_{ij}} y_{ijk}$ et en agrégeant sur les p_{ij}, nous obtenons

$$\pi(\boldsymbol{\mu}, \rho, \theta, \gamma \,|\, \mathbf{y}) \propto \prod_{i=1}^{\ell} \prod_{j=1}^{m_i} \frac{B\left(s_{ij} + \mu_i \frac{1-\rho}{\rho}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho}{\rho}\right)}{B\left(\mu_i \frac{1-\rho}{\rho}, (1-\mu_i)\frac{1-\rho}{\rho}\right)}$$

$$\times \frac{\mu_i^{\theta\frac{1-\gamma}{\gamma}-1}(1-\mu_i)^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\left(\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right)}, \; 0<\mu_i, \rho<1, \quad i=1,\ldots,\ell, \; \frac{\gamma}{1-\gamma}<\theta<\frac{1-2\gamma}{1-\gamma}, \; 0<\gamma<\frac{1}{3}.$$

Parce que $T_{ij}^{(1)} \,|\, p_{ij} \overset{\text{ind}}{\sim} \text{Binomiale}\left(N_{ij} - n_{ij}, p_{ij}\right)$ et $T_{ij}^{(2)} \,|\, p_{ij} \overset{\text{ind}}{\sim} \text{Binomiale}\left(N_{ij}, p_{ij}\right)$ et que, sachant p_{ij}, $T_{ij}^{(1)}$ et $T_{ij}^{(2)}$ sont indépendants, après avoir obtenu les échantillons des p_{ij}, il est facile de faire une inférence bayésienne prédictive. Voir Nandram (2015) pour des renseignements détaillés.

2.2 Un modèle double avec corrélations hétérogènes

Nous étendons le modèle CHO pour pouvoir traiter les corrélations hétérogènes. Nos hypothèses sont

$$y_{ijk} \,|\, p_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}\left(p_{ij}\right), \tag{2.6}$$

$$p_{ij} \,|\, \mu_i, \rho_i \overset{\text{ind}}{\sim} \text{Bêta}\left[\mu_i \frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right], \tag{2.7}$$

$$\mu_i \,|\, \theta, \gamma \overset{\text{iid}}{\sim} \text{Bêta}\left[\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right], \tag{2.8}$$

$$\rho_i \,|\, \phi, \delta \overset{\text{iid}}{\sim} \text{Bêta}\left[\phi\frac{1-\delta}{\delta}, (1-\phi)\frac{1-\delta}{\delta}\right], \tag{2.9}$$

$$\theta, \gamma, \phi, \delta \overset{\text{iid}}{\sim} \text{Uniforme}(0,1). \tag{2.10}$$

Notons que le coefficient de corrélation intragrappe ρ introduit dans le modèle CHO est remplacé par $\rho_i \; (i=1,\ldots,\ell)$ pour fournir le modèle bayésien hiérarchique avec corrélations hétérogènes.

Comme pour le modèle CHO, nous imposons aussi a priori deux ensembles de contraintes d'unimodalité,

$$\frac{\gamma}{1-\gamma}<\theta<\frac{1-2\gamma}{1-\gamma}, \; 0<\gamma<\frac{1}{3} \text{ et } \frac{\delta}{1-\delta}<\phi<\frac{1-2\delta}{1-\delta}, \; 0<\delta<\frac{1}{3}. \tag{2.11}$$

L'annexe B donne des preuves simples des inégalités susmentionnées en tant que critères d'unimodalité et la façon d'intégrer ces contraintes dans nos calculs. Donc, nous dénommons modèle CHE (pour corrélations hétérogènes) le modèle bayésien hiérarchique spécifié par (2.6) à (2.11).

De nouveau, à l'instar de Nandram (2015), nous montrons à l'annexe A que, sous le modèle CHE,

$$\operatorname{cor}\left(y_{ijk}, y_{ijk'} \,\middle|\, \mu_i, \gamma, \rho_i\right) = \rho_i, \quad k \neq k', \tag{2.12}$$

$$\operatorname{cor}\left(y_{ijk}, y_{ij'k'} \,\middle|\, \theta, \gamma, \rho_i\right) = \gamma, \quad j \neq j', \quad k \neq k'. \tag{2.13}$$

En d'autres mots, à l'intérieur du i^{e} domaine, le coefficient de corrélation intragrappe est ρ_i et le coefficient de corrélation intergrappes est γ.

En appliquant le théorème de Bayes dans le modèle CHE, la densité conjointe a posteriori $\pi(\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \gamma, \phi, \delta \,|\, \mathbf{y})$ est facile à écrire. (Il s'agit de la densité sans la constante de normalisation.) Donc, nous pourrions donner à cette densité conjointe a posteriori le nom de posterior CHE.

Pour faire une inférence sur la proportion dans la population finie, P_i, nous tirons des échantillons de $\pi(\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \gamma, \phi, \delta \,|\, \mathbf{y})$ en utilisant la règle de multiplication et l'échantillonneur de Gibbs par blocs. Cette procédure est décrite à la section 2.3.

2.3 Calculs du posterior CHE

En premier lieu, notons que nous agrégeons le posterior CHE sur les p_{ij} et que nous utilisons ensuite l'échantillonneur de Gibbs pour ajuster la densité a posteriori marginale conjointe. Après avoir obtenu les échantillons, nous pouvons tirer des échantillons des p_{ij} à partir de densités a posteriori conditionnelles des p_{ij} en appliquant la règle de multiplication.

Comme dans le modèle CHO, la densité a posteriori conditionnelle des p_{ij} est

$$p_{ij} \,\middle|\, \mu_i, \rho_i, \theta, \gamma, \phi, \delta, \mathbf{y} \overset{\text{ind}}{\sim} \text{Bêta}\left\{s_{ij} + \mu_i \frac{1-\rho_i}{\rho_i}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right\}, \quad 0 < p_{ij} < 1.$$

Donc, il est facile de tirer des échantillons des p_{ij} une fois que les échantillons sont obtenus à partir de la densité a posteriori conjointe de $(\boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \gamma, \phi, \delta)$. Après élimination des p_{ij} du posterior CHE par intégration, la densité a posteriori conjointe marginale est donnée par

$$\pi(\boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \gamma, \phi, \delta \,|\, \mathbf{y}) \propto \prod_{i=1}^{\ell}\prod_{j=1}^{m_i} \frac{B\left(s_{ij} + \mu_i \frac{1-\rho_i}{\rho_i}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i \frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}$$

$$\times \frac{\mu_i^{\theta\frac{1-\gamma}{\gamma}-1}(1-\mu_i)^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\left(\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right)} \times \frac{\rho_i^{\phi\frac{1-\delta}{\delta}-1}(1-\rho_i)^{(1-\phi)\frac{1-\delta}{\delta}-1}}{B\left(\phi\frac{1-\delta}{\delta}, (1-\phi)\frac{1-\delta}{\delta}\right)}, \; 0 < \mu_i, \rho_i < 1, \quad i = 1, \ldots, \ell,$$

$$\frac{\gamma}{1-\gamma} < \theta < \frac{1-2\gamma}{1-\gamma}, \; 0 < \gamma < \frac{1}{3}, \; \frac{\delta}{1-\delta} < \phi < \frac{1-2\delta}{1-\delta}, \; 0 < \delta < \frac{1}{3}.$$

Les densités a posteriori conditionnelles sont

$$\pi\left(\mu_i \,|\, \rho_i, \theta, \gamma, \phi, \delta, \mathbf{y}\right) \propto \prod_{j=1}^{m_i} \frac{B\left(s_{ij}+\mu_i\frac{1-\rho_i}{\rho_i}, n_{ij}-s_{ij}+\left(1-\mu_i\right)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i\frac{1-\rho_i}{\rho_i}, \left(1-\mu_i\right)\frac{1-\rho_i}{\rho_i}\right)} \times \mu_i^{\theta\frac{1-\gamma}{\gamma}-1}\left(1-\mu_i\right)^{(1-\theta)\frac{1-\gamma}{\gamma}-1},$$

$$\pi\left(\rho_i \,|\, \mu_i, \theta, \gamma, \phi, \delta, \mathbf{y}\right) \propto \prod_{j=1}^{m_i} \frac{B\left(s_{ij}+\mu_i\frac{1-\rho_i}{\rho_i}, n_{ij}-s_{ij}+\left(1-\mu_i\right)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i\frac{1-\rho_i}{\rho_i}, \left(1-\mu_i\right)\frac{1-\rho_i}{\rho_i}\right)} \times \rho_i^{\phi\frac{1-\delta}{\delta}-1}\left(1-\rho_i\right)^{(1-\delta)\frac{1-\delta}{\delta}-1},$$

et, en posant $G_1 = \left\{\prod_{i=1}^{\ell}\mu_i\right\}^{1/\ell}$ et $G_2 = \left\{\prod_{i=1}^{\ell}\left(1-\mu_i\right)\right\}^{1/\ell}$,

$$\pi\left(\theta \,|\, \boldsymbol{\mu}, \boldsymbol{\rho}, \gamma, \phi, \delta, \mathbf{y}\right) \propto \left\{\frac{G_1^{\theta\frac{1-\gamma}{\gamma}-1} G_2^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\left(\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right)}\right\}^{\ell},$$

et

$$\pi\left(\gamma \,|\, \boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \phi, \delta, \mathbf{y}\right) \propto \left\{\frac{G_1^{\theta\frac{1-\gamma}{\gamma}-1} G_2^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\left(\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right)}\right\}^{\ell}.$$

De même, en posant $H_1 = \left\{\prod_{i=1}^{\ell}\rho_i\right\}^{1/\ell}$ et $H_2 = \left\{\prod_{i=1}^{\ell}\left(1-\rho_i\right)\right\}^{1/\ell}$,

$$\pi\left(\phi \,|\, \boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \gamma, \delta, \mathbf{y}\right) \propto \left\{\frac{H_1^{\phi\frac{1-\delta}{\delta}-1} H_2^{(1-\phi)\frac{1-\delta}{\delta}-1}}{B\left(\phi\frac{1-\delta}{\delta}, (1-\phi)\frac{1-\delta}{\delta}\right)}\right\}^{\ell},$$

et

$$\pi\left(\delta \,|\, \boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \gamma, \phi, \mathbf{y}\right) \propto \left\{\frac{H_1^{\phi\frac{1-\delta}{\delta}-1} H_2^{(1-\phi)\frac{1-\delta}{\delta}-1}}{B\left(\phi\frac{1-\delta}{\delta}, (1-\phi)\frac{1-\delta}{\delta}\right)}\right\}^{\ell}.$$

Le problème de cette procédure est que θ et γ sont corrélés, parce qu'intuitivement, ils dépendent tous deux uniquement de $\{\mu_i\}$ à travers deux nombres, G_1 et G_2, et non les données, $\mathbf{y}$. Cela donne un mauvais mélange dans l'échantillonneur de Gibbs. Par exemple, $E\left(\mu_i \,|\, \theta, \gamma\right) = \theta$, É.-T.$\left(\mu_i \,|\, \theta, \gamma\right) = \theta\sqrt{\gamma(1-\theta)/\theta}$ et $\mu_i \approx \theta\left\{1 + \mathrm{z}_i\sqrt{\gamma(1-\theta)/\theta}\right\}$, où $E\left(\mathrm{z}_i\right) = 0$ et $\mathrm{Var}\left(\mathrm{z}_i\right) = 1$, Nandram (2015). Autrement dit, $\{\mu_i\}$ est corrélé à θ et γ. Un problème similaire se manifeste dans $(\boldsymbol{\rho}, \phi, \delta)$. Par conséquent, afin de résoudre ces problèmes de faible identifiabilité, nous utilisons l'échantillonneur de Gibbs par blocs pour tirer des échantillons aléatoires de $(\boldsymbol{\mu}, \boldsymbol{\rho}, \theta, \gamma, \phi, \delta)$.

L'échantillonneur de Gibbs par blocs s'obtient en tirant $(\boldsymbol{\mu}, \theta, \gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y})$ et $(\boldsymbol{\rho}, \phi, \delta \,|\, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y})$ à tour de rôle de la densité a posteriori conditionnelle jusqu'à la convergence, comme nous le décrivons plus bas. Les deux densités a posteriori conditionnelles conjointes sont

$$\pi_1(\boldsymbol{\mu}, \theta, \gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y}) \propto \prod_{i=1}^{\ell} \prod_{j=1}^{m_i} \frac{B\left(s_{ij} + \mu_i \frac{1-\rho_i}{\rho_i}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i \frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}$$

$$\times \frac{\mu_i^{\theta\frac{1-\gamma}{\gamma}-1}(1-\mu_i)^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\left(\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right)}, \; 0<\mu_i<1, \quad i=1,\ldots,\ell, \; \frac{\gamma}{1-\gamma}<\theta<\frac{1-2\gamma}{1-\gamma}, \; 0<\gamma<\frac{1}{3}$$

et

$$\pi_2(\boldsymbol{\rho}, \phi, \delta \,|\, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y}) \propto \prod_{i=1}^{\ell} \prod_{j=1}^{m_i} \frac{B\left(s_{ij} + \mu_i \frac{1-\rho_i}{\rho_i}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i \frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}$$

$$\times \frac{\rho_i^{\phi\frac{1-\delta}{\delta}-1}(1-\rho_i)^{(1-\phi)\frac{1-\delta}{\delta}-1}}{B\left(\phi\frac{1-\delta}{\delta}, (1-\phi)\frac{1-\delta}{\delta}\right)}, \; 0<\rho_i<1, \quad i=1,\ldots,\ell, \; \frac{\delta}{1-\delta}<\phi<\frac{1-2\delta}{1-\delta}, \; 0<\delta<\frac{1}{3}.$$

Pour exécuter l'échantillonneur de Gibbs par blocs, nous appliquons la règle de multiplication dans $\pi_1(\boldsymbol{\mu}, \theta, \gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y})$ et $\pi_2(\boldsymbol{\rho}, \phi, \delta \,|\, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y})$; voir, par exemple, Molina et coll. (2014) et Toto et Nandram (2010).

D'abord, nous considérons $\pi_1(\boldsymbol{\mu}, \theta, \gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y})$. Nous éliminons $\boldsymbol{\mu}$ par intégration et obtenons la densité a posteriori conditionnelle conjointe de (θ, γ) sachant $\boldsymbol{\rho}, \phi, \delta$ et $\mathbf{y}$,

$$p(\theta, \gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y}) \propto \prod_{i=1}^{\ell} \left\{ \int_0^1 \left[\prod_{j=1}^{m_i} \frac{B\left(s_{ij} + \mu_i \frac{1-\rho_i}{\rho_i}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i \frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)} \right] \right.$$

$$\left. \times \frac{\mu_i^{\theta\frac{1-\gamma}{\gamma}-1}(1-\mu_i)^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\left(\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right)} d\mu_i \right\}, \; 0<\mu_i<1, \quad i=1,\ldots,\ell,$$

$$\frac{\gamma}{1-\gamma}<\theta<\frac{1-2\gamma}{1-\gamma}, \; 0<\gamma<\frac{1}{3}.$$

Ici, nous utilisons la somme de Riemann par la méthode du point milieu pour éliminer par intégration tous les $\mu_i,\ i = 1,\ldots,\ell.$ Nous subdivisons l'intervalle $(0, 1)$ en G sous-intervalles $(a_0, a_1]$, $(a_1, a_2],\ldots,[a_{G-1}, a_G]$, où $a_0 = 0,\ a_i = i/G,\ i = 1,\ldots G.$ Alors, nous pouvons calculer la distribution a posteriori conditionnelle conjointe de (θ, γ) comme il suit.

$$p(\theta, \gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y}) \propto \prod_{i=1}^{\ell}\left[\lim_{G\to\infty}\sum_{v=1}^{G} g_i\left(\frac{a_{v-1}+a_v}{2}\right)\{F_1(a_{v-1}) - F_1(a_v)\}\right],$$

$$g_i(\mu_i) = \prod_{j=1}^{m_i}\frac{B\left(s_{ij} + \mu_i\frac{1-\rho_i}{\rho_i}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}{B\left(\mu_i\frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right)}$$

et $F_1(\cdot)$ est la fonction de répartition correspondant à $f_1(\cdot)$, qui est une fonction de densité de $\text{Bêta}\left(\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right)$. Ensuite, nous éliminons également θ par intégration en utilisant la quadrature gaussienne au moyen des polynômes orthogonaux de Legendre,

$$p(\gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y}) \approx \sum_{g=1}^{G}\omega_g\left\{\prod_{i=1}^{\ell}\int_0^1 \pi_1\left(\mu_i, x_g, \gamma \,\middle|\, \rho_i, \phi, \delta, \mathbf{y}\right)d\mu_i\right\},$$

où $\{\omega_g\}$ sont les poids et $\{x_g\}$ sont les racines du polynôme de Legendre sur l'intervalle $\left[\frac{\gamma}{1-\gamma}, \frac{1-2\gamma}{1-\gamma}\right]$. Nous avons pris $G = 20$ dans nos calculs (de plus grandes valeurs de G ne font guère de différence).

Maintenant, nous pouvons utiliser une méthode à grille univariée (par exemple, Molina, Nandram et Rao 2014 et Toto et Nandram 2010) en vue de tirer des échantillons de la densité a posteriori de γ conditionnellement à $\boldsymbol{\rho}, \phi, \delta$ et $\mathbf{y}$; voir Ritter et Tanner (1992) pour une description de l'échantillonneur de Gibbs « à grille ». Alors, conditionnellement à γ, nous obtenons la densité a posteriori de θ comme il suit,

$$p(\theta \,|\, \gamma, \boldsymbol{\rho}, \phi, \delta, \mathbf{y}) \approx \sum_{g=1}^{G}\omega_g\left\{\prod_{i=1}^{\ell}\int_0^1 \pi_1\left(\mu_i, \theta \,\middle|\, \gamma, \rho_i, \phi, \delta, \mathbf{y}\right)d\mu_i\right\}.$$

Les échantillons sont tirés de la densité a posteriori conditionnelle de θ en utilisant de nouveau l'échantillonneur à grille univariée. Par la suite, conditionnellement à (θ, γ), $\boldsymbol{\mu}$ est tiré de $p(\boldsymbol{\mu} \,|\, \theta, \gamma, \boldsymbol{\rho}, \phi, \delta, \mathbf{y})$ en utilisant l'échantillonneur à grille univariée.

Pour la méthode à grille, nous divisons l'intervalle unitaire en sous-intervalles de 0,01 de largeur, et nous approximons la densité a posteriori conjointe par une distribution discrète avec probabilités proportionnelles aux hauteurs de la distribution continue aux points milieu de ces sous-intervalles. Notons que nous introduisons un bruit aléatoire (*jittering*) uniforme à l'intérieur de chaque intervalle sélectionné pour permettre différents écarts avec probabilité de un (Nandram 2015). Même quand nous avons utilisé des sous-intervalles plus fins (par exemple, largeur de 0,005), les résultats d'inférence ont été presque les mêmes. Donc, nous utilisons les sous-intervalles de 0,01 de largeur; voir Molina et coll. (2014). Lorsque la plupart de la distribution se trouve près de l'une des bornes (par exemple, 0 ou 1), nous créons des intervalles de plus petite largeur pour saisir les petites ou les grandes valeurs du paramètre.

Deuxièmement, nous considérons $\pi_2(\boldsymbol{\rho}, \phi, \delta \,|\, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y})$. Nous éliminons $\boldsymbol{\rho}$ par intégration et obtenons la densité a posteriori conditionnelle conjointe de (ϕ, δ) sachant $\boldsymbol{\mu}, \theta, \gamma$ et $\mathbf{y}$,

$$p(\phi, \delta \,|\, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y}) \propto \prod_{i=1}^{\ell} \left\{ \int_0^1 \left[\prod_{j=1}^{m_i} \frac{B\left(s_{ij} + \mu_i \frac{1-\rho_i}{\rho_i},\, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i} \right)}{B\left(\mu_i \frac{1-\rho_i}{\rho_i},\, (1-\mu_i)\frac{1-\rho_i}{\rho_i} \right)} \right] \right.$$

$$\left. \times \frac{\rho_i^{\phi\frac{1-\delta}{\delta}-1}(1-\rho_i)^{(1-\phi)\frac{1-\delta}{\delta}-1}}{B\left(\phi\frac{1-\delta}{\delta},\, (1-\phi)\frac{1-\delta}{\delta} \right)} \right\},\ 0<\rho_i<1,\ i=1,\ldots,\ell,\ \frac{\delta}{1-\delta}<\phi<\frac{1-2\delta}{1-\delta},\ 0<\delta<\frac{1}{3}.$$

De nouveau, nous appliquons la somme de Riemann par la méthode du point milieu pour éliminer par intégration tous les $\rho_i,\ i=1,\ldots,\ell$ et calculer la distribution a posteriori conditionnelle conjointe de (ϕ, δ),

$$p(\phi, \delta \,|\, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y}) \propto \prod_{i=1}^{\ell} \left[\lim_{G\to\infty} \sum_{v=1}^{G} h_i\left(\frac{a_{v-1}+a_v}{2} \right) \{F_2(a_{v-1}) - F_2(a_v)\} \right],$$

où

$$h_i(\rho_i) = \prod_{j=1}^{m_i} \frac{B\left(s_{ij} + \mu_i \frac{1-\rho_i}{\rho_i}, n_{ij} - s_{ij} + (1-\mu_i)\frac{1-\rho_i}{\rho_i} \right)}{B\left(\mu_i \frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i} \right)}$$

et $F_2(\cdot)$ est la fonction de répartition correspondant à $f_2(\cdot)$ qui est une fonction de densité de $\text{Bêta}\left(\phi\frac{1-\delta}{\delta}, (1-\phi)\frac{1-\delta}{\delta}\right)$. En utilisant la quadrature gaussienne au moyen des polynômes orthogonaux de Legendre, nous pouvons éliminer ϕ par intégration et obtenir la densité a posteriori conditionnelle de δ,

$$p(\delta \,|\, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y}) \approx \sum_{g=1}^{G} \omega'_g \left\{ \prod_{i=1}^{\ell} \int_0^1 \pi_2\left(\rho_i, x'_g, \delta \,\middle|\, \mu_i, \theta, \gamma, \mathbf{y} \right) d\rho_i \right\},$$

où $\{\omega'_g\}$ sont les poids et $\{x'_g\}$ sont les racines du polynôme de Legendre sur l'intervalle $\left[\frac{\delta}{1-\delta}, \frac{1-2\delta}{1-\delta}\right]$.

Alors, nous appliquons la méthode à grille univariée afin de tirer des échantillons de la densité a posteriori de δ conditionnellement à $\boldsymbol{\mu}, \theta, \gamma$ et $\mathbf{y}$. Par conséquent, nous pouvons représenter la densité a posteriori conditionnelle de ϕ par

$$p(\phi \,|\, \delta, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y}) \approx \sum_{g=1}^{G} \omega'_g \left\{ \prod_{i=1}^{\ell} \int_0^1 \pi_2(\rho_i, \phi \,|\, \delta, \mu_i, \theta, \gamma, \mathbf{y})\, d\rho_i \right\},$$

et obtenir des échantillons de θ en utilisant de nouveau l'échantillonneur à grille univariée. Enfin, conditionnellement à (ϕ, δ), $\boldsymbol{\rho}$ peut être tiré de $p(\boldsymbol{\rho} \,|\, \boldsymbol{\mu}, \theta, \gamma, \phi, \delta, \mathbf{y})$, où nous utilisons également la méthode à grille univariée.

Cet algorithme échantillonne $\pi_1(\boldsymbol{\mu}, \theta, \gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y})$ en tirant d'abord une itération de $\pi_1(\gamma \,|\, \boldsymbol{\rho}, \phi, \delta, \mathbf{y})$, une itération de $\pi_1(\theta \,|\, \gamma, \boldsymbol{\rho}, \phi, \delta, \mathbf{y})$, puis une itération de $\pi_1(\boldsymbol{\mu} \,|\, \theta, \gamma, \boldsymbol{\rho}, \phi, \delta, \mathbf{y})$. Ensuite, il échantillonne

$\pi_2(\boldsymbol{\rho}, \phi, \delta | \boldsymbol{\mu}, \theta, \gamma, \mathbf{y})$ en tirant d'abord une itération de $\pi_2(\delta | \boldsymbol{\mu}, \theta, \gamma, \mathbf{y})$, une itération de $\pi_2(\phi | \delta, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y})$, puis une itération de $\pi_2(\boldsymbol{\rho} | \phi, \delta, \boldsymbol{\mu}, \theta, \gamma, \mathbf{y})$. La procédure complète se poursuit jusqu'à la convergence. Cela revient à utiliser un échantillonneur de Gibbs avec deux densités a posteriori conditionnelles, ce qui est, en fait, l'échantillonneur de Gibbs par blocs. La construction de l'échantillonneur de Gibbs par blocs est très efficace et il s'agit de l'une de nos principales contributions dans le présent article. En fait, nous pourrions donner à l'échantillonneur de Gibbs par blocs le nom d'échantillonneur de Gibbs « à grille » par blocs (Ritter et Tanner 1992).

Nous avons examiné la convergence de l'échantillonneur de Gibbs par blocs en utilisant des tracés, des graphiques d'autocorrélation et le test de stationnarité de Geweke. Les tracés (itérations en fonction du temps) renseignent sur la durée de la période de rodage requise pour éliminer l'effet des valeurs initiales. Les graphiques d'autocorrélation montrent la dépendance dans la chaîne et, par conséquent, ceux présentant de fortes corrélations entre de longs décalages sont le signe d'une mauvaise chaîne de mélange. Le test de Geweke compare les moyennes de la partie initiale et de la partie ultérieure de la chaîne de Markov en utilisant une statistique de score z, où l'hypothèse nulle est que la chaîne est stationnaire; les valeurs p sont toutes supérieures à 0,10. Nous avons utilisé les tracés, les graphiques d'autocorrélation et le test de Geweke pour chaque paramètre afin d'étudier la convergence de chaque exécution de l'échantillonneur de Gibbs par blocs. Pour nos données, nous avons tiré 2 000 échantillons et en avons utilisé 1 000 pour le rodage afin d'obtenir un échantillon de 1 000 itérations pour l'inférence. Cette période de rodage, qui est basée sur les tracés et le test de Geweke, est suffisamment longue pour obtenir des échantillons aléatoires. Les corrélations sont toutes non significatives, et, ce qui est intéressant, nous ne devons pas réduire les itérations. En outre, le test de Geweke donne la preuve de la stationnarité de notre échantillonneur. Donc, nous disposons d'un échantillonneur de Gibbs par blocs très efficace. L'exécution de la procédure en R prend quelques minutes. Nous avons appliqué la même procédure pour notre étude en simulation.

3 Étude numérique et comparaisons

À la présente section, nous procédons à des études empiriques pour évaluer la performance du modèle CHE que nous comparons au modèle CHO. À la section 3.1, nous donnons un exemple et à la section 3.2, nous présentons une étude en simulation.

3.1 Un exemple

Nous utilisons des données recueillies auprès de la population d'élèves de troisième année aux États-Unis; voir Nandram (2015) pour une brève discussion de ces données. L'ensemble de données, recueilli en 1999, a trait à 2 477 élèves qui ont participé à la *Third International Mathematics and Science Study* (TIMSS). Foy, Rust, et Schleicher (1996) ont décrit le plan d'échantillonnage systématique avec probabilité proportionnelle à la taille (PPT) utilisé pour la collecte des données de la TIMSS, et Caslyn, Gonzales et Frase (1999) ont présenté les faits saillants de l'enquête. Les domaines sont formés en recoupant quatre régions (nord-est, sud, centre et ouest) et trois types de collectivité des États-Unis (village ou région rurale, périphérie d'une ville, et proximité du centre d'une ville). Donc, il y a douze domaines. La variable binaire est la question de savoir si la note de mathématique de l'élève est ou non inférieure à la moyenne. Les grappes sont les écoles, tandis que les unités dans les grappes sont les élèves.

Pour évaluer la qualité de l'inférence bayésienne prédictive, comme l'a suggéré un examinateur, Nandram (2015) a pris un échantillon correspondant à la moitié des données originales et l'a appelé échantillon synthétique. L'échantillon original a servi de population, et le demi-échantillon a été utilisé pour l'analyse, ce qui a fourni une méthode pour évaluer le pouvoir prédictif des modèles dans Nandram (2015). Dans le présent article, comme l'a proposé un examinateur, au lieu d'utiliser un demi-échantillon, nous nous servons de l'ensemble de données original à notre disposition; voir le tableau 3.1 pour la description de l'ensemble de données complet que nous analysons dans le présent article. Nous évaluons principalement le pouvoir prédictif du modèle CHE au moyen de l'étude en simulation.

Malheureusement, comme dans le cas de nombreuses enquêtes complexes, les analystes des données secondaires ne connaissent pas les fractions d'échantillonnage. Cependant, pour nombre de ces enquêtes, les fractions d'échantillonnage sont habituellement relativement faibles. Dans le cas des données de la TIMSS, nous supposons que l'ensemble de données est un échantillon de 5 % de la population. Par exemple, si quatre écoles sont échantillonnées pour un domaine, disons le i^{e} domaine $(i = 1, \ldots, \ell)$, le nombre total de grappes, M_i, est supposé être 80. Si 17 élèves sont observés dans une école échantillonnée, disons la j^{e} école, le nombre total d'élèves, $N_{ij}\,(j = 1, \ldots, m_i)$, est supposé être 340. Pour les écoles non échantillonnées, $N_{ij}\,(j = m_i + 1, \ldots, M_i)$ est supposé être la moyenne du nombre total d'élèves dans les écoles échantillonnées pour chaque domaine. En outre, dans de nombreuses écoles, beaucoup d'élèves, voire tous, étaient soit en dessous soit au-dessus de la moyenne. Cet ensemble de données est donc très épars, ce qui rend l'estimation directe difficile.

Tableau 3.1
Nombre d'élèves américains sous la moyenne en mathématique dans les écoles par domaine

Domaine	(*s*, *n*)	*m*	Écoles																
NR	40	4	9	10	11	10													
	74		17	16	21	20													
NP	60	9	8	7	12	3	12	8	7	1	2								
	173		20	21	17	19	16	25	22	14	19								
NC	135	11	9	20	1	22	20	11	26	10	1	12	3						
	222		15	23	16	25	22	25	27	19	16	22	12						
SR	84	8	6	14	14	9	14	10	12	5									
	140		16	21	16	14	23	19	22	9									
SP	164	16	14	9	12	10	18	11	3	0	13	9	13	8	11	10	19	4	
	298		19	14	13	18	22	18	21	16	18	15	26	9	19	22	25	23	
SC	150	13	16	11	13	6	8	9	13	6	11	15	15	18	9				
	225		16	13	17	16	19	16	18	12	19	16	19	21	23				
CR	17	2	7	10															
	39		16	23															
CP	59	7	13	11	5	15	3	2	10										
	140		22	18	9	19	24	23	25										
CC	145	14	21	1	12	9	12	13	16	13	7	12	7	8	4	10			
	259		21	26	22	13	16	18	21	18	17	18	17	19	16	17			
OR	54	7	13	11	4	2	7	11	6										
	118		15	19	10	16	16	20	22										
OP	117	13	8	11	15	9	7	10	1	15	14	9	7	6	5				
	224		13	13	25	16	20	12	20	18	20	17	17	17	16				
OC	331	31	9	17	10	12	15	15	8	22	20	7	18	7	13	15	13	8	
			6	8	17	13	9	6	12	7	11	4	9	8	2	3	7		
	515		18	22	10	14	15	15	8	23	22	7	18	10	26	29	13	17	
			16	14	18	15	13	23	21	26	16	11	14	14	17	15	15		

Nota : (*s*, *n*) représentent *s* (en haut), le nombre d'élèves dont la note est inférieure à la moyenne et *n* (en bas), la taille de l'échantillon [par exemple, NR compte 74 élèves échantillonnés dans $m = 4$ écoles pour un total de 40 élèves dont la note est inférieure à la moyenne]. Les domaines sont formés par le recoupement de la région (N : nord, S : sud, C : centre, O : ouest) et de la collectivité (R : rurale, P : périphérie d'une ville, C : centre d'une ville).

Nous appliquons trois procédures d'évaluation de la qualité de l'ajustement des modèles, à savoir le critère d'information de déviance (DIC pour *Deviance information criterion*), la valeur p prédictive a posteriori bayésienne (BPP pour *Bayesian posterior predictive* $p-value$) et le logarithme de la pseudo-vraisemblance marginale (LPML pour *Log pseudo marginal likelihood*), qui est une mesure fondée sur la même procédure de validation croisée avec suppression d'une unité (*leave-one-out*). Nous pouvons évaluer l'ajustement global des modèles au moyen de ces procédures.

Dans le modèle CHE, $s_{ij} \,|\, p_{ij} \overset{\text{ind}}{\sim} \text{Binomiale}\left(n_{ij}, p_{ij}\right), p_{ij} \overset{\text{ind}}{\sim} \text{Bêta}\left(\mu_i(1-\rho_i)/\rho_i, (1-\mu_i)(1-\rho_i)/\rho_i\right)$. Donc, en éliminant les p_{ij} par intégration, nous pouvons obtenir la fonction de masse de probabilité bêta-binomiale suivante,

$$f(\mathbf{s}\,|\,\boldsymbol{\mu}, \boldsymbol{\rho}) = \prod_{i=1}^{\ell}\prod_{j=1}^{m_i}\binom{n_{ij}}{s_{ij}}\frac{B\left(s_{ij}+\mu_i(1-\rho_i)/\rho_i,\, n_{ij}-s_{ij}+(1-\mu_i)(1-\rho_i)/\rho_i\right)}{B\left(\mu_i(1-\rho_i)/\rho_i, (1-\mu_i)(1-\rho_i)/\rho_i\right)}.$$

Il est également vrai que $E\left(s_{ij}\,|\,\mu_i, \rho_i\right) = n_{ij}\mu_i$ et $\text{Var}\left(s_{ij}\,|\,\mu_i, \rho_i\right) = n_{ij}\left\{1+\left(n_{ij}-1\right)\rho_i\right\}\mu_i(1-\mu_i)$.

Soit $\mu_i^{(h)}$ et $\rho_i^{(h)}$ $(i=1,\ldots,\ell,\, h=1,\ldots,H)$ les itérations provenant de l'échantillonneur de Gibbs par blocs. Soit $\overline{\mu}_i = \sum_{h=1}^{H}\mu_i^{(h)}/H$ $(i=1,\ldots,\ell)$ et $\overline{\rho}_i = \sum_{h=1}^{H}\rho_i^{(h)}/H$. En posant que $D(\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\rho}}) = -2\log\{p(\mathbf{s}\,|\,\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\rho}})\}$ et $\overline{D} = -2\sum_{h=1}^{H}\log\{p(\mathbf{s}\,|\,\boldsymbol{\mu}^{(h)}, \boldsymbol{\rho}^{(h)})\}/H$, le critère d'information de déviance est donné par

$$\text{DIC} = 2\overline{D} - D(\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\rho}}).$$

Les modèles dont le DIC est petit sont préférés à ceux dont le DIC est grand. Cependant, puisque le critère DIC a tendance à sélectionner des modèles surajustés, Nandram (2015) a décrit les valeurs p prédictives bayésiennes comme auxiliaire. Pour le modèle CHE, la fonction de divergence est

$$T(\mathbf{s}; \boldsymbol{\mu}, \boldsymbol{\rho}) = \sum_{i=1}^{\ell}\sum_{j=1}^{m_i}\frac{\left\{s_{ij} - E\left(s_{ij}\,|\,\mu_i, \rho_i\right)\right\}^2}{\text{Var}\left(s_{ij}\,|\,\mu_i, \rho_i\right)}.$$

Soient $\mathbf{s}^{(\text{rep})}$ les échantillons répétés (rep) tirés de la distribution prédictive a posteriori de $\mathbf{s}$. Alors, le critère BPP est $p\{T(\mathbf{s}^{(\text{rep})}; \boldsymbol{\mu}, \boldsymbol{\rho}) \geq T(\mathbf{s}^{(\text{obs})}; \boldsymbol{\mu}, \boldsymbol{\rho})\,|\,\mathbf{s}\}$, ce qui est calculé sur ses itérations correspondantes $(\boldsymbol{\mu}^{(h)}, \boldsymbol{\rho}^{(h)}), h=1,\ldots,H$. Une valeur de cette probabilité proche de 0 ou de 1 indique un mauvais ajustement du modèle. En fait, les modèles dont le BPP est compris dans l'intervalle (0,05; 0,95) sont considérés comme étant raisonnables.

En plus de ces quantités, nous pouvons évaluer la qualité de l'ajustement des modèles au moyen d'une autre mesure, le LPML, qui est une statistique sommaire des valeurs de l'ordonnée prédictive conditionnelle (CPO pour *Conditional predictive ordinate*), et est fondée sur une validation croisée. Contrairement au critère DIC, de grandes valeurs du LPML indiquent un meilleur ajustement des modèles (par exemple, Geisser et Eddy 1979).

Dans le cas du modèle CHE, le critère CPO peut être estimé par

$$\widehat{\text{CPO}}_{ij} = \left[\frac{1}{H} \sum_{h=1}^{H} \frac{1}{f\left(s_{ij} \mid p_{ij}^{(h)}\right)} \right]^{-1}, \quad j = 1, \ldots, m_i, \quad i = 1, \ldots, \ell,$$

où $p_{ij}^{(h)}$ représente les échantillons tirés de $p_{ij} \mid s_{ij}, \mu_i, \rho_i$ et $s_{ij} \mid p_{ij} \overset{\text{iid}}{\sim} \text{Binomiale}\left(n_{ij}, p_{ij}\right)$. Notons que, pour chaque (i, j), $\widehat{\text{CPO}}_{ij}$ est la moyenne harmonique des vraisemblances $f\left(s_{ij} \mid p_{ij}^{(h)}\right), h = 1, \ldots, H$. Alors, le LPML est donné par

$$\text{LPML} = \sum_{i=1}^{\ell} \sum_{j=1}^{m_i} \log\left(\widehat{\text{CPO}}_{ij}\right).$$

Ces trois mesures d'évaluation des modèles ont des formes similaires sous le modèle CHO. Pour le modèle CHO (CHE), DIC = 774,421 (773,173), BPP = 0,349 (0,408), LPML = -352,064 (-346,171), ce qui indique que le modèle CHE donne un meilleur ajustement. À un niveau de détail plus fin, nous avons également examiné les valeurs de CPO individuelles provenant des deux modèles pour chaque école. À la figure 3.1, nous comparons les CPO pour les modèles CHE et CHO, et nous constatons qu'en général, les valeurs de CPO sont plus élevées pour le modèle CHE que pour le modèle CHO. En fait, sous le modèle CHO (CHE), nous avons constaté que le pourcentage des valeurs de CPO inférieures à 0,025 est de 3,70 % (2,96 %) et que le pourcentage des valeurs de CPO inférieures à 0,014 est de 0,74 % (0,00 %). Ces résultats ne donnent aucun indice d'un écart important par rapport aux hypothèses de modélisation; voir Ntzoufras (2009). Par conséquent, à première vue, ces mesures donnent des preuves que le modèle CHE est un peu mieux ajusté aux données de la TIMSS que le modèle CHO.

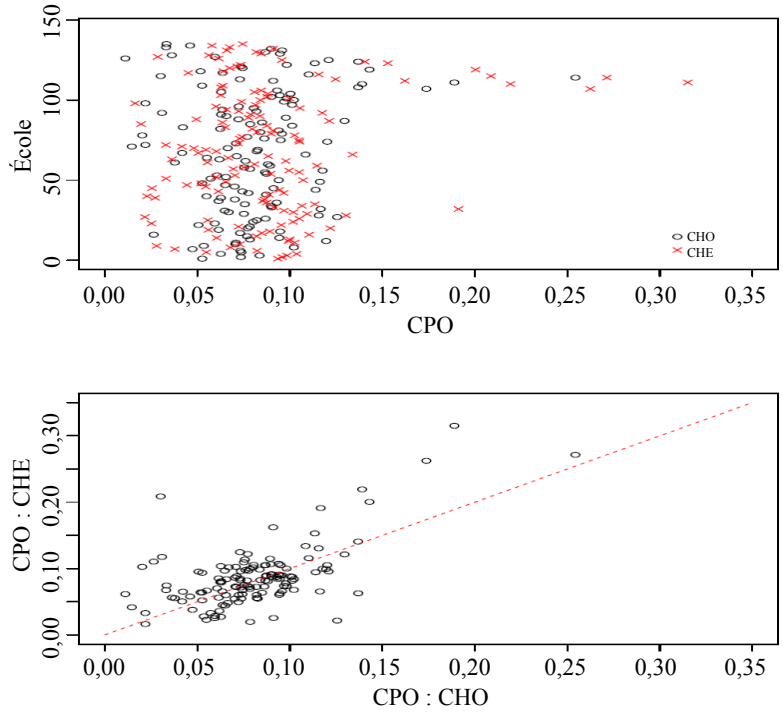

Figure 3.1 Nuages de points des CPO des modèles CHO et CHE (en haut : École en fonction du critère CPO) et (en bas : CHE en fonction de CHO).

Considérons maintenant l'inférence au sujet de θ et γ. Examinons d'abord θ. Sous le modèle CHO, la moyenne a posteriori (MP) vaut 0,519, l'écart-type a posteriori (ETP) vaut 0,068 et l'intervalle de crédibilité à 95 % (Cre) est (0,390; 0,639). Sous le modèle CHE, MP = 0,515, ETP = 0,065 et Cre à 95 % est (0,383; 0,639). Ensuite, examinons γ. Sous le modèle CHO, MP = 0,207, ETP = 0,011 et Cre à 95 % est (0,190; 0,224). Sous le modèle CHE, MP = 0,208, ETP = 0,011 et Cre à 95 % est (0,190; 0,225). Donc, il est bon de constater que les inférences au sujet de θ et γ sont très proches pour les deux modèles concurrents (modèles CHO et CHE).

Au tableau 3.2, nous présentons l'inférence a posteriori sur les proportions dans la population finie pour les notes de mathématique par domaine. Des différences existent entre les moyennes a posteriori sous les modèles CHO et CHE. La plupart sont faibles, mais quelques-unes sont grandes. Pour les domaines NC, SR et CR, nous avons 0,560 (0,543), 0,568 (0,584) et 0,465 (0,445) sous le modèle CHO (CHE), respectivement. Les écarts-types a posteriori sont également proches, mais il existe quelques différences modérément grandes (par exemple, pour NR, nous avons 0,113 sous le modèle CHO et 0,077 sous le modèle CHE). Les intervalles de crédibilité (Cre) et de densité a posteriori la plus grande (DPPG) reflètent ces différences.

Tableau 3.2
Comparaison de l'inférence a posteriori d'après les modèles doubles avec corrélation homogène (CHO) et corrélations hétérogènes (CHE) pour les proportions de la population finie pour les élèves américains sous la moyenne en mathématique par domaine

Domaine	Modèle CHO				Modèle CHE			
	MP	ETP	Cre à 95 %	DPPG à 95 %	MP	ETP	95 % Cre	DPPG à 95 %
NR	0,522	0,113	(0,299; 0,735)	(0,310; 0,741)	0,525	0,077	(0,363; 0,662)	(0,361; 0,658)
NP	0,365	0,075	(0,227; 0,524)	(0,227; 0,520)	0,359	0,072	(0,228; 0,511)	(0,236; 0,516)
NC	0,560	0,070	(0,420; 0,701)	(0,408; 0,680)	0,543	0,082	(0,370; 0,695)	(0,396; 0,710)
SR	0,568	0,080	(0,405; 0,725)	(0,424; 0,731)	0,584	0,062	(0,454; 0,699)	(0,456; 0,699)
SP	0,537	0,058	(0,423; 0,648)	(0,417; 0,639)	0,537	0,063	(0,409; 0,655)	(0,408; 0,653)
SC	0,646	0,064	(0,552; 0,766)	(0,522; 0,766)	0,654	0,059	(0,521; 0,763)	(0,544; 0,774)
CR	0,465	0,137	(0,195; 0,719)	(0,185; 0,709)	0,445	0,125	(0,212; 0,716)	(0,199; 0,700)
CP	0,437	0,085	(0,279; 0,603)	(0,276; 0,596)	0,439	0,091	(0,257; 0,620)	(0,265; 0,620)
CC	0,549	0,064	(0,415; 0,671)	(0,423; 0,672)	0,550	0,066	(0,414; 0,681)	(0,422; 0,685)
OR	0,461	0,086	(0,297; 0,629)	(0,295; 0,626)	0,460	0,085	(0,289; 0,626)	(0,276; 0,611)
OP	0,516	0,066	(0,384; 0,643)	(0,387; 0,644)	0,516	0,058	(0,401; 0,626)	(0,409; 0,633)
OC	0,670	0,042	(0,581; 0,748)	(0,586; 0,749)	0,662	0,047	(0,569; 0,748)	(0,568; 0,746)

Nota : MP est la moyenne a posteriori, ETP est l'écart-type a posteriori, Cre est l'intervalle de crédibilité à queues égales, et DPPG est l'intervalle de densité a posteriori la plus grande.

Le tableau 3.3 donne les valeurs sommaires de MP, ETP et DPPG à 95 % pour les corrélations intragrappe sous le modèle CHE. Nous voyons que les corrélations intragrappe varient d'un domaine à l'autre. L'estimation la plus élevée est 0,337 pour NC et la plus faible est 0,073 pour SR. Ces deux domaines présentent quelques grandes différences entre les moyennes a posteriori sous les modèles CHO et CHE. L'intervalle DPPG à 95 % pour la corrélation commune dans le modèle CHO, qui est (0,160; 0,260), est contenu par tous les intervalles, sauf ceux pour NR, NC, SR et OC. Donc, il est raisonnable d'étudier le modèle CHE.

Tableau 3.3
Valeurs sommaires a posteriori pour les corrélations intragrappe des modèles doubles avec corrélations hétérogènes pour les élèves américains sous la moyenne en mathématique par domaine

Domaine	MP	ETP	Cre à 95 %	DPPG à 95 %
NR	0,076	0,084	(0,002; 0,301)	(0,001; 0,251)
NP	0,184	0,087	(0,053; 0,380)	(0,042; 0,358)
NC	0,337	0,087	(0,190; 0,520)	(0,184; 0,513)
SR	0,073	0,067	(0,003; 0,252)	(0,001; 0,216)
SP	0,237	0,075	(0,113; 0,393)	(0,110; 0,387)
SC	0,176	0,079	(0,055; 0,356)	(0,048; 0,329)
CR	0,149	0,147	(0,003; 0,523)	(0,001; 0,445)
CP	0,233	0,103	(0,079; 0,486)	(0,050; 0,434)
CC	0,235	0,077	(0,105; 0,388)	(0,099; 0,381)
OR	0,181	0,099	(0,033; 0,413)	(0,021; 0,378)
OP	0,181	0,075	(0,059; 0,362)	(0,048; 0,327)
OC	0,301	0,063	(0,191; 0,437)	(0,188; 0,434)

Nota : En utilisant le modèle double avec corrélation homogène, MP = 0,211, ETP = 0,026, Cre à 95 % = (0,162; 0,266), et DPPG à 95 % = (0,160; 0,260). MP est la moyenne a posteriori, ETP est l'écart-type a posteriori, Cre est l'intervalle de crédibilité à queues égales, et DPPG est l'intervalle de densité a posteriori la plus grande.

À la figure 3.2, nous comparons les densités a posteriori des corrélations intragrappe pour le modèle CHE (douze corrélations) et le modèle CHO (une corrélation). Les distributions sous le modèle CHE sont plus variables, et elles se situent principalement à la gauche ou à la droite de celles sous le modèle CHO, le chevauchement étant faible pour certains domaines (par exemple, NR, NC et SR).

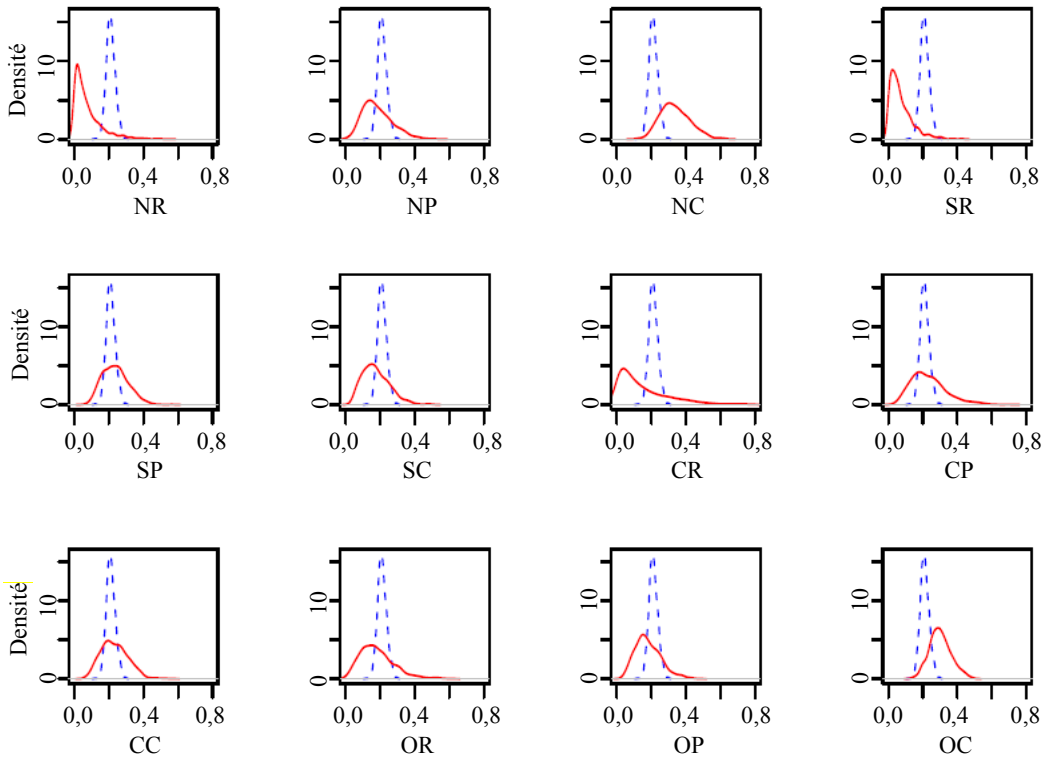

Figure 3.2 Courbes des densités a posteriori des corrélations intragrappes pour les notes de mathématique par domaine (trait plein : modèle CHE, trait pointillé : modèle CHO).

Aux figures 3.3, 3.4 et 3.5, nous comparons les courbes des densités a posteriori des proportions de la population finie pour les notes de mathématique par domaine pour les deux modèles. Des différences appréciables s'observent entre les modèles CHO et CHE (par exemple, domaines NR, NC, SR, CR et OC).

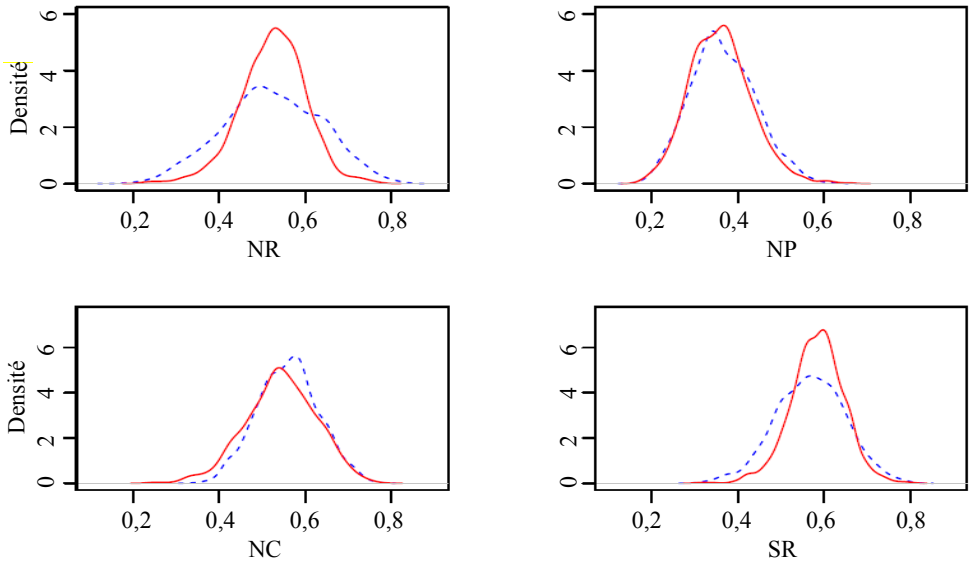

Figure 3.3 Courbes des densités a posteriori des proportions dans la population finie pour les notes de mathématique par domaine (NR, NP, NC, SR) (trait plein : modèle CHE, trait pointillé : modèle CHO).

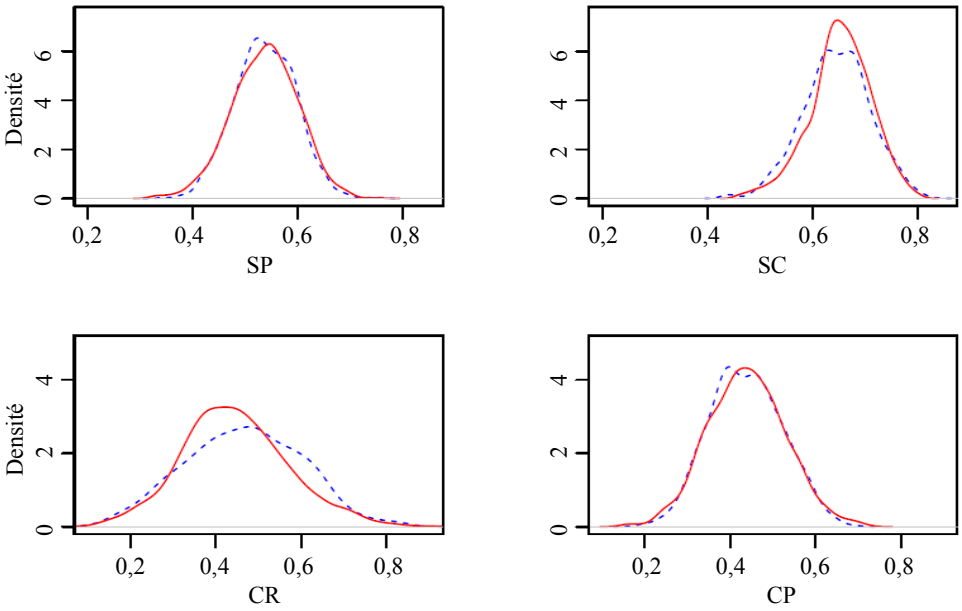

Figure 3.4 Courbes des densités a posteriori des proportions dans la population finie pour les notes de mathématique par domaine (SP, SC, CR, CP) (trait plein : modèle CHE, trait pointillé : modèle CHO).

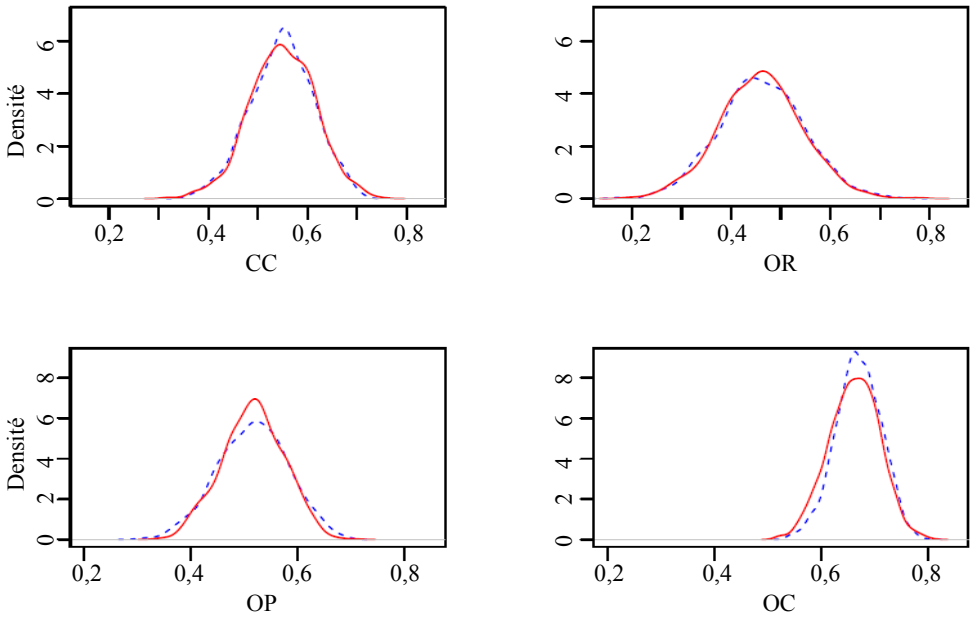

Figure 3.5 Courbes des densités a posteriori des proportions dans la population finie pour les notes de mathématique par domaine (CC, OR, OP, OC) (trait plein : modèle CHE, trait pointillé : modèle CHO).

3.2 Étude en simulation

Une étude en simulation nous permet de poursuivre l'évaluation de la performance du modèle CHE en vue de la comparer à celle du modèle CHO. Ici, nous utilisons deux facteurs, présentant chacun trois niveaux, pour obtenir neuf points de référence.

Nous avons fixé à 100 le nombre de grappes (écoles) dans chaque domaine et à 15 le nombre d'individus (élèves) dans chaque grappe. Autrement dit, nous prenons $N_{ij} = 15,\ j = 1,\ldots,M_i,\ M_i = 100,\ i = 1,\ldots,\ell$ où $\ell = 12$. Désignons par $\mathbf{a}$ un vecteur de moyennes a posteriori et par $\mathbf{b}$, le vecteur des écarts-types a posteriori correspondant aux μ_i ou aux ρ_i. Plus précisément, pour les ρ_i nous utilisons $\mathbf{a}_1$ et $\mathbf{b}_1$, et pour les μ_i nous utilisons $\mathbf{a}_2$ et $\mathbf{b}_2$. Quand nous simulons les données à partir du modèle CHE, les niveaux des ρ_i sont $(1:\mathbf{a}_1 - 0{,}5\mathbf{b}_1;\ 2:\mathbf{a}_1;\ 3:\mathbf{a}_1 + 0{,}5\mathbf{b}_1)$ et les niveaux des μ_i sont $(1:\mathbf{a}_2 - 0{,}5\mathbf{b}_2;\ 2:\mathbf{a}_2;\ 3:\mathbf{a}_2 + 0{,}5\mathbf{b}_2)$. Pour les douze domaines, $\mathbf{a}_1$ prend les valeurs 0,09; 0,19; 0,32; 0,08; 0,22; 0,18; 0,15; 0,22; 0,23; 0,17; 0,18; 0,30; $\mathbf{b}_1$ prend les valeurs 0,08; 0,09; 0,08; 0,06; 0,07; 0,08; 0,13; 0,09; 0,07; 0,09; 0,07; 0,06; $\mathbf{a}_2$ prend les valeurs 0,53; 0,37; 0,54; 0,58; 0,54; 0,65; 0,46; 0,44; 0,55; 0,46; 0,52; 0,66; et $\mathbf{b}_2$ prend les valeurs 0,08; 0,08; 0,08; 0,06; 0,06; 0,06; 0,12; 0,09; 0,07; 0,08; 0,06; 0,05.

Nous tirons aussi un échantillon aléatoire simple de cinq grappes parmi les 100 grappes de la population, ainsi qu'un échantillon aléatoire simple de dix individus dans chaque grappe échantillonnée (c'est-à-dire $m_i = 5$ et $n_{ij} = 10$). Ces nombres sont nettement plus faibles que ceux pour les données utilisées à la section 3.1, ce qui rend l'inférence un peu plus difficile (Nandram 2015). Notons que l'ensemble de données contient environ 7 % de grappes échantillonnées où tous les élèves étaient soit en dessous ou au-dessus de la moyenne. Nous nommons cette quantité le pourcentage de données éparses. La configuration de la

présente étude en simulation donne lieu à des données encore plus éparses. Pour neuf points de référence, tous les pourcentages moyens de données éparses sont supérieurs à 7 % et la plupart sont de l'ordre de 10 %. La figure 3.6 montre les histogrammes des pourcentages de données éparses pour chaque point de référence.

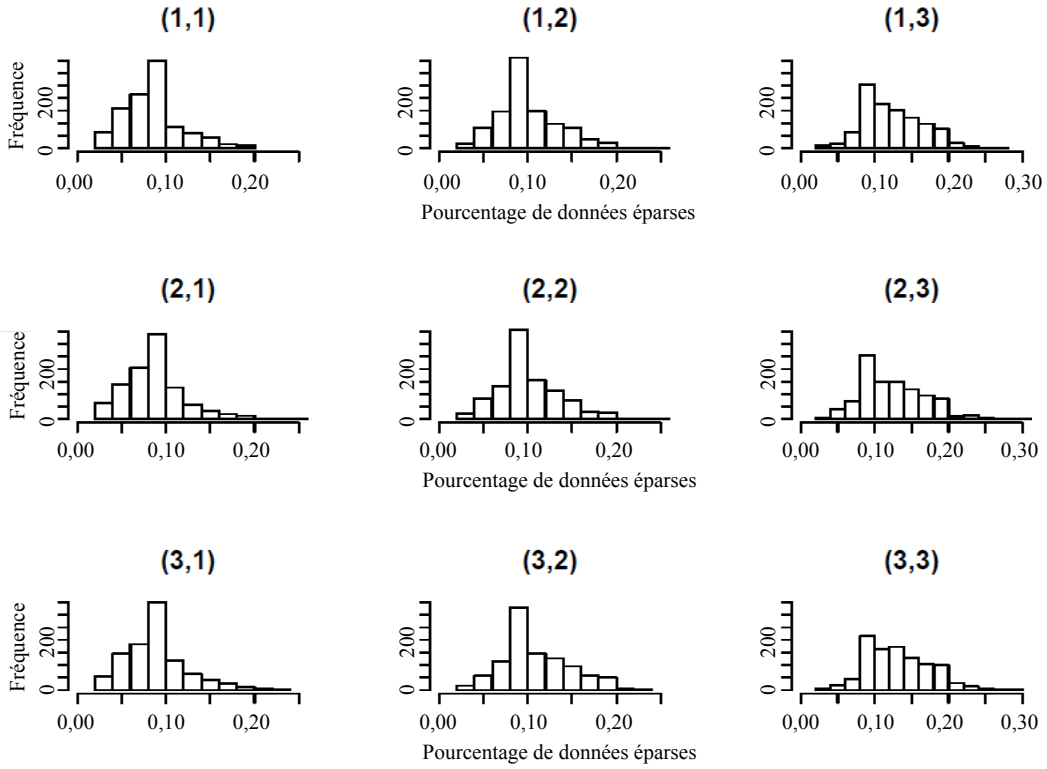

Figure 3.6 Histogrammes du pourcentage de données éparses quand les données sont tirées du modèle CHE par point de référence [(*i*, *j*) : *i*, *j* = 1, 2, 3] où le premier facteur correspond à ρ et le second, à μ.

Nous étudions deux scénarios. Dans le premier, nous générons des données à partir du modèle CHE et ajustons les deux modèles, et dans le second, nous générons des données à partir du modèle CHO et ajustons les deux modèles. Dans le cas des données simulées à partir du modèle CHE, nous avons neuf points de référence $[(1,1),(1,2),(1,3),\ldots,(3,1),(3,2),(3,3)]$, et le premier facteur correspond à ρ_i. Quand nous simulons les données à partir du modèle CHO, nous avons trois points de référence $(1:\mathbf{a}_2 - 0{,}5\mathbf{b}_2;\ 2:\mathbf{a}_2;\ 3:\mathbf{a}_2 + 0{,}5\mathbf{b}_2)$ pour les trois niveaux pour les μ_i; la valeur de ρ est maintenue fixe à sa moyenne a posteriori.

Dans le premier scénario, à chaque point de référence, nous simulons des données binaires à partir du modèle CHE,

$$p_{ij} \mid \mu_i, \rho_i \overset{\text{ind}}{\sim} \text{Bêta}\left[\mu_i \frac{1-\rho_i}{\rho_i}, (1-\mu_i)\frac{1-\rho_i}{\rho_i}\right],$$

$$y_{ijk} \mid p_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}\left(p_{ij}\right), \quad k = 1,\ldots,N_{ij}.$$

Donc, nous avons les vraies valeurs de $P_i = \sum_{j=1}^{M_i} \sum_{k=1}^{N_{ij}} y_{ijk} \Big/ \sum_{j=1}^{M_i} N_{ij}$ pour $i = 1, \ldots, \ell$. Nous tirons 1 000 échantillons à chacun des neuf points de référence. Pour chaque échantillon, nous exécutons l'échantillonneur de Gibbs « à grille » par blocs de la même façon que pour les données.

Comme Nandram (2015), nous calculons $\mathrm{BA}_{ih} = \left| \mathrm{MP}_{ih} - P_{ih} \right|$, $\mathrm{BAR}_{ih} = \mathrm{BA}_{ih} / P_{ih}$ et $\mathrm{REQMP}_{ih} = \sqrt{\mathrm{ETP}_{ih}^2 + \mathrm{BA}_{ih}^2}$ pour étudier les propriétés fréquentistes de notre procédure $(i = 1, \ldots, \ell,\ h = 1, \ldots, 1\,000)$. Nous obtenons aussi les intervalles de crédibilité et DPPG à 95 % pour chacune des 1 000 exécutions de la simulation, et nous étudions la largeur W_{ih} et l'incidence de crédibilité I_{ih}. Si l'intervalle de crédibilité (ou DPPG) à 95 % de la h^{e} exécution contient la valeur réelle P_i, I_{ih} est égale à un, sinon elle est nulle. Donc, le contenu probabiliste estimé de l'intervalle de crédibilité à 95 % pour le i^{e} domaine est $C_i = \sum_{h=1}^{1\,000} I_{ih} / 1\,000$.

Le tableau 3.4 donne la comparaison des modèles CHO et CHE. Les couvertures sont clairement plus élevées sous le modèle CHE que sous le modèle CHO. Notons que les couvertures des intervalles DPPG pour le modèle CHE sont nettement plus proches de la valeur nominale de 95 % et sont conservatrices. Cependant, les intervalles de crédibilité et DPPG à 95 % sont plus larges que sous le modèle CHO. Ces effets deviennent beaucoup plus importants à mesure que $\boldsymbol{\rho}$ augmente. Les mesures BA, BAR et REQMP sont toutes plus petites sous le modèle CHE que sous le modèle CHO. Donc, en s'appuyant sur ces mesures, la préférence est donnée au modèle CHE plutôt qu'au modèle CHO.

Tableau 3.4
Simulation sous le modèle CHE : Comparaison des modèles CHE et CHO en utilisant la couverture moyenne et la largeur des intervalles de crédibilité à 95 % et le biais absolu, le biais absolu relatif et la racine carrée de l'erreur quadratique moyenne a posteriori pour les proportions de la population finie par point de référence

Point de référence	Modèle	C-Cre	L-Cre	C-DPPG	L-DPPG	BA	BAR	REQMP
(1,1)	CHE	0,989	0,620	0,961	0,603	0,112	0,227	0,206
	CHO	0,930	0,555	0,893	0,541	0,130	0,266	0,207
(1,2)	CHE	0,984	0,622	0,960	0,603	0,112	0,227	0,206
	CHO	0,926	0,558	0,889	0,545	0,132	0,249	0,209
(1,3)	CHE	0,980	0,623	0,955	0,608	0 ,120	0,211	0,210
	CHO	0,923	0,558	0,892	0,546	0,134	0,236	0,212
(2,1)	CHE	0,982	0,621	0,953	0,603	0,119	0,242	0,212
	CHO	0,922	0,564	0,879	0,549	0,137	0,281	0,215
(2,2)	CHE	0,980	0,625	0,952	0,609	0,122	0,228	0,214
	CHO	0,918	0,566	0,879	0,552	0,139	0,264	0,217
(2,3)	CHE	0,981	0,628	0,956	0,611	0,121	0,211	0,214
	CHO	0,930	0,570	0,895	0,556	0,135	0,239	0,214
(3,1)	CHE	0,982	0,627	0,949	0,608	0,121	0,245	0,215
	CHO	0,934	0,583	0,892	0,566	0,136	0,278	0,218
(3,2)	CHE	0,980	0,628	0,947	0,610	0,123	0,242	0,217
	CHO	0,928	0,583	0,885	0,566	0,138	0,274	0,220
(3,3)	CHE	0,976	0,632	0,951	0,614	0,124	0,218	0,218
	CHO	0,928	0,581	0,889	0,565	0,139	0,246	0,220

Nota : Dans le point de référence [(i, j) : i, j = 1, 2, 3], le premier facteur correspond à ρ et le second, à μ. C-Cre et C-DPPG sont les contenus probabilistes d'un intervalle de crédibilité et d'un intervalle DPPG; L-Cre et L-DPPG sont les largeurs d'un intervalle de crédibilité et d'un intervalle DPPG. BA, BAR et REQMP sont le biais absolu, le biais absolu relatif et la racine carrée de l'erreur quadratique moyenne a posteriori.

Dans le tableau 3.5, nous comparons les données sommaires pour les critères DIC, BPP et LPML. Toutes les valeurs de DIC sous le modèle CHE sont plus faibles que les valeurs correspondantes sous le modèle CHO, et toutes les valeurs de LPML sous le modèle CHE sont plus grandes que celles sous le modèle CHO.

Sous le modèle CHO, toutes les valeurs de BPP varient dans l'intervalle (0,06; 0,09), mais sous le modèle CHE, elles varient dans l'intervalle (0,2; 0,4). De nouveau, ces mesures montrent que le modèle CHE donne de meilleurs résultats que le modèle CHO.

De façon similaire, pour le deuxième scénario, nous générons des données binaires à partir de

$$p_{ij} \mid \mu_i, \rho \overset{\text{ind}}{\sim} \text{Bêta}\left[\mu_i \frac{1-\rho}{\rho}, (1-\mu_i)\frac{1-\rho}{\rho}\right],$$

$$y_{ijk} \mid p_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}\left(p_{ij}\right), \quad k = 1, \ldots, N_{ij}.$$

Dans le tableau 3.6, nous comparons les modèles CHO et CHE. Ici, les critères BA, BAR et REQMP ne sont que légèrement plus faibles sous le modèle CHO. Les couvertures des intervalles de crédibilité et DPPG sous le modèle CHE s'approchent davantage de la valeur nominale de 95 %, tandis que celles sous le modèle CHO sont plus petites. Le tableau 3.7 donne les données sommaires pour les critères DIC, BPP et LPML. Toutes les valeurs de DIC sous le modèle CHE sont plus petites que sous le modèle CHO, tandis que les valeurs de BPP et de LPML sont similaires pour les deux modèles, celles obtenues sous le modèle CHO étant légèrement meilleures.

Tableau 3.5
Simulation sous le modèle CHE : Comparaison des modèles CHE et CHO en utilisant le critère d'information de déviance (DIC), la valeur *p* prédictive bayésienne (BPP) et le logarithme de la pseudo-vraisemblance marginale (LPML) par point de référence

Point de référence	Modèle CHO			Modèle CHE		
	DIC	BPP	LPML	DIC	BPP	LPML
(1,1)	419,275	0,090	-285,452	402,044	0,429	-267,990
(1,2)	418,351	0,091	-286,250	400,647	0,439	-266,377
(1,3)	416,784	0,088	-286,290	400,414	0,446	-267,203
(2,1)	436,980	0,067	-307,028	416,264	0,300	-292,756
(2,2)	437,306	0,062	-308,816	414,955	0,318	-292,404
(2,3)	430,531	0,080	-302,258	410,436	0,351	-285,206
(3,1)	441,204	0,090	-316,126	424,010	0,227	-308,825
(3,2)	442,165	0,083	-318,223	424,363	0,235	-309,815
(3,3)	438,305	0,071	-315,159	418,827	0,260	-306,619

Nota : Dans le point de référence $[(i, j) : i, j = 1, 2, 3]$, le premier facteur correspond à ρ et le second, à μ. Les valeurs sommaires de DIC, BPP et LPML sont calculées sur les 1 000 exécutions de la simulation.

Tableau 3.6
Simulation sous le modèle CHO : Comparaison des modèles CHE et CHO en utilisant la couverture moyenne et la largeur des intervalles de crédibilité à 95 % et le biais absolu, le biais absolu relatif et la racine carrée de l'erreur quadratique moyenne a posteriori pour les proportions de la population finie par point de référence

Point de référence	Modèle	C-Cre	L-Cre	C-DPPG	L-DPPG	BA	BAR	REQMP
1	CHE	0,985	0,627	0,969	0,608	0,117	0,242	0,212
	CHO	0,944	0,575	0,919	0,559	0,107	0,240	0,210
2	CHE	0,988	0,634	0,952	0,616	0,122	0,234	0,216
	CHO	0,938	0,585	0,917	0,568	0,115	0,214	0,211
3	CHE	0,977	0,628	0,940	0,611	0,126	0,222	0,218
	CHO	0,933	0,572	0,908	0,556	0,113	0,202	0,208

Nota : Le point de référence $[i : i = 1, 2, 3]$ correspond à μ avec ρ maintenue fixe à sa moyenne a posteriori. C-Cre et C-DPPG sont les contenus probabilistes d'un intervalle de crédibilité et d'un intervalle DPPG; L-Cre et L-DPPG sont les largeurs d'un intervalle de crédibilité et d'un intervalle DPPG. BA, BAR et REQMP sont le biais absolu, le biais absolu relatif et la racine carrée de l'erreur quadratique moyenne a posteriori.

Tableau 3.7
Simulation sous le modèle CHO : Comparaison des modèles CHE et CHO en utilisant le critère d'information de déviance (DIC), la valeur *p* prédictive bayésienne (BPP) et le logarithme de la pseudo-vraisemblance marginale (LPML) par point de référence

Point de référence	Modèle CHO			Modèle CHE		
	DIC	BPP	LPML	DIC	BPP	LPML
1	428,647	0,308	-300,526	416,626	0,302	-303,001
2	430,113	0,371	-295,191	417,557	0,317	-296,531
3	429,598	0,379	-295,613	414,877	0,335	-297,250

Nota : Le point de référence [i : i = 1, 2, 3] correspond à μ avec ρ maintenue fixe à sa moyenne a posteriori. Les valeurs sommaires de DIC, BPP et LPML sont calculées sur les 1 000 exécutions de la simulation.

Donc, quand les données sont effectivement issues du modèle CHE, nous constatons certaines différences importantes entre les deux modèles, la préférence étant donnée au modèle CHE. Par contre, quand les données proviennent du modèle CHO, les différences constatées entre les deux modèles sont mineures. Bien entendu, le modèle CHE (corrélations inégales) contient plus de paramètres que le modèle CHO (une corrélation).

4 Conclusion

Afin d'ajouter un degré de flexibilité à nos analyses de données, nous avons étendu un modèle double homogène, décrit dans Nandram (2015), à un modèle double hétérogène. Ces modèles contiennent des paramètres faiblement identifiés qui posent de sérieux problèmes de calcul. Par conséquent, nous avons fait deux autres ajouts. Premièrement, nous avons introduit une contrainte unimodale sur les distributions bêta a priori. Deuxièmement, nous avons utilisé un échantillonneur de Gibbs par blocs pour effectuer les calculs. Pour comparer les modèles, nous avons procédé à une inférence bayésienne prédictive. À titre d'exemple, nous avons utilisé des données provenant de la TIMSS, une étude de la performance en mathématique des élèves américains de troisième année. En outre, nous avons effectué une étude en simulation pour comparer les deux modèles doubles.

Il est important de se servir du modèle hétérogène pour modéliser le plan d'échantillonnage double, car dans de nombreuses applications, les corrélations intragrappe peuvent varier d'un domaine à l'autre, ce qui rend ce modèle plus approprié que le modèle double homogène. En effet, à l'aide d'un exemple et d'une étude en simulation avec application de plusieurs critères diagnostiques, nous avons montré qu'il convient de préférer le modèle double hétérogène au modèle double homogène quand les corrélations varient considérablement.

Nos travaux peuvent s'étendre afin de prendre en compte des données binaires multivariées. Cela peut se concevoir comme un problème de groupement de données provenant de distributions multinomiales pour faire des inférences sur des proportions de la population finie. Par exemple, dans le cas de la TIMSS, nous pouvons utiliser les notes de mathématique et de science en tant que réponses binaires bivariées (corrélation). Nous pouvons alors élaborer un modèle hiérarchique bayésien pour les réponses multinomiales et une distribution de Dirichlet a priori pour modéliser les probabilités dans les cellules. Dans le cadre de cette étude, nous pouvons nous attaquer à deux questions. D'abord, nous pouvons examiner dans quelle mesure la prédiction sera améliorée si l'on utilise les données multivariées. Nous pouvons également

étudier dans quelle mesure la précision de l'inférence augmentera si l'on privilégie un modèle avec corrélations intragrappe hétérogènes plutôt qu'un modèle avec corrélation homogène en ce qui a trait aux données multivariées.

Remerciements

Les auteurs remercient les deux examinateurs pour leur lecture attentive du manuscrit et leurs suggestions. Leurs travaux de recherche ont bénéficié du soutien financier du *Basic Science Research Program* par l'entremise de la *National Research Foundation of Korea* (NRF) financée par le ministère de l'Éducation (NRF-2014R1A1A2058954). Les travaux ont également été financés par une bourse de la *Simons Foundation* (#353953, Balgobin Nandram).

Annexe A

Preuves des formules (2.12) et (2.13)

Il est facile de montrer que

$$\begin{aligned}\mathrm{Cov}\left(y_{ijk}, y_{ijk'} \,\middle|\, \mu_i, \gamma, \rho_i\right) &= \mathrm{Var}\left(p_{ij} \,\middle|\, \mu_i, \gamma, \rho_i\right) = \mu_i(1-\mu_i)\rho_i,\\ \mathrm{Var}\left(y_{ijk} \,\middle|\, \mu_i, \gamma, \rho_i\right) &= E\left[\mathrm{Var}\left(y_{ijk} \,\middle|\, p_{ij}, \mu_i, \gamma, \rho_i\right)\right] + \mathrm{Var}\left[E\left(y_{ijk} \,\middle|\, p_{ij}, \mu_i, \gamma, \rho_i\right)\right],\\ &= E\left(p_{ij} \,\middle|\, \mu_i, \gamma, \rho_i\right)\left[1 - E\left(p_{ij} \,\middle|\, \mu_i, \gamma, \rho_i\right)\right] = \mu_i(1-\mu_i).\end{aligned}$$

Donc, $\mathrm{Cor}\left(y_{ijk}, y_{ijk'} \,\middle|\, \mu_i, \gamma, \rho_i\right) = \rho_i \,(k \neq k')$, ce qui prouve (2.12).

De même, il est facile de montrer que

$$\begin{aligned}\mathrm{Cov}\left(y_{ijk}, y_{ij'k'} \,\middle|\, \theta, \gamma, \rho_i\right) &= E\left[\mathrm{Cov}\left(p_{ij}, p_{ij'} \,\middle|\, \mu_i, \theta, \gamma, \rho_i\right)\right]\\ &\quad + \mathrm{Cov}\left[E\left(p_{ij} \,\middle|\, \mu_i, \theta, \gamma, \rho_i\right), E\left(p_{ij'} \,\middle|\, \mu_i, \theta, \gamma, \rho_i\right)\right]\\ &= \mathrm{Var}\left(\mu_i \,\middle|\, \theta, \gamma\right) = \theta(1-\theta)\gamma,\\ \mathrm{Var}\left(y_{ijk} \,\middle|\, \theta, \gamma, \rho_i\right) &= E\left(\mu_i \,\middle|\, \theta, \gamma\right) - \left[E\left(\mu_i \,\middle|\, \theta, \gamma\right)\right]^2 = \theta(1-\theta).\end{aligned}$$

Donc, $\mathrm{Cor}\left(y_{ijk}, y_{ij'k'} \,\middle|\, \theta, \gamma, \rho_i\right) = \gamma \,(j \neq j', k \neq k')$, ce qui prouve (2.13).

Annexe B

Calculs avec contraintes d'unimodalité

Il est bien connu qu'une densité de probabilité bêta de paramètres α et β est unimodale si $\alpha > 1$ et $\beta > 1$. Cela peut s'établir facilement en faisant appel au calcul infinitésimal. Dans notre cas, $\mu \,|\, \theta, \gamma \sim \text{Bêta}\left\{\theta\frac{(1-\gamma)}{\gamma}, (1-\theta)\frac{(1-\gamma)}{\gamma}\right\}$. Donc, nous avons deux inégalités,

$$\theta\frac{(1-\gamma)}{\gamma} > 1 \text{ et } (1-\theta)\frac{(1-\gamma)}{\gamma} > 1,$$

et des calculs algébriques simples donnent

$$\frac{\gamma}{1-\gamma} < \theta < \frac{1-2\gamma}{1-\gamma}, \quad 0 < \gamma < \frac{1}{3}.$$

Ensuite, décrivons brièvement la façon d'appliquer ces contraintes aux calculs dans le modèle double avec corrélations hétérogènes. Rappelons la distribution marginale conditionnelle a posteriori de γ,

$$p(\gamma\,|\boldsymbol{\rho}, \phi, \delta, \mathbf{y}) \approx \sum_{g=1}^{G} \omega_g\, p\left(x_g, \gamma\,|\boldsymbol{\rho}, \phi, \delta, \mathbf{y}\right),$$

où $\{\omega_g\}$ sont les poids et $\{x_g\}$ sont les racines du polynôme de Legendre. Ici, nous utilisons la méthode à grille univariée pour échantillonner γ. Donc, nous divisons l'intervalle $(0, \frac{1}{3})$, la première contrainte, en G1 sous-intervalles $[\gamma_0, \gamma_1), [\gamma_1, \gamma_2), \ldots, [\gamma_{G1-1}, \gamma_{G1}]$. Pour un nombre aléatoire uniforme, u^*, provenant de toute grille, disons, $[\gamma_{\nu-1}, \gamma_\nu)$, nous calculons la hauteur, c'est-à-dire la valeur de la fonction de densité marginale conditionnelle a posteriori de γ sous la forme

$$\frac{1-3u^*}{1-u^*}\sum_{g=1}^{G^*} \omega_g^*\, p\left(x_g^*, u^*\,|\boldsymbol{\rho}, \phi, \delta, \mathbf{y}\right),$$

où $\{\omega_g^*\}$ sont les poids et $\{x_g^*\}$, les racines du polynôme de Legendre sur l'intervalle $[\frac{u^*}{1-u^*}, \frac{1-2u^*}{1-u^*}]$, la deuxième contrainte. De même, nous pouvons appliquer le critère d'unimodalité à l'échantillon (ϕ, δ).

Bibliographie

Brier, S.S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika,* 67, 591-595.

Caslyn, C., Gonzales, P. et Frase, M. (1999). Highlights from TIMSS. *National Center for Education Statistics,* Washington, DC.

Damien, P., Laud, P.W. et Smith, A.F.M. (1997). Bayesian estimation of unimodal distributions. *Communications in Statistics,* 26(2), 429-440.

Foy, P., Rust, K. et Schleicher, A. (1996). Sample design. Dans *TIMSS Technical Report, Volume I: Design and Development,* (Éds., M.O. Martin et D.L. Kelly), Chestnut Hill, MA: Boston College.

Fuller, W.A., et Battese, G.E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association,* 68, 626-632.

Ghosh, M., et Lahiri, P. (1988). Bayes and empirical Bayes analysis in multistage sampling. Dans *Statistical Decision Theory and Related Topics IV,* (Éds., S.S. Gupta et J.O. Berger), New York: Springer, Vol. 1, 195-212.

Molina, I., Nandram, B. et Rao, J.N.K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *Annals of Applied Statistics,* 8(2), 852-885.

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

Nandram, B. (2015). Bayesian predictive inference of a proportion under a two-fold small area model. *Journal of Official Statistics* (accepté).

Nandram, B., et Sedransk, J. (1993). Bayesian predictive inference for a finite population proportion: Two-stage cluster sampling. *Journal of the Royal Statistical Society, Series B,* 55, 399-408.

Nandram, B., Bhatta, D., Sedransk, J. et Bhadra, B. (2013). A Bayesian test of independence in a two-way contingency table using surrogate sampling. *Journal of Statistical Planning and Inference,* 143, 1392-1408.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. New Jersey: Wiley, Hoboken.

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

Rao, J.N.K., et Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association,* 76, 221-230.

Rao, J.N.K., et Scott, A.J. (1984). On chi-squared tests for multi-way tables with cell proportions estimated from survey data. *Annals of Statistics,* 12, 46-60.

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

Scott, A., et Smith, T.M.F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association,* 101, 1387-1397.

Stukel, D.M., et Rao, J.N.K. (1997). Estimation of regression models with nested error regression structure and unequal variances under two and three stage cluster sampling. *Statistics & Probability Letters,* 35, 401-407.

Stukel, D.M., et Rao, J.N.K. (1999). On small-area estimation under two-fold nested error regression models. *Journal of Statistical Planning and Inference,* 78, 131-147.

Toto, M.C.S., et Nandram, B. (2010). A Bayesian predictive inference for small area means incorporating covariates and sampling weights. *Journal of Statistical Planning and Inference,* 140, 2963-2979.

Techniques d'enquête, juin 2017
Vol. 43, nº 1, pp. 103-117
Statistique Canada, nº 12-001-X au catalogue

Répartition de l'échantillon pour une estimation efficace sur petits domaines par modélisation

Mauno Keto et Erkki Pahkinen[1]

Résumé

Nous présentons les résultats de notre recherche sur les modes de répartition d'échantillons qui permettent de faire une estimation efficace sur petits domaines par modélisation dans les cas où les domaines d'intérêt coïncident avec les strates. Les méthodes d'estimation assistées d'un modèle et celles fondées sur un modèle sont répandues dans la production de statistiques relatives aux petits domaines, mais l'utilisation du modèle et de la méthode d'estimation sous-jacents est rarement intégrée au plan de répartition de l'échantillon entre les domaines. C'est pourquoi nous avons conçu un nouveau mode de répartition fondée sur un modèle que nous avons appelé répartition *g*1. Aux fins de comparaison, nous décrivons un autre mode de répartition fondée sur un modèle qui a récemment vu le jour. Ces deux répartitions sont fondées sur une mesure ajustée de l'homogénéité qui se calcule à l'aide d'une variable auxiliaire et constitue une approximation de la corrélation intraclasse à l'intérieur des domaines. Nous avons choisi cinq solutions de répartition par domaine sans modèle, adoptées par le passé dans le cadre d'études spécialisées, comme méthodes de référence. Pour une répartition égale ou proportionnelle, il nous faut connaître le nombre de domaines ainsi que le nombre d'unités statistiques de base dans chacun d'eux. Les répartitions de Neyman et de Bankier et la répartition par programmation non linéaire (PNL), nécessitent des paramètres au niveau du domaine comme l'écart-type, le coefficient de variation ou les totaux. En règle générale, on peut caractériser les méthodes de répartition en fonction des critères d'optimisation et de l'utilisation de données auxiliaires. On évalue alors les propriétés statistiques des diverses méthodes retenues au moyen d'expériences de simulation d'échantillon faisant appel aux données réelles du registre de population. Selon les résultats de simulation, on peut conclure que l'intégration du modèle et de la méthode d'estimation à la méthode de répartition a pour effet d'améliorer les résultats de l'estimation.

Mots-clés : Taille d'échantillon optimale de domaine; critères; information auxiliaire; mesure d'homogénéité.

1 Introduction

Dans le présent document, nous exposons une nouvelle méthode de répartition fondée sur un modèle dans un échantillonnage stratifié où les domaines d'intérêt coïncident avec les strates. L'étude cible les éléments d'une répartition par domaine efficace. Nous parvenons à un point de départ clair pour le processus de répartition si les domaines d'intérêt sont définis dès l'étape de conception de la recherche et si nous savons déjà quel est l'ordre de grandeur de l'échantillon compte tenu des ressources disponibles (temps, budget, etc.). Le choix d'un mode de répartition dépend de divers facteurs comme le modèle choisi, la méthode d'estimation, l'information antérieure disponible sur la population et les critères d'optimisation relatifs au domaine, à la population, ou aux deux en même temps.

Nous avons choisi six méthodes de répartition existantes, puis nous en avons élaboré une nouvelle que nous avons qualifiée de mode de répartition fondée sur un modèle. Nous examinons les propriétés générales de ces méthodes à la section 2 ainsi qu'à la section 3. On peut considérer que cinq d'entre elles sont sans modèle; deux utilisent uniquement des renseignements de dénombrement (par exemple, le nombre de domaines et le nombre d'unités de base dans chacun des domaines); trois exigent non seulement des données

1. Mauno Keto, Université de Jyväskylä. Courriel : mauno.j.keto@student.jyu.fi; Erkki Pahkinen, Département de mathématiques et de statistique de l'Université de Jyväskylä. Courriel : pahkinen@maths.jyu.fi.

de dénombrement, mais aussi des paramètres de domaine comme les totaux, les écarts-types ou les coefficients de variation (CV). Comme nous ne disposons pas de ces renseignements sur la variable étudiée, une solution courante est de remplacer l'information par une variable de substitution pertinente. Le dernier des modes de répartition nous servant de référence vient de Molefe et Clark (MC) (2015). Cette répartition fondée sur un modèle recourt à un estimateur composite et à un modèle à deux niveaux. Nous l'avons appelée répartition MC.

Les critères d'optimisation des cinq modes de répartition sans modèle diffèrent les uns des autres. Il est facile de calculer une répartition qui se fait uniquement au moyen de données de dénombrement, mais ce choix n'est pertinent que dans des circonstances restreintes. Le critère d'optimisation est particulier dans chacun des modes de répartition fondée sur des paramètres. Il peut se situer au niveau des estimations de population (répartition de Neyman) ou à celui des estimations de domaine en moyenne (répartition de Bankier). Une troisième solution qui s'écarte des deux premières est la répartition PNL, où les tolérances des estimations s'établissent au niveau de la population ainsi qu'à celui du domaine.

Au départ, nous posons l'hypothèse que, si on recourt à une estimation assistée d'un modèle ou fondée sur un modèle dans le cadre d'une enquête, on se doit de tenir compte du modèle et de la méthode d'estimation sous-jacents au moment de concevoir la répartition de l'échantillon entre les domaines. Cela a été notre point de départ lorsque nous avons mis au point le nouveau mode de répartition $g1$ fondée sur un modèle (voir la section 2). Il faut aussi dire qu'un des modes de répartition utilisés comme référence, fondé sur un modèle, repose sur un modèle déterminé.

Nous avons soumis à une comparaison de rendement les différentes méthodes de répartition en situation réelle en procédant à des expériences de simulation (voir la section 4). Notre population consiste en un registre finnois officiel de logements d'immeubles d'appartements en vente. Nous décrivons la structure de ce registre à la section 4.1. Nous avons remplacé la variable étudiée par une variable auxiliaire au moment de calculer les tailles d'échantillon de domaine pour chaque répartition sauf en ce qui concerne la répartition égale ou proportionnelle. La comparaison démontre clairement que les différentes répartitions donnent une distribution différente de l'échantillon. Des différences s'observent aussi sur le plan du rendement. Nous avons soumis les modes de répartition à une estimation EBLUP (pour *Empirical best linear unbiased predictor*) fondée sur un modèle pour dégager les totaux de domaine de la variable étudiée. Pour mesurer et comparer le rendement des répartitions, nous avons employé la racine carrée de l'erreur quadratique moyenne relative (REQMR) et le biais relatif absolu (BRA) en pourcentage.

À la section 5, nous concluons en examinant les résultats empiriques de simulation. Ces résultats confortent l'idée que, dans un mode de répartition, on devrait déterminer non seulement l'information auxiliaire, mais aussi le modèle et la méthode d'estimation dès l'étape de la conception d'une enquête. Une bonne illustration est la répartition $g1$ présentée à la section 2.2. C'est au moyen de cette méthode que nous avons obtenu les estimations les plus fidèles des totaux de domaine.

2 Répartitions par modélisation

2.1 Choix du modèle

Pfeffermann (2013) présente une grande variété de modèles et de méthodes pour l'estimation de petits domaines. Notre modèle constitue l'un de cette collection, soit un modèle mixte au niveau de l'unité.

$$y_{dk} = \mathbf{x}'_{dk}\boldsymbol{\beta} + v_d + e_{dk};\ \ k = 1,\ldots,N_d;\ \ d = 1,\ldots,D, \tag{2.1}$$

où les v_d sont des effets aléatoires de domaine à moyenne zéro et à variance σ_v^2 et où les e_{dk} sont des effets aléatoires à moyenne zéro et à variance σ_e^2. De plus, $E(y_{dk}) = \mathbf{x}'_{dk}\boldsymbol{\beta}$ et $V(y_{dk}) = \sigma_v^2 + \sigma_e^2$ (variance totale). La matrice $\mathbf{V}$ est une matrice des variances-covariances de la variable étudiée y. Le modèle peut être employé quand on dispose de valeurs au niveau de l'unité pour les variables auxiliaires $\mathbf{x}$. Dans notre étude, nous utilisons une seule variable auxiliaire.

Il nous faut deux mesures importantes pour réaliser un de ces types de répartition, soit la corrélation intradomaine commune ρ et le rapport δ entre les composantes de la variance. Le tout se définit ainsi :

$$\rho = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2) \quad \text{et} \quad \delta = \sigma_e^2 / \sigma_v^2 = 1/\rho - 1. \tag{2.2}$$

Avant d'estimer les paramètres de domaine, nous devons estimer les composantes de la variance, les coefficients de régression et les effets de domaine à partir des données de l'échantillon. Nous obtenons l'estimateur BLUE (pour Best Linear Unbiased Estimator) de $\boldsymbol{\beta}$, noté $\tilde{\boldsymbol{\beta}}$, selon la théorie du modèle linéaire généralisé, et nous le remplaçons par sa contrepartie EBLUP (meilleur prédicteur linéaire sans biais empirique) $\hat{\boldsymbol{\beta}}$.

L'estimation EBLUP (valeur prévue) du total de domaine Y_d de la variable étudiée est la somme des valeurs y observées et des valeurs y estimées pour les unités hors échantillon :

$$\hat{Y}_{d,\text{Eblup}} = \sum_{k\in s_d} y_{dk} + \sum_{k\in \overline{s}_d} \hat{y}_{dk} = \sum_{k\in s_d} y_{dk} + \sum_{k\in \overline{s}_d} \mathbf{x}'_{dk}\hat{\boldsymbol{\beta}} + (N_d - n_d)\hat{v}_d. \tag{2.3}$$

Nous utilisons l'approximation de Prasad-Rao (voir Rao 2003) de l'erreur quadratique moyenne (EQM) pour des populations finies :

$$\text{eqm}\left(\hat{Y}_{d,\text{Eblup}}\right) = g_{1d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) + g_{2d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) + 2g_{3d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) + g_{4d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right), \tag{2.4}$$

où les quatre composantes g_{1d}, g_{2d}, g_{3d} et g_{4d} se définissent de la manière suivante :

$$\begin{aligned}
g_{1d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) &= \left(N_d - n_d^*\right)^2 \left(1 - \hat{\gamma}_d\right)\hat{\sigma}_v^2, \\
g_{2d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) &= \left(N_d - n_d^*\right)^2 \left(\overline{\mathbf{x}}_d^* - \hat{\gamma}_d \overline{\mathbf{x}}_d\right)' \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \left(\overline{\mathbf{x}}_d^* - \hat{\gamma}_d \overline{\mathbf{x}}_d\right), \\
g_{3d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) &= \left(N_d - n_d^*\right)^2 \left(n_d^*\right)^{-2} \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \left(n_d^*\right)^{-1}\right)^{-3} \left[\hat{\sigma}_e^4 V\left(\hat{\sigma}_v^2\right)\right. \\
&\qquad \left. + \hat{\sigma}_v^4 V\left(\hat{\sigma}_e^2\right) - 2\hat{\sigma}_e^2 \hat{\sigma}_v^2 \,\text{Cov}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)\right], \\
g_{4d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) &= \left(N_d - n_d^*\right)\hat{\sigma}_e^2.
\end{aligned} \tag{2.5}$$

Les tailles d'échantillon de domaine n_d^* dépendent de l'échantillon et ne sont pas fixes. La composante g_{1d} contient le rapport spécifique de domaine $\hat{\gamma}_d = \hat{\sigma}_v^2 / \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 / n_d^*\right)$. D'après Nissinen (2009, page 53), la composante g_{1d} (que nous appellerons simplement $g1$) contribue généralement pour plus de 90 % à l'EQM estimée. Elle représente l'incertitude quant à la variation entre les domaines. Bien sûr, la variation doit être assez ample pour que la proportion soit forte pour $g1$.

Malheureusement, le choix d'une solution analytique, où on minimise la somme des EQM sur les domaines sous réserve de $n = \sum_{d=1}^{D} n_d$, est difficile et longue à réaliser, puisque les composantes de l'approximation de l'EQM (2.5) comprennent l'information inconnue de l'échantillon et que certaines composantes consistent en un traitement sur matrice complexe et en des opérations sur variances-covariances. Nous avons examiné ce problème de répartition pour la première fois dans une étude expérimentale (Keto et Pahkinen 2009). Nous avons élaboré un mode de répartition reposant seulement sur la composante $g1$ et la variable auxiliaire x. Cette solution se justifie parce que x et y sont en corrélation et que la variation interdomaine dans x est transférée à y.

2.2 Répartition $g1$ fondée sur un modèle

La répartition $g1$ se fait au moyen de la variable auxiliaire x et le coefficient ajusté d'homogénéité (Keto et Pahkinen 2014). Ce coefficient est une approximation d'une corrélation intraclasse (CIC) connue dans l'échantillonnage en grappes. Dans le présent cas, nous considérons un domaine comme une grappe. D'abord, nous faisons une analyse de variance classique (ANOVA) entre domaines, ce qui nous amène à calculer une mesure ajustée d'homogénéité de la variation entre domaines :

$$R_{ax}^2 = 1 - R^2(x) = 1 - \mathrm{CMI} / S_x^2, \tag{2.6}$$

où $R^2(x)$ est le coefficient de détermination de l'analyse de régression, où CMI (carré moyen intragroupe) est la moyenne de la somme des carrés pour les domaines et où S_x^2 est la variance de la variable auxiliaire x.

Comme l'EQM du total de domaine est complexe, nous employons seulement la composante $g1$ en (2.4) et (2.5), pour la raison mentionnée à la section 2.1. Nous recherchons le minimum pour la somme des $g1$ sur les domaines :

$$\sum_{d=1}^{D} g_{1d}\left(\sigma_v^2, \sigma_e^2\right) = \sum_{d=1}^{D} \left(N_d - n_d\right)^2 \left(n_d / \sigma_e^2 + 1/\sigma_v^2\right)^{-1} \tag{2.7}$$

sous réserve de $n = \sum_{d=1}^{D} n_d$.

Nous prenons la méthode des multiplicateurs de Lagrange pour dégager la solution. Nous définissons donc la fonction F des tailles d'échantillon $\mathbf{n}' = \left(n_1, n_2, \ldots, n_D\right)$ et de λ :

$$F(\mathbf{n}, \lambda) = \sum_{d=1}^{D} g_{1d}\left(\sigma_v^2, \sigma_e^2\right) = \sum_{d=1}^{D} \left(N_d - n_d\right)^2 \left(n_d / \sigma_e^2 + 1/\sigma_v^2\right)^{-1} + \lambda\left(\sum_{d=1}^{D} n_d - n\right). \tag{2.8}$$

Nous fixons à zéro la dérivée de F par rapport à la taille d'échantillon de domaine n_d et résolvons pour n_d. Voici la formule qui s'applique à la taille d'échantillon de domaine n_d^{g1} :

$$n_d^{g1} = \frac{(N_d + \delta)(n + \delta D)}{N + \delta D} - \delta = \frac{N_d n - (N - N_d D - n)(1/\rho - 1)}{N + D(1/\rho - 1)}, \tag{2.9}$$

où le rapport δ et la corrélation intradomaine ρ sont définis en (2.2). Le seul membre inconnu en (2.9) est la corrélation intradomaine ρ. Nous remplaçons donc ρ par la mesure d'homogénéité connue (2.6) de la variable auxiliaire x. Voici la forme que prend finalement l'expression dans le calcul des tailles d'échantillon de domaine :

$$n_d^{g1} = \frac{N_d n - (N - N_d D - n)(1/R_{ax}^2 - 1)}{N + D(1/R_{ax}^2 - 1)}. \tag{2.10}$$

Il est facile de démontrer que $\sum_{d=1}^{D} n_d^{g1} = n$. Les tailles d'échantillon calculées sont arrondies à l'entier le plus proche. Parfois, des compromis sont inévitables. De l'examen de (2.10), nous pouvons conclure que la taille d'échantillon augmente avec la taille de domaine N_d, mais non proportionnellement. Dans certains cas comme lorsque le coefficient d'homogénéité, la taille d'échantillon globale n ou la taille de domaine N_d est faible, le calcul peut donner une valeur négative de taille d'échantillon de domaine n_d^{g1}. Nous remplaçons alors la valeur négative par une valeur nulle. Un cas d'espèce se présente si la variation totale intervient seulement entre les domaines, auquel cas la mesure d'homogénéité (2.6) se ramène à l'unité et (2.10), à la répartition proportionnelle.

2.3 Répartition MC assistée d'un modèle

Molefe et Clark (2015) ont utilisé l'estimateur composite suivant de la moyenne de la variable étudiée y pour le domaine d :

$$\tilde{y}_d^C = (1 - \varphi_d)\,\bar{y}_{dr} + \varphi_d\,\hat{\boldsymbol{\beta}}'\bar{\mathbf{X}}_d. \tag{2.11}$$

Cet estimateur en combine deux, à savoir l'estimateur synthétique $\hat{\bar{Y}}_{d(\text{syn})} = \hat{\boldsymbol{\beta}}'\bar{\mathbf{X}}_d$, où $\hat{\boldsymbol{\beta}}$ est le coefficient de régression estimé et $\bar{\mathbf{X}}_d$, la moyenne de population de domaines des variables auxiliaires $\mathbf{x}$, et l'estimateur direct $\bar{y}_{dr} = \bar{y}_d + \hat{\boldsymbol{\beta}}'(\bar{\mathbf{x}}_d - \bar{\mathbf{X}}_d)$, où $\bar{y}_d$ et $\bar{\mathbf{x}}_d$ sont les moyennes d'échantillon de domaines d pour y et $\mathbf{x}$. Nous employons une seule variable auxiliaire dans notre étude. Les coefficients φ_d visent à minimiser l'EQM de l'estimateur (2.11). L'EQM approchée basée sur le plan pour l'estimateur dans certaines conditions et hypothèses est donnée par l'expression

$$\text{EQM}_p\left(\tilde{y}_d^C; \bar{Y}_d\right) \approx (1 - \varphi_d)^2\, v_{d(\text{syn})} + \varphi_d^2 B_d^2, \tag{2.12}$$

où $v_{d(\text{syn})}$ est la variance d'échantillonnage de l'estimateur synthétique $\hat{\bar{Y}}_{d(\text{syn})}$ et où $B_d = \boldsymbol{\beta}_U'\,\bar{\mathbf{X}}_d - \bar{Y}_d$ est le biais lorsque $\hat{\bar{Y}}_{d(\text{syn})}$ sert à estimer $\bar{Y}_d$, $\boldsymbol{\beta}_U$ désignant l'espérance approchée basée sur le plan de $\hat{\boldsymbol{\beta}}$.

La population contient N unités et D strates définies comme domaines, et on emploie un échantillonnage stratifié. Nous prélevons un échantillon aléatoire EASSR (échantillonnage aléatoire simple

sans remise) de n_d unités dans la strate $d\,(d = 1,\ldots,D)$ contenant N_d unités. La taille relative de domaine d est $P_d = N_d / N$.

Nous posons un modèle linéaire ξ à deux niveaux conditionnel aux valeurs de $\mathbf{x}$ avec des effets aléatoires de strate sans corrélation u_d et des effets aléatoires ε_i :

$$\left.\begin{aligned} y_i &= \boldsymbol{\beta}'\mathbf{x}_i + u_d + \varepsilon_i \\ E_\xi(u_d) &= E_\xi(\varepsilon_i) = 0 \\ V_\xi(u_d) &= \sigma_{ud}^2 \\ V_\xi(\varepsilon_i) &= \sigma_{ed}^2 \end{aligned}\right\}, \tag{2.13}$$

où i renvoie à toutes les unités de la strate d. Ce modèle implique que $V_\xi(y_i) = \sigma_{ud}^2 + \sigma_{ed}^2$ pour toutes les unités de population et que $\operatorname{cov}_\xi(y_i, y_j)$ est égal à $\rho_d\,\sigma_d^2$ pour les unités $i \neq j$ dans la même strate et à zéro pour les unités d'autres strates, où $\rho_d = \sigma_{ud}^2 / (\sigma_{ud}^2 + \sigma_{ed}^2)$. Nous définissons une hypothèse simplifiée selon laquelle il y a égalité $\rho_d = \rho$ pour toutes les strates.

Après avoir posé un certain nombre d'autres hypothèses de simplification et calculé le poids optimal φ_d en (2.12), nous dégageons par anticipation l'EQM approchée optimale finale ou procédons avec modèle :

$$\mathrm{EQMA}_d = E_\xi \mathrm{EQM}_p\left(\tilde{y}_d^C\left[\varphi_{d(\mathrm{opt})}\right]; \bar{Y}_d\right) \approx \sigma_d^2 \rho(1-\rho)\left[1 + (n_d - 1)\rho\right]^{-1}. \tag{2.14}$$

Ensuite, nous définissons et élaborons le critère F en formule approchée finale à l'aide des EQM par anticipation des estimateurs de la moyenne de petit domaine et de la moyenne globale pour la répartition fondée sur un modèle :

$$\begin{aligned} F &= \sum_{d=1}^{D} N_d^q\,\mathrm{EQMA}_d + G N_+^{(q)} E_\xi \operatorname{var}_p\left(\hat{\bar{Y}}_r\right) \\ &\approx \sum_{d=1}^{D} N_d^q\,\sigma_d^2\,\rho(1-\rho)\left[1 + (n_d - 1)\rho\right]^{-1} + G N_+^{(q)} \sum_{d=1}^{D} \sigma_d^2 P_d^2 n_d^{-1}(1-\rho). \end{aligned} \tag{2.15}$$

Nous obtenons des tailles d'échantillon optimales pour les domaines en minimisant (2.15) sous réserve de $\sum_d n_d = n$. L'expression (2.15) suit l'idée de Longford (2006). Le poids N_d^q traduit la priorité inférentielle (importance) du domaine d avec $0 \le q \le 2$, et $N_+^{(q)} = \sum_{d=1}^{D} N_d^q$. La quantité G est un coefficient de priorité relative au niveau de la population. Si nous excluons comme but l'estimation de la moyenne de population, nous avons $G = 0$, et l'attention se porte alors seulement sur l'estimation au niveau du domaine. Par ailleurs, plus la valeur de G augmente, plus la deuxième composante en (2.15) domine et plus l'estimation au niveau du domaine est écartée.

Nous supposons d'abord que l'estimation de population n'est pas prioritaire $(G = 0)$ et que le coût d'enquête par unité est fixe. Dans ce cas, la minimisation de (2.15) par rapport à n_d a une solution unique

$$n_{d,\mathrm{opt}} = \frac{n\sqrt{\sigma_d^2\,N_d^q}}{\sum_{d=1}^{D}\sqrt{\sigma_d^2\,N_d^q}} + \frac{1-\rho}{\rho}\left(\frac{\sqrt{\sigma_d^2\,N_d^q}}{D^{-1}\sum_{d=1}^{D}\sqrt{\sigma_d^2\,N_d^q}} - 1\right). \tag{2.16}$$

La formule (2.16) contient deux paramètres inconnus, à savoir la corrélation intraclasse ρ et la variance spécifique de domaine σ_d^2. Nous remplaçons ρ par un coefficient ajusté d'homogénéité de la variable auxiliaire x. Ce coefficient approche la valeur CIC de corrélation intraclasse (section 2.2). Nous substituons au paramètre σ_d^2 la variance de x dans le domaine d. Si nous optons pour ce double remplacement, c'est

que y est en corrélation avec x. Si en plus l'estimation de population est prioritaire $(G > 0)$, (2.16) ne s'applique pas et F doit être minimisé numériquement par la méthode PNL, comme nous l'avons fait (Excel Solver, option PNL), par exemple.

Tableau 2.1
Récapitulation des répartitions fondées sur un modèle et assistées d'un modèle

Méthode	Calcul de la taille d'échantillon n_d pour le domaine d	Niveau d'optimalité
Méthode $g1$ fondée sur un modèle	$n_d^{g1} = \dfrac{N_d n - (N - N_d D - n)(1/R_{ax}^2 - 1)}{N + D(1/R_{ax}^2 - 1)}$, où R_{ax}^2 est la mesure ajustée d'homogénéité de la variable auxiliaire x.	Domaine
Méthode MCG0 assistée d'un modèle Méthode MCG50 assistée d'un modèle	$n_{d,\text{opt}} = \dfrac{n\sqrt{\sigma_d^2 N_d^q}}{\sum_{d=1}^D \sqrt{\sigma_d^2 N_d^q}} + \dfrac{1-\rho}{\rho}\left(\dfrac{\sqrt{\sigma_d^2 N_d^q}}{D^{-1}\sum_{d=1}^D \sqrt{\sigma_d^2 N_d^q}} - 1\right)$ Minimisation de $F = \sum_{d=1}^D N_d^q \sigma_d^2 \rho(1-\rho)[1+(n_d-1)\rho]^{-1} + GN_+^{(q)} \sum_{d=1}^D \sigma_d^2 P_d^2 n_d^{-1}(1-\rho)$ par rapport à n_d. Le paramètre ρ est remplacé par R_{ax}^2 et σ_d^2, par $S_d^2(x)$.	Domaine et population à la fois

3 Quelques modes de répartition par domaine sans modèle

Notre but dans la présente section est de passer en revue, à titre de référence, les cinq modes de répartition déjà présentés. Selon le genre d'information auxiliaire utilisée dans chaque cas, ces méthodes se répartissent en deux groupes (répartitions fondées sur les données de dénombrement et répartitions fondées sur les paramètres).

3.1 Répartitions fondées sur les données de dénombrement

La répartition égale et la répartition proportionnelle sont deux modes de répartition de base couramment utilisés. Aucune n'applique de critères particuliers au niveau du domaine ou de la population. Leur application exige seulement de l'information sur le nombre de strates D et le nombre d'unités N_d dans chaque strate.

Dans la répartition égale, la taille d'échantillon n_d est simplement le quotient

$$n_d^{\text{Equ}} = n/D. \tag{3.1}$$

Il est recommandé de choisir la taille d'échantillon globale n pour que le quotient soit un nombre entier. Dans le mode de répartition égale, on ne tient aucunement compte des différences entre les domaines, ce qui donne des estimations imprécises de domaine. Une borne inférieure naturelle pour la taille d'échantillon est min $n = 2D$.

Le mode de répartition proportionnelle est fréquemment employé. On calcule alors les tailles d'échantillon de domaines par

$$n_d^{\text{Pro}} = n\left(N_d / N\right). \tag{3.2}$$

Si les domaines sont de taille très variable, cela peut donner une situation où la taille d'échantillon attribuée est $n_d^{\text{Pro}} < 2$ pour un ou plusieurs domaines. Il s'agit d'un obstacle au calcul d'estimations directes fondées sur le plan de l'estimateur pour les erreurs-types. Une solution est d'appliquer le mode de répartition mixte proposé par Costa, Satorra et Ventura (2004). Ce mode est une solution pondérée joignant la répartition égale à la répartition proportionnelle selon la situation. La taille combinée d'échantillon de domaines est alors

$$n_d^{\text{Com}} = k n_d^{\text{Pro}} + (1-k)\, n_d^{\text{Equ}} \tag{3.3}$$

pour une constante spécifiée $k\,(0 \le k \le 1)$. Un problème secondaire se présente quand, pour un certain nombre de domaines, $n/D > N_d$. Une solution modifiée existe en pareil cas.

3.2 Répartitions fondées sur les paramètres

Dans ces modes de répartition, on utilise l'information au niveau du domaine de la variable étudiée y et, dans certains cas, de la variable auxiliaire x en corrélation avec y. Les valeurs de x sont disponibles pour toutes les unités de population. Dans la pratique, on remplace y inconnu par une variable appropriée de substitution y^* comme une variable étudiée venant d'une recherche antérieure sur le même sujet. Autre possibilité, on obtient les valeurs de y^* à l'aide d'un modèle approprié dans un petit échantillon préalable. On peut également substituer x à y. Les critères de répartition peuvent être fixés au niveau de la population seulement, au niveau du domaine seulement ou aux deux niveaux combinés.

Dans la répartition de Neyman, on recherche une précision optimale pour les paramètres de population $\text{ET}(y)_d$ (Tschuprow 1923). On doit alors connaître l'écart-type de la variable étudiée y ou d'une certaine variable de substitution et le nombre d'unités dans chaque domaine. Cette méthode privilégie les grands domaines à forte variation.

La répartition « spectrale » de Bankier (1988) est fondée sur un critère fixé au niveau du domaine. On pondère les valeurs CV de domaine de y au moyen de transformations de total de domaine X_d^q comportant une constante de cadrage q. Dans la pratique, on doit utiliser y^* ou x à la place de y. Cette méthode privilégie surtout les grands domaines à fort coefficient de variation.

Choudhry, Rao et Hidiroglou (2012) proposent la méthode de répartition PNL pour l'estimation directe. Ils ont recours à la programmation non linéaire pour la recherche d'une solution. Ils définissent les critères de répartition en fixant des bornes supérieures aux valeurs CV de la variable étudiée y dans chaque domaine et dans la population. Dans la pratique, y^* ou x remplace y. Ce programme recherche alors la taille d'échantillon minimale $n = \sum_d n_d$ satisfaisant à ces conditions. Nous avons employé la procédure PNL dans SAS (pour *Statistical Analysis System*) avec l'option Newton-Raphson pour trouver la solution. Cette méthode privilégie les domaines à fort coefficient de variation sans égard à la taille de domaine N_d.

Le tableau 3.1 récapitule les modes de répartition sans modèle et les formules de calcul des tailles d'échantillon de domaine.

Tableau 3.1
Récapitulation des répartitions fondées sur des données de dénombrement et sur des paramètres

Répartition	Calcul de la taille d'échantillon de domaine n_d	Niveau d'optimalité
Égale	$n_d^{\text{Equ}} = n/D$	Domaine
Proportionnelle	$n_d^{\text{Pro}} = n(N_d/N)$	Population
Neyman	$n_d^{\text{Ney}} = n\left(N_d S_d \Big/ \sum_{d=1}^{D} N_d S_d\right)$, où S_d est l'écart-type de y (y^* ou x dans la pratique) dans le domaine d.	Population
Bankier	$n_d^{\text{Ban}} = n\left(X_d^q \text{CV}(y)_d \Big/ \sum_{d=1}^{D} X_d^q \text{CV}_d(y)\right)$, où X_d est le total de domaine de x, où $\text{CV}_d(y) = S_d/\overline{Y}_d$ et où q est une constante d'ajustement. Dans la pratique, y^* ou x remplace y.	Domaine
PNL	$n_{st}^{\text{PNL}} = \min\left(\sum_{d=1}^{D} n_d\right)$, où les tolérances $\text{CV}(\overline{y}_d) \le \text{CV}_{0d}$ et $\text{CV}(\overline{y}_{st}) \le \text{CV}_0$ sont respectées. Dans la pratique, y^* ou x remplace y.	Population et domaine à la fois

Mentionnons brièvement d'autres modes de répartition fondée sur des paramètres. Ainsi, Longford (2006) a introduit les priorités inférentielles P_d pour les strates d et G pour la population et a fait intervenir ces contraintes dans la répartition. Une autre solution est proposée par Falorsi et Righi (2008). Celle-ci n'impose pas directement des quotas, mais essaie d'aménager l'ensemble des données à l'aide d'un plan d'échantillonnage à plusieurs degrés, de sorte que l'estimation sur domaine puisse s'effectuer efficacement.

4 Comparaison du rendement des modes de répartition

Dans la présente section, nous examinons le rendement des méthodes de répartition présentées aux sections 2 et 3. Les paramètres estimés sont les totaux de domaine et de population de la variable étudiée y. La taille d'échantillon globale est $n = 112$. À la section 4.1, nous décrivons les données de recherche et, à la section 4.3, les expériences de simulation et les comparaisons de méthodes de répartition.

4.1 Données empiriques

Nos données de recherche viennent d'un registre finnois national de logements d'immeubles d'appartements en vente. Ce registre est tenu par une société privée, Alma Mediapartners Ltd, dont les clients sont des agences immobilières. Toute l'information nécessaire sur les appartements est sauvegardée

dans ce registre dès qu'un mandat est reçu des propriétaires. La population que nous avons utilisée comprend 9 815 appartements (ce sont nos unités d'échantillonnage) en vente figurant au registre. Au total, 14 districts finlandais, surtout des villes, y sont représentés au printemps de 2011. Les tailles du domaine le plus petit et du plus grand étaient respectivement de 112 et 1 333. La variable étudiée (y) mesure le prix de l'appartement (1 000 €) et la variable auxiliaire (x), la taille (m^2). Le tableau 4.1 présente les tailles de domaine (N_d) et les statistiques sommaires de population (totaux, moyennes, écarts-types et coefficients de variation) pour y et x, ainsi que les valeurs de corrélation entre x et y. Les caractéristiques des domaines varient amplement. Le domaine le plus divergent est Helsinki.

Tableau 4.1
Statistiques sommaires de population

Domaine		Variable étudiée y				Variable auxiliaire x				Corrélation
Désignation	N_d	Y_d	$\bar{Y}_d$	$S_d(y)$	$\mathrm{CV}_d(y)$	X_d	$\bar{X}_d$	$S_d(x)$	$\mathrm{CV}_d(x)$	r_{yx}
Ville de Porvoo	112	25 409	226,86	207,82	0,916	8 940	79,82	50,67	0,635	0,877
District de Pirkkala	148	30 323	204,88	87,82	0,429	11 149	75,33	23,78	0,316	0,823
Comté de Savo Sud	493	64 863	131,57	72,90	0,554	32 644	66,22	20,25	0,306	0,437
Ville de Jyväskylä	494	89 941	182,07	69,65	0,383	40 000	80,97	17,62	0,218	0,509
Comté de Lappi	555	62 143	111,97	50,15	0,448	30 805	55,50	16,22	0,292	0,207
Sud-Est de la Finlande	585	98 504	168,38	106,78	0,634	47 750	81,62	21,68	0,266	0,601
Helsinki (capitale)	621	437 902	705,16	562,38	0,798	76 931	123,88	57,98	0,468	0,753
District de la côte ouest	655	108 339	165,40	75,85	0,459	50 903	77,71	36,39	0,468	0,439
District « Trackside »	818	148 845	181,96	65,08	0,358	59 220	72,40	23,84	0,321	0,517
District de Kuopio	871	126 867	145,66	75,79	0,520	64 103	73,60	23,27	0,324	0,580
District de Turku	958	166 613	173,92	131,62	0,757	79 970	83,48	25,71	0,308	0,635
District d'Oulu	1 072	133 591	124,62	50,19	0,403	59 210	55,23	16,92	0,306	0,392
Région métropolitaine	1 100	263 293	239,36	117,84	0,492	80 034	72,76	26,37	0,362	0,754
District de Lahti-Tampere	1 333	262 400	196,85	110,76	0,563	105 804	79,37	25,54	0,322	0,602
Population	9 815	2 019 031	205,71	215,52	1,048	747 462	76,16	31,76	0,417	0,674

La mesure ajustée d'homogénéité de la variable auxiliaire x est $R_{ax}^2 = 0{,}231$, ce qui indique une variabilité plutôt marquée entre les domaines.

4.2 Modes de répartition

En général, la taille d'échantillon globale dépend des délais et des ressources financières dont on dispose dans un projet de recherche. Cet aspect n'est pas pris en compte dans la présente étude, la question étant celle d'une étude expérimentale. Le taux d'échantillonnage a été déterminé par $f(\%) = 100 \times (112/9\,815) = 1{,}14\ \%$. Nous avons dégagé les valeurs de répartition par méthode selon les formules présentées aux tableaux 2.1 et 3.1. Nous avons tenu compte de certains détails. Dans la répartition de Bankier, la valeur d'une constante de cadrage q est de 0,5. Dans la répartition PNL, les limites choisies de CV sont de 0,1258 (12,58 %) pour les domaines et de 0,0375 (3,75 %) pour la population, ce qui donne une taille d'échantillon globale de 112. Nous appliquons la procédure d'Excel Solver avec l'option de

programmation non linéaire pour résoudre le problème de répartition PNL. Nous recourons à une répartition proportionnelle modifiée pour obtenir une taille d'échantillon de domaine d'au moins deux. Nous attribuons d'abord une unité à chaque domaine, puis le reste (98 unités) par répartition proportionnelle. Nous remplaçons y par x dans chaque répartition fondée sur des paramètres. Dans les répartitions assistées d'un modèle, nous fixons la valeur de q à l'unité et la quantité G, à zéro ou à 50. Les tailles d'échantillon finales figurent au tableau 4.2 pour les diverses répartitions. La variation des tailles d'échantillon au niveau du domaine est très forte entre les modes de répartition.

Tableau 4.2
Tailles d'échantillon de domaine selon la répartition

Domaine		Fondée sur un modèle	Estimation composite assistée d'un modèle		Répartitions fondées sur des données de dénombrement		Répartitions fondées sur des paramètres		
Désignation	N_d	$g1^*$	**MCG0***	**MCG50***	**ÉGALE**	**PROP.**	**Ney_X**	**Ban_X**	**PNL_X**
Ville de Porvoo	112	0	6	3	8	2	2	6	20
District de Pirkkala	148	0	2	2	8	2	2	4	6
Comté de Savo Sud	493	5	4	4	8	6	4	6	6
Ville de Jyväskylä	494	5	3	4	8	6	4	5	3
Comté de Lappi	555	6	3	4	8	6	4	5	5
Sud-Est de la Finlande	585	6	6	5	8	7	6	6	4
Helsinki (capitale)	621	7	21	16	8	7	16	14	14
District de la côte ouest	655	7	12	11	8	8	10	11	14
District « Trackside »	818	10	8	8	8	9	9	8	7
District de Kuopio	871	11	8	9	8	10	9	8	6
District de Turku	958	12	10	11	8	11	11	9	6
District d'Oulu	1 072	13	6	8	8	12	8	8	6
Région métropolitaine	1 100	13	11	12	8	12	13	11	8
District Lahti-Tampere	1 333	17	12	15	8	14	14	11	7
Total	9 815	112	112	112	112	112	112	112	112

* En fonction du coefficient ajusté d'homogénéité (valeur 0,231) qui est calculé pour x.

4.3 Comparaison du rendement des modes de répartition

Dans la présente section, nous présentons les résultats selon les expériences de simulation de conception. Pour chaque méthode, nous avons mis en simulation 1 500 échantillons stratifiés indépendants EASSR dans le programme SAS et procédé aux calculs nécessaires à partir des échantillons simulés dans le programme SPSS (pour *Statistical Package for the Social Sciences*). Nous avons soumis à une estimation EBLUP fondée sur un modèle les échantillons de chaque répartition. À des fins de comparaison des modes de répartition, nous avons calculé deux mesures de qualité en pourcentage, à savoir la REQMR et le BRA % dans chaque cas.

Nous supposons que r échantillons simulés sont tirés dans chaque répartition. Soit $\hat{Y}_{di,\mathrm{EBLUP}}$ l'estimation EBLUP du total de domaine Y_d dans le i^{e} échantillon $(i = 1,\ldots,r)$. La REQMR et le BRA en pourcentage se définissent ainsi :

$$\text{REQMR}_d\% = 100 \times \sqrt{1/r \sum_{i=1}^{r} \left(\hat{Y}_{di,\text{EBLUP}} - Y_d\right)^2} \Big/ Y_d ,$$

$$\text{BRA}_d\% = 100 \times \left| 1/D \sum_{i=1}^{r} \left(\hat{Y}_{di,\text{EBLUP}} \Big/ Y_d - 1\right) \right|,$$

Les moyennes sur les domaines se calculent de la manière suivante :

$$\text{REQMRM}\% = 1/D \sum_{d=1}^{D} \text{REQMR}_d\% \quad \text{et} \quad \text{BRAM}\% = 1/D \sum_{d=1}^{D} \text{BRA}_d\%.$$

L'estimation du total de population dans le i^{e} échantillon simulé $(i = 1, \ldots, r)$ est la somme des estimations des totaux de domaine : $\hat{Y}_{i,\,\text{EBLUP}} = \sum_{d=1}^{D} \hat{Y}_{di,\,\text{EBLUP}}$. La REQMR en pourcentage du total de population se calcule ainsi :

$$\text{REQMR}_{\text{pop}}\% = 100 \times \sqrt{1/r \sum_{i=1}^{r} \left(\hat{Y}_{i,\text{EBLUP}} - Y\right)^2} \Big/ Y ,$$

où Y est la valeur vraie du total de population pour laquelle le BRA en pourcentage se calcule comme

$$\text{BRA}_{\text{pop}}\% = 100 \times \left| 1/r \sum_{i=1}^{r} \left(\hat{Y}_{i,\text{EBLUP}} \Big/ Y - 1\right) \right|.$$

Les tableaux 4.3 et 4.4 présentent les valeurs REQMR et BRA en pourcentage des domaines, leurs moyennes sur les domaines et les REQMR et les BRA de population pour chaque mode de répartition. L'évaluation des résultats est à deux arguments, à savoir la valeur moyenne de la mesure de qualité pour le niveau du domaine et la valeur de la mesure de qualité pour le niveau de la population.

Tableau 4.3
REQMR en pourcentage de domaine et de population par mode de répartition

Domaine	N_d	$g1$	MCG0	MCG50	ÉGALE	PROP.	Ney_X	Ban_X	PNL_X
Ville de Porvoo	112	8,08	14,63	15,93	13,41	19,79	16,49	14,78	10,10
District de Pirkkala	148	6,60	9,72	10,77	8,35	12,04	10,60	9,76	8,97
Comté de Savo Sud	493	22,29	22,77	23,20	18,63	20,70	23,20	20,16	20,88
Ville de Jyväskylä	494	15,36	24,55	20,70	13,61	14,43	20,83	18,33	21,98
Comté de Lappi	555	21,72	28,19	26,19	19,91	21,34	25,45	23,97	22,59
Sud-Est de la Finlande	585	20,76	27,25	25,93	19,68	19,64	24,37	24,31	27,81
Helsinki (capitale)	621	22,72	12,68	14,97	21,92	23,15	14,35	16,02	16,43
District de la côte ouest	655	21,15	22,43	21,57	20,35	19,92	21,75	20,67	18,91
District « Trackside »	818	11,93	12,86	13,63	12,31	11,38	13,73	12,76	13,47
District de Kuopio	871	16,22	23,22	20,70	19,21	16,37	20,84	20,82	23,49
District de Turku	958	17,56	24,75	21,66	20,94	17,74	21,57	22,70	26,44
District d'Oulu	1 072	14,39	25,40	21,14	16,96	14,34	21,22	19,00	19,81
Région métropolitaine	1 100	9,59	11,31	10,86	12,14	9,78	10,16	10,78	11,55
District de Lahti-Tampere	1 333	10,54	13,43	11,66	13,35	10,64	12,76	12,87	14,98
Moyenne sur les domaines (%)		15,65	19,51	18,59	16,48	16,52	18,38	17,64	18,39
Valeur de population (%)		6,15	6,53	5,88	6,13	5,97	6,07	5,89	6,62

Nous avons obtenu en pourcentage le REQMR moyen le plus bas (15,65 %) pour le mode de répartition $g1$ conçu pour la présente étude. Helsinki faisait exception au niveau du domaine parce que sa REQMR en pourcentage était nettement plus élevée dans cette répartition que dans les répartitions assistées d'un modèle et fondées sur des paramètres. Les répartitions égale et proportionnelle donnaient aussi de bons résultats au

niveau du domaine avec des valeurs moyennes 16,48 % et 16,52 %. Nous avons obtenu la moyenne la plus élevée pour les répartitions MC assistées d'un modèle. Au niveau de la population, nous avons dégagé la valeur la plus basse de la mesure de qualité pour la répartition MCG50 assistée d'un modèle (5,88 %) et la deuxième la plus basse pour la répartition de Bankier (5,89 %), mais les différences étaient généralement légères entre les répartitions à ce niveau.

Tableau 4.4
BRA en pourcentage de domaine et de population par mode de répartition

Domaine	N_d	$g1$	MCG0	MCG50	ÉGAL	PROP.	Ney _ X	Ban _ X	PNL _ X
Ville de Porvoo	112	2,28	2,20	0,97	0,04	1,26	1,28	0,98	0,79
District de Pirkkala	148	0,17	2,10	1,08	0,19	0,79	0,85	0,86	1,15
Comté de Savo Sud	493	8,08	11,81	10,87	6,76	7,29	11,47	9,09	9,81
Ville de Jyväskylä	494	6,09	19,78	15,36	6,10	5,82	14,33	12,16	16,31
Comté de Lappi	555	2,08	5,27	3,14	1,45	2,70	2,44	1,22	1,44
Sud-Est de la Finlande	585	9,05	20,62	18,28	9,53	8,11	15,69	15,96	20,41
Helsinki (capitale)	621	9,71	6,38	7,93	10,95	11,59	7,43	8,80	9,45
District de la côte ouest	655	7,83	12,34	11,60	9,07	8,16	12,69	10,52	10,87
District « Trackside »	818	1,21	3,11	1,78	1,76	0,96	2,61	2,10	2,94
District de Kuopio	871	6,00	14,90	10,68	9,37	6,53	11,33	11,77	15,56
District de Turku	958	5,26	16,46	12,59	8,48	5,78	11,54	13,27	16,91
District d'Oulu	1 072	0,81	10,17	6,08	1,88	1,84	6,47	4,71	4,00
Région métropolitaine	1 100	3,06	5,84	5,11	5,29	3,37	4,39	5,12	5,76
District de Lahti-Tampere	1 333	1,86	6,14	3,97	3,62	1,79	4,65	4,37	6,10
Moyenne sur les domaines (%)		4,53	9,79	7,82	5,32	4,71	7,66	7,21	9,15
Valeur de population (%)		0,01	3,33	2,05	0,18	0,50	2,26	1,83	3,01

La répartition $g1$ était la seule pour laquelle le biais relatif absolu était de moins de 10 % dans chaque domaine. Ce biais était pratiquement nul au niveau de la population. De plus, les répartitions égale et proportionnelle présentaient de faibles biais au double niveau du domaine et de la population, mais les répartitions assistées d'un modèle et fondées sur des paramètres étaient d'un rendement nettement inférieur. Un détail intéressant dans le cas de la répartition $g1$ est que la précision des estimations de domaine est plutôt bonne et que le biais relatif est bas également dans le cas de deux domaines avec taille d'échantillon nulle. Un trait commun à ces domaines est que les moyennes des variables y et x sont proches des moyennes de population correspondantes. En tout cas, il est essentiel que l'estimation fondée sur un modèle puisse produire des estimations fiables pour les domaines sans représentation dans l'échantillon aléatoire.

5 Observations en conclusion

Notre recherche a porté sur sept modes de répartition classés en trois groupes selon les données auxiliaires nécessaires à la mise en œuvre des solutions. La quantité d'information auxiliaire est la moindre dans les répartitions égale et proportionnelle qui reposent sur le nombre de domaines et le nombre d'unités

statistiques dans chacun. Les répartitions de Neyman et Bankier et la répartition PNL sont fondées sur des critères d'optimisation préétablis et l'application de ces méthodes fait intervenir des valeurs de paramètres par domaine comme l'écart-type ou le coefficient de variation de la variable étudiée. Dans la répartition de Bankier, les totaux de domaine d'au moins une variable auxiliaire doivent être connus. Comme la variable étudiée est inconnue, elle doit être remplacée par une variable substitutive ou auxiliaire appropriée pour que ces trois méthodes puissent s'appliquer. Un trait commun aux répartitions fondées sur des données de dénombrement et sur des paramètres est qu'elles ne font appel à aucun modèle, alors que les trois autres utilisent non seulement un modèle, mais aussi des données de dénombrement.

À en juger par nos résultats empiriques, nous pouvons penser que la répartition $g1$ fondée sur un modèle est la meilleure de toutes les répartitions analysées dans notre recherche. Il faut aussi dire que les répartitions égale et proportionnelle ont donné de bons résultats, alors que les répartitions assistées d'un modèle et fondées sur des paramètres étaient d'un rendement nettement moindre. Les trois derniers modes de répartition ont été conçus à l'origine pour une estimation directe selon le plan, et leurs résultats peuvent se comprendre de ce point de vue. Si on les compare à la répartition $g1$, les répartitions MC font appel à un modèle différent, ce qui peut influer sur leurs résultats.

Une caractéristique de la répartition $g1$ est que le modèle et la méthode d'estimation sont intégrés au départ à l'élaboration du plan d'échantillonnage et constituent donc une information préalable donnée. Pour ce mode de répartition qui repose sur un modèle linéaire mixte au niveau de l'unité et sur la méthode d'estimation EBLUP, nous avons seulement besoin du coefficient d'homogénéité entre domaines qui se calcule par les valeurs de la variable auxiliaire. À cet égard, cette répartition diffère des autres dans notre comparaison. Ajoutons que le point de départ est différent au moment de choisir la méthode d'estimation finale, puisque cette répartition privilégie une estimation fondée sur un modèle, et non une estimation directe selon le plan où interviennent les poids d'échantillonnage. Le choix d'une estimation fondée sur un modèle se justifie également par son caractère courant dans l'estimation sur petits domaines. Par ailleurs, la répartition $g1$ permet d'utiliser de petites tailles d'échantillon, l'information pouvant être empruntée entre domaines dans l'application du modèle. Cela peut avoir de l'importance dans les enquêtes ou les études rapides que peuvent faire les organismes d'études de marché là où même une mesure unique coûte cher. Il importe toutefois d'examiner les caractéristiques des domaines, et des petits en particulier, avant d'établir les tailles d'échantillon finales.

Nous recommanderions d'entreprendre une recherche plus large pour constater quels sont les avantages et les inconvénients de la détermination, dès la conception du plan de recherche, de la technique de calcul à employer pour produire les statistiques de domaines.

Remerciements

Les auteurs remercient le rédacteur en chef, le rédacteur adjoint et les deux examinateurs, ainsi que le professeur Risto Lehtonen, de leurs observations et leurs suggestions constructives.

Bibliographie

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician,* 42, 174-177.

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

Costa, A., Satorra, A. et Ventura, E. (2004). Improving both domain and total area estimation by composition. *SORT,* 28(1), 69-86.

Falorsi, P.D., et Righi, P. (2008). Une approche d'échantillonnage équilibré pour des plans de sondage à stratification multidimensionnelle pour l'estimation pour petits domaines. *Techniques d'enquête,* 34, 2, 247-259. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2008002/article/10763-fra.pdf.

Keto, M., et Pahkinen, E. (2009). On sample allocation for effective EBLUP estimation of small area totals – "Experimental Allocation". Dans *Survey Sampling Methods in Economic and Social Research,* (Éds., J. Wywial et W. Gamrot), 2010. Katowice: Katowice University of Economics.

Keto, M., et Pahkinen, E. (2014). On sample allocation for efficient small area estimation. *Book of Abstracts.* SAE 2014, Poland: Poznan University of Economics, page 50.

Longford, N.T. (2006). Calcul de la taille de l'échantillon pour l'estimation pour petits domaines. *Techniques d'enquête,* 32, 1, 97-106. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2006001/article/9259-fra.pdf.

Molefe, W.B., et Clark, R.G. (2015). Répartition optimale assistée par modèle pour des domaines planifiés en utilisant l'estimation composite. *Techniques d'enquête,* 41, 2, 399-410. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2015002/article/14230-fra.pdf.

Nissinen, K. (2009). *Small Area Estimation with Linear Mixed Models from Unit-Level Panel and Rotating Panel Data*. Thèse de doctorat, Université de Jyväskylä, Département de mathématiques et de statistique, Rapport 117, https://jyx.jyu.fi/dspace/handle/123456789/21312.

Pfefferman, D. (2013). New important developments in small area estimation. *Statistical Science,* 28, 40-68.

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

Tschuprow, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron,* Vol. 2, 3, 461-493; 4, 646-683.

Techniques d'enquête, juin 2017
Vol. 43, nº 1, pp. 119-138
Statistique Canada, nº 12-001-X au catalogue

Une approche markovienne mixte à classes latentes pour estimer la mobilité sur le marché du travail au moyen d'indicateurs multiples et d'une interrogation rétrospective

Francesca Bassi, Marcel Croon et Davide Vidotto[1]

Résumé

Les erreurs de mesure peuvent provoquer un biais de l'estimation des transitions, donnant lieu à des conclusions erronées au sujet de la dynamique du marché du travail. La littérature traditionnelle sur l'estimation des mouvements bruts est basée sur la supposition que les erreurs de mesure ne sont pas corrélées au fil du temps. Cette supposition n'est pas réaliste dans bien des contextes, en raison du plan d'enquête et des stratégies de collecte de données. Dans le présent document, nous utilisons une approche basée sur un modèle pour corriger les mouvements bruts observés des erreurs de classification au moyen de modèles markoviens à classes latentes. Nous nous reportons aux données recueillies dans le cadre de l'enquête italienne continue sur la population active, qui est transversale et trimestrielle et qui comporte un plan de renouvellement de type 2-2-2. Le questionnaire nous permet d'utiliser plusieurs indicateurs des états de la population active pour chaque trimestre : deux recueillis au cours de la première interview, et un troisième recueilli un an plus tard. Notre approche fournit une méthode pour estimer la mobilité sur le marché du travail, en tenant compte des erreurs corrélées et du plan par renouvellement de l'enquête. Le modèle qui convient le mieux est un modèle markovien mixte à classes latentes, avec des covariables touchant les transitions latentes et des erreurs corrélées parmi les indicateurs; les composantes mixtes sont de type mobile-stable. Le caractère plus approprié de la spécification du modèle mixte est attribuable à des transitions latentes estimées avec une plus grande précision.

Mots-clés : Mouvements bruts; marché du travail; modèles mixtes; modèles à classes latentes.

1 Introduction

Les analystes peuvent exploiter les données par panels pour estimer les mouvements bruts de la population active, c'est-à-dire les transitions d'un état à un autre au fil du temps. Les mouvements nets mesurent les variations de divers états sur le marché au fil du temps, tandis que les mouvements bruts nous renseignent sur la dynamique du marché du travail.

Beaucoup de documents sur l'estimation des mouvements bruts sont basés sur l'hypothèse que les erreurs sont non corrélées au fil du temps, c'est-à-dire qu'il s'agit d'erreurs de classification indépendantes (ECI). L'hypothèse d'ECI suppose que : (i) les erreurs de classification désignant deux occasions différentes sont indépendantes l'une de l'autre en fonction des états réels, et (ii) les erreurs dépendent exclusivement de l'état réel actuel. Par conséquent, les erreurs de classification produisent de fausses transitions et provoquent donc une surestimation des changements.

Cependant, dans bien des contextes, l'hypothèse de l'ECI s'avère irréaliste, en raison du plan d'enquête et des stratégies de collecte de données. En pareilles circonstances, les erreurs de classification peuvent être corrélées : les états observés peuvent également dépendre des états réels à d'autres moments ou des transitions réelles, ou il peut y avoir des effets directs entre les états observés (Bound, Brown et Mathiowetz 2001).

1. Francesca Bassi, Department of Statistical Sciences, Université de Padoue, Italie, Via C. Battisti 241, 35121, Padoue, Italie. Courriel : francesca.bassi@unipd.it; Marcel Croon et Davide Vidotto, département de la méthodologie et des statistiques, Université de Tilburg, aux Pays-Bas.

Dans le présent document, nous utilisons une approche basée sur un modèle pour rajuster les mouvements bruts observés pour les erreurs de classification. Cette approche agence un sous-modèle structurel pour les taux de transition réels non observés et un sous-modèle de mesure corrélant les états réels aux états observés. Un cadre utile pour formuler notre modèle est fourni par l'analyse des classes latentes (CL).

Nous appliquons notre approche aux mouvements bruts observés au sein des trois états de la population active – Occupant un emploi (E), En chômage (C) et Inactif (I) – tirés de l'enquête italienne continue sur la population active (EICPA), une enquête trimestrielle reposant sur un plan de renouvellement de type 2-2-2 entraînant des panels à deux cycles à un trimestre, trois trimestres et un an d'intervalle. Nous considérons que les données sont recueillies de 2005 à 2009.

Le questionnaire nous permet d'utiliser plusieurs indicateurs des états de la population active pour chaque trimestre : (i) tous les répondants sont classifiés comme Occupant un emploi, En chômage ou Inactifs, d'après la définition du Bureau international du travail (BIT) en fonction des réponses données à une série de questions; (ii) on demande aux répondants de s'autodéclarer comme occupant un emploi, en chômage ou inactifs, ce qui constitue l'auto-évaluation de l'état; (iii) une question rétrospective porte sur l'état des répondants sur le marché du travail un an avant l'interview. Cette approche offre une façon d'estimer la mobilité sur le marché du travail en tenant compte des erreurs de mesure corrélées et du plan par renouvellement de l'enquête.

Plus précisément, le modèle qui convient le mieux est un modèle markovien à classes latentes (MMCL), dont les covariables touchant les transitions latentes et les erreurs corrélées font partie des indicateurs. On obtient le modèle mixte en présumant de l'existence de deux sous-populations non observables : les répondants mobiles, c'est-à-dire ceux qui changent d'état sur le marché du travail pendant la période à l'étude, et les répondants stables. Notre recherche a eu comme résultat secondaire que le modèle des répondants mobiles/stables et le MMCL estiment la même quantité d'erreur de mesure dans les données. Le meilleur ajustement de la spécification du modèle mixte est attribuable à des transitions latentes estimées avec une plus grande précision. Magidson, Vermunt et Tran (2007) ont également constaté que le modèle markovien mixte par CL convient mieux aux données que le modèle traditionnel. Toutefois, dans ce cas-ci, la différence pour ce qui est de l'ajustement était attribuable au fait que, comme l'hétérogénéité n'a pas été prise en compte, il s'en est suivi une surestimation de l'erreur de mesure.

Notre document examine les contributions récentes aux ouvrages scientifiques sur la question de l'estimation des mouvements bruts au moyen de la chaîne markovienne cachée et d'indicateurs multiples. Une description exacte du modèle se trouve dans Langeheine (1994). La méthode ne s'appliquait pas seulement à l'estimation des mouvements bruts sur le marché du travail, mais aussi à bien d'autres contextes, des données longitudinales étant disponibles. Paas, Vermunt et Bijmolt (2007), par exemple, ont estimé un MMCL pour étudier les tendances d'acquisition sur le marché des produits financiers; plusieurs indicateurs de détention de produits financiers ont été utilisés pour déterminer des segments du marché non observables directement, au sein desquels les clients pouvaient se déplacer à différentes reprises consécutives de mesure.

Bartolucci, Lupparelli et Montanari (2009) ont estimé le même modèle pour les changements suivants de l'état de santé dans un échantillon de patients au fil du temps. Manzoni, Vermunt, Luijkx et Muffels (2010) ont appliqué un MMCL pour estimer les mouvements bruts sur le marché du travail suédois. Dans une étude plus récente, Pavlopoulos et Vermunt (2015) ont utilisé un modèle markovien pour estimer l'ampleur de l'erreur de mesure dans les renseignements de l'enquête sur la population active hollandaise et l'institut hollandais sur l'assurance pour les employés sur le type d'emploi (permanent ou temporaire).

Grâce à la contribution du présent document à la littérature scientifique au sujet de l'estimation des mouvements bruts, nous avons maintenant trois indicateurs, l'un d'eux étant recueilli rétrospectivement, sur l'état sur le marché du travail, et nous pouvons également tenir compte du plan par renouvellement de l'enquête. Le document contribue également à la littérature sur la qualité des données de l'EICPA (Bassi, Padoan et Trivellato 2012).

La présentation de l'article est la suivante. La section 2 présente le modèle traditionnel (ou standard) et le MMCL mixte. La section 3 décrit l'enquête et ses données. La section 4 compare les rendements du modèle traditionnel et du MMCL mixte. La section 5 présente les résultats, en s'appuyant sur le modèle convenant le mieux pour corriger les mouvements bruts sur le marché du travail à partir des erreurs de mesure. Les conclusions sont présentées à la section 6.

2 Le modèle markovien à classes latentes

L'analyse des classes latentes a été appliquée dans plusieurs études sur les données par panel pour séparer les changements réels des changements observés touchés par des mesures non fiables. Parmi les contributions relativement récentes, mentionnons celles de Bassi, Torelli et Trivellato (1998), de Biemer et Bushery (2000), de Bassi, Croon, Hagenaars et Vermunt (2000) et de Bassi et Trivellato (2009).

L'état réel sur le marché du travail est traité comme une variable latente, et l'état observé est traité comme son indicateur. Ce modèle comprend deux parties :

a) structure, décrivant la dynamique réelle parmi les variables latentes;

b) mesure, appariant chaque variable latente à son indicateur ou ses indicateurs.

Examinons la formulation la plus simple des modèles markoviens à classes latentes (MMCL) (Wiggins 1973), qui suppose que les transitions non observables réelles suivent une chaîne markovienne de premier ordre. Comme pour toutes les spécifications du MMCL standard, l'indépendance locale parmi les indicateurs est présumée, c'est-à-dire que les indicateurs sont indépendants en fonction des variables latentes. Dans le MMCL comportant un indicateur par variable latente, la supposition de l'indépendance locale coïncide avec l'état des erreurs de classification indépendantes.

Si l'on suppose que X_{it} indique l'état réel sur le marché du travail au moment t pour une personne de l'échantillon générique $i, i = 1,\ldots,n$; Y_{it} est l'état observé correspondant; $P\left(X_{i1} = l_1\right)$ est la probabilité de

l'état initial de la chaîne markovienne latente et $P\left(X_{it+1} = l_{t+1} \mid X_{it} = l_t\right)$ est la probabilité de transition entre l'état l_t et l'état l_{t+1} du moment t à $t+1$, sachant que $t = 1,\ldots,T-1$, où T représente le nombre total de périodes consécutives, séparées par des intervalles identiques, pendant lesquelles une personne est observée. En outre, $P\left(Y_{it} = j_t \mid X_{it} = l_t\right)$ est la probabilité d'observer l'état j au moment t, sachant que la personne i au moment t est dans l'état réel l_t : on parle également ici de la composante de mesure du modèle.

Il en résulte que $P(Y(1),\ldots,Y(T))$ est la proportion d'unités observées dans une cellule générique du tableau de contingence à $T-$ entrées. Pour une personne de l'échantillon générique i, un MMCL se définit comme suit :

$$P\left(\mathbf{Y}_i = \mathbf{y}\right) = \sum_{l_1}^{K} \ldots \sum_{l_T}^{K} P\left(X_{i1} = l_1\right) \prod_{t=2}^{T} P\left(X_{it} = l_t \mid X_{it-1} = l_{t-1}\right) \prod_{t=1}^{T} P\left(Y_{it} = j_t \mid X_{it} = l_t\right) \tag{2.1}$$

où $\mathbf{y}$ est le vecteur renfermant les valeurs observées pour la personne i, l_t et j_t varient sur K classes (dans notre application, trois états de la population active). L'équation (2.1) précise la proportion d'unités dans la cellule générique d'un tableau de contingence à $T-$ entrées comme produit de probabilités marginales et conditionnelles.

Dans un MMCL comportant des variables concomitantes, l'appartenance à des classes latentes et les transitions latentes sont exprimées comme des fonctions de covariables avec des distributions connues (Dayton et McReady 1988). $P\left(X_{i1} = l_1 \mid \mathbf{Z}_{i1} = \mathbf{z}_1\right)$, où $\mathbf{z}_1$ est un vecteur renfermant les valeurs des covariables pour le répondant i au moment 1, estime les effets des covariables sur l'état initial, et $P\left(X_{it} = l_t \mid X_{it-1}, \mathbf{Z}_{it} = \mathbf{z}_t\right)$, où $\mathbf{z}_t$ est un vecteur renfermant les valeurs des covariables pour le répondant i au moment t, estime les effets des covariables sur les transitions latentes.

En fonction des composantes qui précèdent, le modèle complet pour la personne i est donné par :

$$P\left(\mathbf{Y}_i = \mathbf{y} \mid \mathbf{Z}_i = \mathbf{z}\right) = \sum_{l_1}^{K} \ldots \sum_{l_T}^{K} P\left(X_{i1} = l_1 \mid \mathbf{Z}_1 = \mathbf{z}_1\right) \prod_{t=2}^{T} P\left(X_{it} = l_t \mid X_{it-1} = l_{t-1}, \mathbf{Z}_{it} = \mathbf{z}_t\right) \prod_{t=1}^{T} P\left(Y_{it} = j_t \mid X_{it} = l_t\right) \tag{2.2}$$

Lorsque plusieurs indicateurs (M) par variable latente sont observés, la formulation du modèle devient la suivante (Vermunt 2010) :

$$P\left(\mathbf{Y}_i = \mathbf{y} \middle| \mathbf{Z}_i = \mathbf{z}\right) = \sum_{l_1}^{K} \ldots \sum_{l_T}^{K} P\left(X_{i1} = l_1 \middle| \mathbf{Z}_1 = \mathbf{z}_1\right)$$

$$\prod_{t=2}^{T} P\left(X_{it} = l_t \middle| X_{it-1} = l_{t-1}, \mathbf{Z}_{it} = \mathbf{z}_t\right)$$

$$\prod_{m=1}^{M} \prod_{t=1}^{T} P\left(Y_{mit} = j_t \middle| X_{it} = l_t\right) \tag{2.3}$$

Dans notre application, les indicateurs M sont donnés par les trois éléments d'information recueillis pour tous les répondants sur leur état sur le marché du travail.

Habituellement, les probabilités conditionnelles sont paramétrées et limitées par des modèles de régression logistique. Les paramètres sont estimés au moyen du maximum de vraisemblance (Vermunt et Magidson 2013). L'identification est un problème bien connu dans les modèles comportant des variables latentes et, bien que le nombre de paramètres indépendants ne doive pas dépasser le nombre de fréquences observées, ce n'est pas une condition suffisante. D'après Goodman (1974), une condition suffisante pour l'identifiabilité locale est que la matrice d'information soit définie positive. Le logiciel Latent Gold (Vermunt et Magidson 2008) fournit de l'information sur l'identification des paramètres. Un autre problème lié à l'estimation est celui des maxima locaux; pour y remédier, nous avons estimé nos modèles plusieurs fois avec différents ensembles de valeurs de départ.

Un MMCL mixte présume de l'existence dans la population de groupes non directement observables qui se déplacent au fil du temps, en suivant des chaînes latentes ayant différentes probabilités d'état initial et différentes probabilités de transition; on peut également présumer que les groupes ont des probabilités de réponse différentes (van de Pol et Langeheine 1990). Un tel modèle peut être élargi de manière à inclure des covariables variables dans le temps et des covariables constantes dans le temps (Vermunt, Tran et Magidson 2008). Un cas particulier d'un MMCL mixte à deux classes est le modèle mobile-stable : le groupe des mobiles a des probabilités positives de passer d'un état à un autre au fil du temps, et le groupe des stables ne change pas. Pour ce dernier groupe, les probabilités de transition entre les différents états sont imposées à zéro. Un MMCL mixte à deux classes comportant des variables concomitantes prend la forme suivante :

$$P\left(\mathbf{Y}_i = \mathbf{y} \mid \mathbf{Z}_i = \mathbf{z}\right) = \sum_{w=1}^{2} \sum_{l_1}^{K} \ldots \sum_{l_T}^{K} P\left(W = w\right) P\left(X_{i1} = l_1 \middle| \mathbf{Z}_1 = \mathbf{z}_1, W = w\right)$$

$$\prod_{t=2}^{T} P\left(X_{it} = l_t \middle| X_{it-1} = l_{t-1}, \mathbf{Z}_{it} = \mathbf{z}_t, W = w\right)$$

$$\prod_{j_t=1}^{K} \prod_{t=1}^{T} P\left(Y_{it} = j_t \middle| X_{it} = l_t, W = w\right) \tag{2.4}$$

où W est une variable latente binaire. Le modèle mobile-stable est obtenu en présumant, pour $l_t \neq l_{t-1}$, que $P\left(X_{it} = l_t \middle| X_{it-1} = l_{t-1}, W = 2\right) = 0$ et donc pour $l_t = l_{t-1}$ que $P\left(X_{it} = l_t \middle| X_{it-1} = l_{t-1}, W = 2\right) = 1$.

La fonction de probabilité d'un modèle de CL peut également être estimée si l'information ne se trouve pas dans les variables de réponse. Nous profitons de cette occasion pour tenir compte des tendances de réponse générées par le plan d'enquête par renouvellement. Les ménages échantillonnés sont interviewés sur deux trimestres consécutifs, ne participent pas à l'enquête pendant les deux trimestres suivants et sont ensuite interviewés de nouveau, à deux autres reprises (voir le tableau 3.1). Nous avons présumé que l'information manquante en raison du plan d'enquête est manquante de façon aléatoire. Dans ce cas-ci, chaque unité contribue seulement à la fonction de vraisemblance d'après l'information disponible (Vermunt 1997).

3 Les données

L'enquête italienne continue sur la population active (EICPA), réalisée par l'ISTAT (*Italian Institute of Statistics*), est la principale source officielle de documentation statistique sur le marché du travail italien. L'EICPA est réalisée depuis 1969 et a été modifiée bien des fois. En 2004, une importante mise à jour a été effectuée, principalement en raison de l'obligation d'adapter l'enquête aux nouvelles normes de l'Union européenne (UE). Les principaux changements mettaient en cause des interviews réparties au fil des années de l'étude, de nouveaux critères pour classifier l'état des répondants sur le marché du travail, des techniques de collecte de données assistée par ordinateur et des interviews avec rétroinformation. Chaque année, l'enquête recueille des renseignements sur environ 280 000 ménages, ce qui donne au total environ 700 000 personnes. La population de référence se compose de tous les membres des ménages résidant officiellement en Italie.

Le plan d'échantillonnage de l'EICPA comporte deux étapes : 1) les municipalités ont été désignées comme les unités primaires d'échantillonnage (UPE) avec stratification, et les ménages ont été désignés comme les unités finales d'échantillonnage (UFE) avec renouvellement. Les UPE ont été stratifiées en fonction de la taille de la population. Les grandes municipalités, dont la population dépasse un seuil donné (également appelées les municipalités auto-représentatives), étaient toujours incluses dans l'échantillon; les petites municipalités (non auto-représentatives) étaient regroupées en strate, de manière à ce qu'une municipalité par strate soit sélectionnée avec une probabilité proportionnelle à sa population; 2) les ménages étaient sélectionnés au hasard à partir des registres de population dans toutes les municipalités tirées à l'étape 1.

L'enquête a été réalisée tous les trimestres en fonction d'un plan par renouvellement de type 2-2-2. Les membres des ménages ont été interviewés pendant deux trimestres consécutifs. Après une pause de deux trimestres, ils ont été interviewés à nouveau, deux fois pendant les deux trimestres correspondants de l'année suivante. Par conséquent, chaque ménage a été inclus dans quatre cycles de l'enquête pendant une période de 15 mois. Ce système par renouvellement signifiait que la moitié de l'échantillon est demeuré inchangé au cours de deux trimestres consécutifs et des trimestres à un an d'intervalle, et que 25 % de l'échantillon est demeuré inchangé pendant trois trimestres.

Toutes les analyses statistiques suivantes sont faites sur la population longitudinale. L'EICPA n'est pas conçue comme une enquête par panel en bonne et due forme : la population initiale change pendant la période à l'étude en raison d'événements démographiques et de migrations. Bien que l'ISTAT ait proposé une procédure pour calculer les poids longitudinaux (Boschetto, Discenza, Lucarelli, Rosati et Fiori 2009), ces derniers ne sont pas à la disposition des chercheurs, ce qui fait que nous n'avons pas pu prendre en considération le plan d'échantillonnage complexe. Toutefois, nous considérons qu'il était raisonnable de présumer que les répondants appartenant aux mêmes ménages étaient indépendants.

L'information sur l'état sur le marché du travail dans un trimestre de référence a été recueillie trois fois : (i) chaque répondant était classifié comme occupant un emploi, en chômage ou inactif, conformément à la définition du BIT en fonction des réponses données à un groupe de questions en particulier; (ii) dans une section subséquente du questionnaire, tous les répondants devaient se classifier sur le marché du travail, afin de recueillir l'état « auto-évalué »; (iii) au bout d'un an, dans une question rétrospective, on demandait aux répondants quel était leur état sur le marché du travail un an avant la première interview.

D'après la définition du BIT, les répondants ont été classifiés comme occupant un emploi pendant le trimestre de référence s'ils avaient 15 ans ou plus pendant la période de référence et ont effectué un type quelconque de travail, pendant au moins une heure, contre rémunération, bénéfices ou gain familial, ou s'ils n'étaient pas au travail, mais ils avaient un emploi ou une entreprise, dont ils étaient absents temporairement en raison d'une maladie, de vacances, d'un conflit de travail ou d'études ou de formation. Les répondants ont été classifiés comme en chômage s'ils avaient 15 à 74 ans et étaient : (a) sans travail pendant la semaine de référence; (b) actuellement disponibles pour travailler dans les deux semaines suivant la semaine de référence; (c) activement à la recherche de travail, c'est-à-dire qu'ils avaient pris des mesures particulières pour trouver du travail, pendant la période de quatre semaines se terminant par la semaine de référence, ou s'ils n'étaient pas à la recherche de travail, mais avaient trouvé un emploi devant commencer plus tard, dans une période allant jusqu'à trois mois (Organisation internationale du travail (OIT) 2008).

L'auto-évaluation actuelle et la question rétrospective classifiaient les répondants dans huit catégories : occupant un emploi; en chômage et à la recherche d'un nouvel emploi; en chômage et à la recherche d'un premier emploi; s'acquittant de tâches ménagères; étudiants; retraités; ayant un handicap les empêchant de travailler; autres.

Le tableau 3.1 démontre le plan par renouvellement de l'enquête pour deux années civiles consécutives. Les lettres indiquent les groupes de renouvellement : quatre groupes de renouvellement ont été interviewés à chaque trimestre. En ce qui concerne une année civile, l'information sur l'état sur le marché du travail provenait de neuf groupes de renouvellement. Toutefois, le plan par renouvellement produit une tendance particulière de données manquantes. Par exemple, pour les unités du groupe de renouvellement A qui sont interviewées pour la quatrième fois au premier trimestre de l'année 1, seuls l'indicateur du BIT (B) et l'autoévaluation (A) de l'état sur le marché du travail au cours du premier trimestre de l'année 1 sont disponibles. Pour les unités du groupe de renouvellement F, qui ont été interviewées pour la première fois au premier trimestre de l'année 1, nous avons seulement de l'information sur l'état sur le marché du travail

en fonction de la définition du BIT, de l'auto-évaluation et de la question rétrospective (R) pour les deux premiers trimestres de l'année 1.

Tableau 3.1
Plan par renouvellement de l'EICPA

Groupe de renouvellement	Année 1				Année 2			
	Trimestre I	Trimestre II	Trimestre III	Trimestre IV	Trimestre I	Trimestre II	Trimestre III	Trimestre IV
A	B-A							
B	B-A	B-A						
C		B-A	B-A					
D			B-A	B-A				
E	B-A-R			B-A	B-A			
F	B-A-R	B-A-R			B-A	B-A		
G		B-A-R	B-A-R			B-A	B-A	
H			B-A-R	B-A-R			B-A	B-A
I				B-A-R	B-A-R			B-A
L					B-A-R	B-A-R		
M						B-A-R	B-A-R	
N							B-A-R	B-A-R
O								B-A-R

B = Indicateur du BIT, A = auto-évaluation de l'état sur le marché du travail, R = indicateur rétrospectif.

Nous avons examiné les données recueillies de 2005 à 2010. (Sont exclues de ces analyses les données recueillies en 2004, première année de la mise en œuvre de la nouvelle enquête sur la population active, parce que les données n'étaient pas complètement fiables; en ce qui concerne 2010, nous n'utilisons ici que l'information recueillie au moyen de la question rétrospective et portant sur l'état sur le marché du travail en 2009.) Le tableau 3.2 énumère la composition du marché du travail au premier trimestre à partir de données regroupées au cours de la période de cinq ans. L'indicateur du BIT compte manifestement un pourcentage plus faible de personnes en chômage et un plus fort pourcentage de personnes inactives que les deux autres indicateurs. Les deux mesures basées sur l'auto-évaluation donnent un plus haut taux de chômage, parce que le BIT applique une définition très stricte du chômage. Pour être réputés en chômage, les répondants de 15 à 74 ans ne doivent pas occuper un emploi au moment de l'interview, mais ils doivent être prêts à accepter des emplois convenables au cours des deux prochaines semaines si l'occasion se présente, et ils doivent avoir cherché activement des façons de décrocher un emploi dans les deux semaines précédentes. Le BIT fournit ces lignes directrices afin de faciliter les comparaisons du rendement sur le marché du travail au fil du temps et d'un pays à un autre (OIT 2008). Cependant, ce cadre a été établi à un moment où le type d'emploi prédominant était à temps plein et sous contrat permanent; depuis, la situation d'emploi est devenue plus souple, les emplois à temps partiel et temporaires s'étant multipliés, en particulier pour les personnes sur le point d'entrer sur le marché du travail.

Tableau 3.2
Composition du marché du travail 2005 – Trimestre I de 2009, % – données regroupées

	E	C	I
BIT	43,07	3,60	53,33
A	41,73	6,73	51,54
R	41,55	6,49	51,96

E = Personnes ayant un emploi, C = Personnes en chômage, I = Personnes inactives.

D'autres études dans la littérature démontrent que la distinction entre les états sur le marché du travail n'est pas toujours évidente : les personnes ne connaissent peut-être pas les définitions officielles ou elles peuvent considérer leur état sur le marché du travail comme différent des critères standard (par exemple, voir Clark et Summer 1979; Flinn et Heckman 1983; Gonul 1992). Dans la plupart des cas, il est difficile de faire la distinction entre les personnes en chômage et les personnes inactives : la condition la plus importante semble être la recherche active d'un emploi, puisque les répondants peuvent se considérer comme en chômage même s'ils ne cherchent pas activement un emploi. Des incohérences peuvent donc survenir entre l'information recueillie dans le cadre d'enquêtes et le comportement réel. Une autre explication des différences entre les classifications du BIT et de l'auto-évaluation est que les répondants ayant un emploi temporaire pour ce qui est du nombre d'heures par semaine peuvent ne pas se considérer comme ayant un emploi.

Le tableau 3.3 énumère les incohérences, c'est-à-dire les différents états sur le marché du travail observés pour le même répondant avec deux indicateurs, parmi les trois indicateurs pour la période en question. Les données au fil des trimestres et des années ont été regroupées pour des motifs d'espace. Le nombre d'incohérences est manifestement plus élevé pour l'état de chômage que pour les deux autres états, et la plupart des mauvaises classifications ont tendance à désigner des personnes hors de la population active plutôt que des personnes occupant un emploi, comme l'ont démontré bien des études précédentes (par exemple, voir Poterba et Summers 1986). En comparant l'état sur le marché du travail conformément à la définition du BIT à celui déclaré d'après les réponses à la question rétrospective, on a obtenu le plus grand nombre d'incohérences. En examinant les cohérences au fil des trimestres et des années pour deux des trois indicateurs (non déclarés ici faute d'espace), nous constatons que la cohérence a tendance à augmenter légèrement au fil du temps, peut-être parce que tous les intervenants participant au processus d'enquête – intervieweurs, répondants, etc. – apprennent comment recueillir et fournir des renseignements de bonne qualité pendant qu'ils participent à l'enquête. Bien que nous n'ayons pas observé d'effets saisonniers pour ce qui est du nombre d'incohérences, le nombre d'incohérences indiquait une erreur de mesure non négligeable dans les données, ce qui veut dire qu'un des deux indicateurs a été mal déclarés, voire les deux.

Tableau 3.3
Incohérences 2005 – 2009, % – données regroupées

	EC	EI	CE	CI	IE	IC
BIT – Auto-évaluation	0,97	1,72	0,44	13,02	0,17	5,80
BIT – Rétrospective	1,14	2,06	5,22	16,76	1,00	5,76
Auto-évaluation – Rétrospective	0,92	1,62	6,03	8,73	1,00	0,89

EC = Considéré comme ayant un emploi avec le premier indicateur, mais en chômage avec le deuxième indicateur.
EI = Considéré comme ayant un emploi avec le premier indicateur, mais inactif avec le deuxième indicateur.
CE = Considéré comme en chômage avec le premier indicateur, mais ayant un emploi avec le deuxième indicateur.
CI = Considéré comme en chômage avec le premier indicateur, mais inactif avec le deuxième indicateur.
IE = Considéré comme inactif avec le premier indicateur, mais ayant un emploi avec le deuxième indicateur.
IC = Considéré comme inactif avec le premier indicateur, mais en chômage avec le deuxième indicateur.

Cependant, les incohérences ressorties des tableaux 3.2 et 3.3 peuvent également se produire parce que les trois indicateurs sont exposés à une erreur de mesure. Des études antérieures ont examiné les causes de la mauvaise perception de l'état sur le marché du travail et conclu que la perception est influencée par des facteurs sociaux, démographiques, économiques et institutionnels (par exemple Richiardi 2002). Les incohérences entre les deux auto-évaluations (réelle et rétrospective) peuvent être en grande partie

attribuables à l'érosion de la mémoire (Bound, Brown et Mathiowetz 2001). Enfin, la plus forte cohérence entre les indicateurs d'auto-évaluation suggère la possibilité d'erreurs de mesure corrélées.

Le tableau 3.4 indique les probabilités de transition trimestrielles parmi les trois états de la population active, du premier au deuxième trimestre des années 2005 à 2009 avec les trois indicateurs. L'indicateur du BIT décrit un marché du travail beaucoup plus dynamique, en particulier pour les répondants en chômage, que les indicateurs basés sur une auto-évaluation et une rétrospective. Cette différence est une autre preuve qui signale une erreur de mesure dans les données. D'après la littérature existante, nous savons que même de petits degrés d'erreur de classification peuvent entraîner un biais grave de l'estimation des probabilités de transition (Hagenaars 1994; Pavlopoulos, Muffles et Vermunt 2012). Si les erreurs ne sont pas corrélées au fil du temps, nous pouvons nous attendre à observer un marché du travail plus dynamique que le vrai marché, et le contraire si la corrélation des erreurs au fil du temps existe également.

Le tableau 3.5 compare les mouvements bruts observés, par exemple du premier au deuxième trimestre de 2005, selon le sexe et l'âge. On a obtenu les trois intervalles d'âge en divisant les échantillons en trois groupes de dimensions égales (c'est-à-dire 33^{e} et 66^{e} centiles). Plus précisément, pour l'année 2005, à l'âge 1, nous trouvons des répondants de 16 à 36 ans; à l'âge 2, ils ont de 36 à 55 ans, et à l'âge 3, ils ont de 56 à 75 ans. On constate que les femmes sont plus dynamiques que les hommes, en particulier pour ce qui est du chômage. Lorsqu'elles quittent une situation de chômage, les femmes ont tendance à quitter le marché du travail plus souvent que de commencer un emploi. Il y a également des variations importantes des mouvements bruts observés au fil du temps. Les répondants les plus âgés étaient plus stables lorsqu'ils étaient inactifs et avaient de plus fortes probabilités de quitter le marché du travail que de commencer un emploi après avoir occupé un emploi. Les jeunes répondants avaient de plus faibles probabilités de quitter une situation de chômage et d'inactivité en trouvant un emploi que ceux du deuxième groupe d'âge. Ces constatations portent à croire que le sexe et l'âge devraient être inclus comme covariables dans notre modèle, afin d'estimer les mouvements bruts corrigés sur le marché du travail.

Tableau 3.4
Mouvements bruts observés, du trimestre I au trimestre II de 2005 à 2009, %, Bureau international du travail (BIT), indicateur auto-évaluation (A) et rétrospective (R)

		EE	EC	EI	CE	CC	CI	IE	IC	II
2005	BIT	96,49	0,87	2,63	18,97	50,50	30,53	1,49	1,99	96,52
	A	96,99	1,33	1,69	15,32	69,85	14,83	1,29	1,50	97,21
	R	95,32	2,10	2,58	20,96	59,56	19,48	1,96	2,22	95,81
2006	BIT	96,13	0,78	3,09	20,40	45,21	34,39	2,42	1,74	95,84
	A	96,11	1,74	2,16	19,84	63,66	16,50	1,88	1,75	96,37
	R	95,55	1,72	2,73	17,93	66,57	15,50	2,00	1,75	96,25
2007	BIT	96,22	0,68	3,10	21,45	40,41	38,14	2,21	1,78	96,02
	A	96,08	1,74	2,16	19,84	63,66	16,50	1,88	1,75	96,37
	R	95,66	1,78	2,56	19,95	60,67	19,38	2,26	1,93	95,80
2008	BIT	97,05	0,80	2,16	19,82	48,50	31,68	1,87	1,87	96,26
	A	96,92	1,54	1,53	15,25	70,84	13,92	1,56	1,69	96,75
	R	95,76	2,13	2,11	19,04	62,60	18,36	2,02	2,26	95,72
2009	BIT	96,58	0,88	2,54	18,41	48,10	33,49	2,08	1,83	96,09
	A	96,14	1,76	2,10	15,17	70,09	14,75	1,59	1,61	96,80
	R	95,45	1,88	2,66	16,88	67,15	15,97	1,78	1,89	96,33

EE = Occupant un emploi au cours des deux trimestres.
EC = Occupant un emploi au premier trimestre, et en chômage au deuxième.
EI = Occupant un emploi au premier trimestre, et inactif au deuxième.
CE = En chômage au premier trimestre, et occupant un emploi au deuxième.
CC = En chômage au cours des deux trimestres.
CI = En chômage au premier trimestre, et inactif au deuxième.
IE = Inactif au cours du premier trimestre, et occupant un emploi au deuxième.
IC = Inactif au cours du premier trimestre, et au chômage au deuxième.
II = Inactif au cours des deux trimestres.

Tableau 3.5
Mouvements bruts observés du trimestre I au trimestre II de 2005, selon le sexe et l'âge, %, Bureau international du travail (BIT), indicateurs auto-évaluation (A) et rétrospective (R)

		EE	EC	EI	CE	CC	CI	IE	IC	II
Hommes	BIT	97,20	0,78	2,02	22,73	51,60	25,68	1,93	2,07	96,00
	A	97,63	1,08	1,29	18,97	73,80	7,23	1,36	1,10	97,53
	R	96,13	1,84	2,03	26,14	65,27	8,60	2,13	1,50	96,37
Femmes	BIT	95,43	1,01	3,57	15,70	49,31	34,99	1,23	1,98	96,79
	A	96,00	1,69	2,31	11,93	65,73	22,34	1,26	1,81	96,93
	R	94,14	2,46	3,40	16,24	53,56	30,19	1,86	2,71	95,43
Âge 1	BIT	88,27	0,46	11,27	21,16	27,50	51,35	0,26	0,06	99,67
	A	89,66	0,56	9,78	10,20	60,09	29,71	0,31	0,10	99,60
	R	83,36	0,45	16,19	20,78	42,54	36,68	0,51	0,13	99,36
Âge 2	BIT	97,65	0,55	1,80	21,62	43,01	35,37	2,72	2,95	94,33
	A	97,87	0,92	1,20	16,83	64,65	18,52	2,52	2,61	94,87
	R	97,04	1,23	1,74	24,60	53,42	21,98	4,05	4,24	91,70
Âge 3	BIT	96,18	1,32	2,50	17,54	51,14	31,32	3,81	6,75	89,44
	A	96,83	1,89	1,28	14,77	71,97	13,27	3,17	4,82	92,01
	R	94,82	3,29	1,89	19,62	63,60	16,78	4,52	6,68	88,80

EE = Occupant un emploi au cours des deux trimestres.
EC = Occupant un emploi au premier trimestre, et en chômage au deuxième.
EI = Occupant un emploi au premier trimestre, et inactif au deuxième.
CE = En chômage au premier trimestre, et occupant un emploi au deuxième.
CC = En chômage au cours des deux trimestres.
CI = En chômage au premier trimestre, et inactif au deuxième.
IE = Inactif au cours du premier trimestre, et occupant un emploi au deuxième.
IC = Inactif au cours du premier trimestre, et en chômage au deuxième.
II = Inactif au cours des deux trimestres.

4 Résultats : Comparaisons des MMCL mixte et standard

Nous estimons plusieurs spécifications des MMCL standard et mixtes. Le modèle standard comporte deux composantes : structure, décrivant la dynamique réelle parmi les variables latentes (vrais états) par une chaîne markovienne de premier ordre; et mesure, qui apparie chaque variable latente à ses indicateurs (états observés sur le marché du travail). Certaines restrictions comportant de l'information a priori et/ou des suppositions sont imposées aux paramètres de la composante de mesure, en fonction de la preuve ressortie des données observées (incohérences et transitions) et des conclusions de la littérature sur la méthodologie de l'enquête et la psychologie cognitive sur le mécanisme entraînant des erreurs. Seulement quatre des neuf groupes de renouvellement fournissant de l'information visant une année civile ont été interviewés à tous les trimestres, et nous avons les trois indicateurs des états sur le marché du travail seulement pour deux de ces groupes (voir le tableau 3.1). Pour les deux autres groupes, nous ne recueillons pas l'information avec la question rétrospective. La tendance de l'information manquante en raison du plan par renouvellement de l'enquête est incluse dans les MMCL estimés comme des données manquantes sur une base aléatoire.

Tous les modèles estimés partagent les caractéristiques suivantes : les transitions réelles suivent une chaîne markovienne de premier ordre. En raison du plan d'enquête, aucune personne n'était observée pendant trois cycles consécutifs, c'est-à-dire qu'une chaîne markovienne de deuxième ordre ne peut pas être estimée, puisque les statistiques suffisantes relatives sont absentes. Cependant, bien que l'état sur le marché du travail au cours d'un trimestre donné risque fort de toucher l'état au cours du trimestre subséquent, il est beaucoup moins probable que l'incidence soit significative au bout de deux trimestres. On présume que les erreurs de classification sont constantes au fil du temps pour chaque indicateur; la supposition de l'ECI est incluse. L'ajustement du modèle est évaluée par l'indice BIC (pour *Bayesian information criterion*) en raison de la grande taille de l'échantillon (moyenne de 250 000 unités par année; voir le tableau 4.1).

La spécification d'un MMCL mixte est également recommandée par le fait que l'échantillon peut renfermer plusieurs groupes de répondants ayant des comportements différents sur le marché du travail. Comme susmentionné, la littérature récente démontre qu'en ne tenant pas compte de l'hétérogénéité non observée des transitions pour estimer les MMCL, on risque d'entraîner des estimations biaisées de l'erreur de mesure (Magidson et coll. 2007). De plus, un MMCL mixte peut améliorer l'ajustement des données.

Nous estimons un MMCL mobile-stable en supposant des erreurs de mesure constantes dans les deux groupes latents. Il convient de souligner que tous les modèles estimés ont été relevés, et que, pour réduire le risque de détection des maxima locaux, l'estimation a été effectuée à plusieurs reprises avec différents ensembles de valeurs de départ. Le logiciel Latent Gold 4.5 a été mis en œuvre (Vermunt et Magidson 2008).

Le tableau 4.1 compare les MMCL mixte et standard adaptés à nos cinq échantillons de données, qui visent les années de 2005 à 2009, et au moyen de l'indice BIC. Le modèle mixte démontre une meilleure correspondance pour tous les échantillons. Le tableau 4.2 indique les pourcentages de personnes mobiles et stables au premier trimestre de 2005, ainsi que la répartition de deux groupes non observés au premier trimestre de chaque année. De toute évidence, l'hétérogénéité non observée est fortement corrélée à l'état initial et, comme il fallait s'y attendre, les personnes stables occupent un emploi ou sont inactives, c'est-à-dire qu'un très faible pourcentage d'entre elles sont en chômage.

Tableau 4.1
Comparaison des MMCL standard et mixte : indice BIC

Année	n	Standard	Mixte
2005	220 051	650 241	649 401
2006	206 037	587 794	587 058
2007	274 484	748 788	748 654
2008	277 363	667 399	666 335
2009	274 723	747 997	746 991

Tableau 4.2
MMCL mixte : proportion de personnes mobiles et stables et répartition de l'état initial en 2005, trimestre I, %

	Proportion	E	C	I
Mobiles	10,23	39,85	39,09	21,06
Stables	81,79	41,79	3,36	54,85

E = Personnes ayant un emploi, C = Personnes en chômage, I = Personnes inactives.

Comme le démontrent les données dans les tableaux 4.3-4.5 (la composition du marché du travail, les transitions estimées et les erreurs de mesure estimées démontrent la même tendance au cours des trois autres trimestres de chaque année), la meilleure correspondance aux données du modèle mixte est entièrement attribuable aux différents taux de transition estimés; la composition du marché du travail et les erreurs de mesure estimées sont les mêmes dans les deux modèles. Ce résultat vient contredire celui qui avait été obtenu par Magidson et coll. (2007), qui ont comparé le modèle des personnes mobiles-stables et le MMCL standard appliqué aux transitions sur le marché du travail de l'enquête sur la population active. Les auteurs susmentionnés ont constaté que le MMCL mixte convient mieux aux données que le MMCL standard, et que ce dernier, qui ne tient pas compte de l'hétérogénéité non observée, surestime le degré d'erreur de mesure en ce qui concerne le modèle des personnes mobiles-stables. Plus précisément, les auteurs

susmentionnés ont utilisé des résultats simulés pour estimer une infraction aux probabilités de transition homogènes, de sorte que l'hétérogénéité corrélée avec l'état initial produise des estimations exagérées des erreurs de mesure dans un MMCL standard.

Tableau 4.3
Comparaison des MMCL standard et mixte : composition du marché du travail au trimestre I de 2005, %

		E	C	I
2005	Standard	41,67	7,00	51,33
	Mixte	41,59	7,02	51,39

E = Personnes ayant un emploi, C = Personnes en chômage, I = Personnes inactives.

Tableau 4.4
Comparaison des MMCL standard et mixte : transitions estimées du trimestre I au trimestre II de 2005 – 2009, %

		EE	EC	EI	CE	CC	CI	IE	IC	II
2005	Standard	97,36	1,32	1,32	15,59	76,18	8,23	0,57	0,74	98,69
	Mixte	96,46	1,68	1,86	19,61	69,65	10,74	0,91	1,09	98,00
2006	Standard	96,75	1,68	1,56	19,52	71,27	9,21	1,01	0,99	90,00
	Mixte	96,22	1,92	1,87	22,11	66,96	10,93	1,25	1,22	97,54
2007	Standard	96,69	1,67	1,64	18,84	70,56	10,60	1,01	0,99	98,00
	Mixte	96,42	1,80	1,78	20,22	67,80	11,98	1,10	1,45	95,45
2008	Standard	97,56	1,41	1,03	15,86	79,73	4,42	0,53	0,62	98,85
	Mixte	96,45	1,89	1,66	19,56	73,25	7,19	0,83	0,89	98,28
2009	Standard	96,85	1,71	1,44	14,04	75,33	9,63	1,04	1,01	97,95
	Mixte	96,27	1,95	1,78	17,09	71,16	11,75	1,30	1,22	97,48

EE = Occupant un emploi au cours des deux trimestres.
EC = Occupant un emploi au premier trimestre, et en chômage au deuxième.
EI = Occupant un emploi au premier trimestre, et inactif au deuxième.
CE = En chômage au premier trimestre, et occupant un emploi au deuxième.
CC = En chômage au cours des deux trimestres.
CI = En chômage au premier trimestre, et inactif au deuxième.
IE = Inactif au cours du premier trimestre, et occupant un emploi au deuxième.
IC = Inactif au cours du premier trimestre, et en chômage au deuxième.
II = Inactif au cours des deux trimestres.

Tableau 4.5
Comparaison des MMCL standard et mixte : erreurs de mesure estimées au trimestre I de 2005 – 2009, %, indicateur du BIT

		EE	EC	EI	CE	CC	CI	IE	IC	II
2005	Standard	99,82	0,01	0,17	6,17	45,04	48,80	0,89	0,50	98,61
	Mixte	99,82	0,01	0,17	6,16	45,06	48,78	0,90	0,51	98,59
2006	Standard	99,83	0,01	0,16	6,50	41,92	51,58	0,75	0,45	98,80
	Mixte	99,87	0,01	0,13	5,17	37,28	57,55	0,68	0,40	98,92
2007	Standard	99,75	0,01	0,24	6,84	39,83	53,34	0,75	0,47	98,79
	Mixte	99,75	0,01	0,24	6,77	39,92	53,31	0,77	0,47	98,76
2008	Standard	99,83	0,01	0,17	3,81	42,45	53,74	0,61	0,38	99,02
	Mixte	99,83	0,01	0,17	3,82	42,41	53,76	0,62	0,38	99,00
2009	Standard	95,34	0,98	3,68	18,30	41,17	40,53	2,06	1,61	96,33
	Mixte	95,22	2,34	2,44	15,60	68,02	16,37	1,74	2,14	96,13

EE = Réellement occupé et réputé occupé par l'indicateur du BIT.
EC = Réellement occupé, mais réputé en chômage par l'indicateur du BIT.
EI = Réellement occupé, mais réputé inactif par l'indicateur du BIT.
CE = Réellement en chômage, mais réputé occupé par l'indicateur du BIT.
CC = Réellement en chômage et réputé en chômage par l'indicateur du BIT.
CI = Réellement en chômage, mais réputé inactif par l'indicateur du BIT.
IE = Réellement inactif, mais réputé occupé par l'indicateur du BIT.
IC = Réellement inactif, mais réputé en chômage par l'indicateur du BIT.
II = Réellement inactif et réputé inactif par l'indicateur du BIT.

Le modèle des personnes mobiles-stables décrit un marché du travail plus dynamique, en particulier pour les répondants en chômage : la probabilité de demeurer en chômage pendant le trimestre est plus faible que prévu par le modèle standard.

5 Résultats : MMCL mixte avec covariables et erreurs de mesure corrélées

Les résultats présentés à la section précédente indiquaient qu'un MMCL mixte convient mieux à nos données. Comme le MMCL standard, le MMCL mixte tient compte de la mauvaise classification et de la tendance des données manquantes à présumer de cette dernière de façon aléatoire, et il comprend également l'hétérogénéité non observée. La supposition que les données sont manquantes de façon aléatoire s'explique par le fait que chaque groupe de renouvellement est observé au cours de deux trimestres, mais pas au cours des deux trimestres subséquents, et que ces données sont manquantes en raison du plan et qu'elles ne dépendent pas de l'état réel ou déclaré des répondants ou d'autres variables non observées. Pour estimer nos modèles, nous avons utilisé simultanément de l'information de tous les groupes de renouvellement, c'est-à-dire une technique de calcul du maximum de vraisemblance à information complète. Les données des mouvements bruts observés, en particulier le fait que la mobilité observée est très différente entre les sexes et les âges (tableau 3.5), indiquaient que l'estimation d'un MMCL mixte avec ces deux covariables avait une incidence sur les transitions latentes.

Divers modèles ont été estimés avec la caractéristique commune suivante : les transitions mobile-stable et les transitions latentes suivent une chaîne markovienne de premier ordre. Afin de préciser le modèle de mesure, les facteurs suivants ont été pris en considération : (i) la réponse à la question sur l'état autoévalué sur le marché du travail est donnée pendant la même interview une fois que les répondants ont répondu aux questions sur lesquelles repose l'indicateur du BIT; (ii) toutefois, l'indicateur du BIT est déterminé par l'ISTAT en fonction des réponses données à une série de questions respectant les lignes directrices du BIT, tandis que A représente les auto-évaluations des répondants : il est plausible que les répondants ne soient pas au courant de la classification de l'ISTAT; (iii) l'indicateur A et l'indicateur découlant de l'interrogation rétrospective décrivent un marché du travail plus stable que celui du BIT et montrent le plus haut niveau de cohérence : les répondants peuvent être influencés par les réponses données au trimestre précédent; (iv) l'information pour R est recueillie un an après les réponses au BIT et A; (v) pour les personnes qui sont dans un état stable, la déclaration correcte de l'état sur le marché du travail est une tâche cognitive plus facile que pour les personnes qui traversent au moins un changement, et elle peut donc indiquer de plus fortes probabilités de donner des réponses incorrectes.

Parmi les diverses spécifications possibles, le modèle convenant le mieux, pour toutes les années analysées, consistait à présumer que les personnes stables déclarent leur état sur le marché du travail correctement et que, pour les personnes mobiles, les erreurs de mesure sont constantes au fil du temps et que les deux indicateurs basés sur l'auto-évaluation, A et R, sont corrélés, c'est-à-dire qu'un effet direct entre ces deux indicateurs est intégré à la spécification du modèle. (Tous les modèles estimés ont été cernés et, pour éviter les maxima locaux, une estimation a été effectuée plusieurs fois avec différents ensembles de valeurs de départ; pour estimer des modèles plus parcimonieux, les trois interactions des variables ont été

établies à 0.) À titre d'exemple, les tableaux 5.1 à 5.3 indiquent quelques-uns des résultats des estimations : la composition du marché du travail et les mouvements estimés pour l'ensemble de la population, les personnes mobiles et stables collectivement (l'ensemble complet des résultats des estimations peut être obtenu auprès des auteurs) et les erreurs de mesure estimées. En moyenne, au cours des cinq ans, le pourcentage de personnes mobiles était de 17,69.

Tableau 5.1
Composition estimée du marché du travail au trimestre I de 2005 – 2009, %

	2005	2006	2007	2008	2009
E	42,01	42,36	40,72	40,92	40,00
C	5,93	5,64	5,75	5,27	6,46
I	52,07	52,00	53,53	53,81	53,53

E = Personnes ayant un emploi, C = Personnes en chômage, I = Personnes inactives.

Tableau 5.2
Mouvements bruts estimés du trimestre I au trimestre II de 2005 - 2009, %, erreurs types entre parenthèses

	EE	EC	EI	CE	CC	CI	IE	IC	II
2005	96,70	1,60	1,61	17,41	71,80	10,78	0,97	0,70	98,29
	(0,0017)	(0,0012)	(0,0012)	(0,0133)	(0,0142)	(0,0079)	(0,0013)	(0,0011)	(0,0017)
2006	96,10	1,93	1,93	19,16	67,04	13,80	1,71	0,89	97,41
	(0,0027)	(0,0020)	(0,0020)	(0,0112)	(0,0150)	(0,0136)	(0,0011)	(0,0015)	(0,0018)
2007	96,30	1,79	1,89	18,11	67,95	13,94	1,42	1,24	97,34
	(0,0023)	(0,0016)	(0,0017)	(0,0145)	(0,0158)	(0,0094)	(0,0018)	(0,0018)	(0,0025)
2008	96,88	1,77	1,35	18,00	74,57	7,43	1,61	1,03	97,37
	(0,0037)	(0,0027)	(0,0028)	(0,0118)	(0,0157)	(0,0138)	(0,0013)	(0,0017)	(0,0020)
2009	96,50	1,83	1,62	15,04	71,62	13,35	1,55	1,10	97,35
	(0,0024)	(0,0019)	(0,0016)	(0,0153)	(0,0168)	(0,0092)	(0,0019)	(0,0014)	(0,0024)

EE = Occupant un emploi au cours des deux trimestres.
EC = Occupant un emploi au premier trimestre, et en chômage au deuxième.
EI = Occupant un emploi au premier trimestre, et inactif au deuxième.
CE = En chômage au premier trimestre, et occupant un emploi au deuxième.
CC = En chômage pendant les deux trimestres.
CI = En chômage au premier trimestre, et inactif au deuxième.
IE = Inactif au cours du premier trimestre, et occupant un emploi au deuxième.
IC = Inactif au cours du premier trimestre, et en chômage au deuxième.
II = Inactif pendant les deux trimestres.

Tableau 5.3a
Erreurs de mesure estimées de 2005 – 2009, indicateur du BIT, %, erreurs types entre parenthèses

	EE	EC	EI	CE	CC	CI	IE	IC	II
2005	99,75	0,02	0,23	0,93	89,72	9,36	0,97	1,04	98,00
	(0,0002)	(0,0001)	(0,0001)	(0,0028)	(0,0050)	(0,0051)	(0,0004)	(0,0003)	(0,0005)
2006	99,75	0,01	0,24	1,17	89,39	9,44	0,55	0,99	98,46
	(0,0007)	(0,0004)	(0,0005)	(0,0025)	(0,0042)	(0,0035)	(0,0003)	(0,0002)	(0,0004)
2007	99,82	0,01	0,24	0,84	88,28	10,88	0,58	0,87	98,55
	(0,0002)	(0,0001)	(0,0002)	(0,0028)	(0,0050)	(0,0051)	(0,0004)	(0,0003)	(0,0005)
2008	99,44	0,10	0,46	1,16	89,36	9,48	0,57	1,38	90,05
	(0,0007)	(0,0004)	(0,0005)	(0,0025)	(0,0042)	(0,0035)	(0,0003)	(0,0002)	(0,0004)
2009	99,77	0,01	0,22	0,43	88,98	10,57	0,33	0,86	98,79
	(0,0001)	(0,0000)	(0,0001)	(0,0025)	(0,0038)	(0,0039)	(0,0003)	(0,0002)	(0,0003)

EE = Réellement occupé et réputé occupé par l'indicateur du BIT.
EC = Réellement occupé, mais réputé en chômage par l'indicateur du BIT.
EI = Réellement occupé, mais réputé inactif par l'indicateur du BIT.
CE = Réellement en chômage, mais réputé occupé par l'indicateur du BIT.
CC = Réellement en chômage et réputé en chômage par l'indicateur du BIT.
CI = Réellement en chômage, mais réputé inactif par l'indicateur du BIT.
IE = Réellement inactif, mais réputé occupé par l'indicateur du BIT.
IC = Réellement inactif, mais réputé en chômage par l'indicateur du BIT.
II = Réellement inactif et réputé inactif par l'indicateur du BIT.

Tableau 5.3b
Erreurs de mesure estimées de 2005 – 2009, indicateurs A et R, %, erreurs types entre parenthèses

	État réel	AR								
		EE	EC	EI	CE	CC	CI	IE	IC	II
2005	E	94,83	1,17	2,28	0,22	0,18	0,11	0,44	0,07	0,70
		(0,0008)	(0,0006)	(0,0005)	(0,0002)	(0,0001)	(0,0002)	(0,0003)	(0,0004)	(0,0003)
	C	0,01	0,00	0,00	0,97	97,16	1,11	0,09	0,31	0,35
		(0,0001)	(0,0001)		(0,0006)	(0,0008)	(0,0004)	(0,0009)	(0,0004)	(0,0003)
	I	0,00	0,00	0,01	0,12	0,70	0,70	0,78	0,98	96,72
				(0,0001)	(0,0005)	(0,0009)	(0,0008)	(0,0004)	(0,0006)	(0,0008)
2006	E	94,86	0,96	2,21	0,16	0,11	0,10	0,45	0,06	1,06
		(0,0052)	(0,0006)	(0,0005)	(0,0001)	(0,0002)	(0,0009)	(0,0001)	(0,0004)	(0,0003)
	C	0,00	0,01	0,00	0,86	97,98	0,50	0,11	0,32	0,22
			(0,0001)		(0,0001)	(0,0006)	(0,0001)	(0,0002)	(0,0003)	(0,0003)
	I	0,01	0,00	0,01	0,13	0,82	0,74	0,71	0,74	96,83
		(0,0001)		(0,0001)	(0,0006)	(0,0005)	(0,0004)	(0,0004)	(0,0001)	(0,0005)
2007	E	95,17	1,06	1,06	0,16	0,11	0,10	0,45	0,06	0,82
		(0,0009)	(0,0003)	(0,0005)	(0,0002)	(0,0004)	(0,0005)	(0,0006)	(0,0004)	(0,0004)
	C	0,00	0,01	0,00	0,90	97,74	0,73	0,09	0,31	0,21
			(0,0001)		(0,0005)	(0,0009)	(0,0003)	(0,0005)	(0,0004)	(0,0002)
	I	0,01	0,01	0,01	0,15	0,59	0,66	1,10	0,89	96,59
		(0,0001)	(0,0001)	(0,0001)	(0,0005)	(0,0006)	(0,0008)	(0,0004)	(0,0004)	(0,0020)
2008	E	94,65	1,48	1,83	0,16	0,02	0,14	0,72	0,04	0,96
		(0,0006)	(0,0009)	(0,0005)	(0,0003)	(0,0006)	(0,0004)	(0,0003)	(0,0004)	(0,0002)
	C	0,00	0,03	0,00	1,32	97,39	0,82	0,05	0,33	0,05
			(0,0001)		(0,0002)	(0,0010)	(0,0009)	(0,0005)	(0,0004)	(0,0004)
	I	0,01	0,02	0,01	0,17	0,45	1,34	1,05	1,50	95,45
		(0,0001)	(0,0001)	(0,0001)	(0,0009)	(0,0005)	(0,0003)	(0,0006)	(0,0004)	(0,0003)
2009	E	96,11	0,65	1,21	0,12	0,24	0,10	0,42	0,10	1,04
		(0,0004)	(0,0002)	(0,0001)	(0,0002)	(0,0003)	(0,0008)	(0,0009)	(0,0008)	(0,0009)
	C	0,01	0,01	0,00	0,59	98,23	0,55	0,08	0,26	0,25
		(0,0001)	(0,0001)		(0,0004)	(0,0004)	(0,0002)	(0,0005)	(0,0006)	(0,0006)
	I	0,01	0,00	0,01	0,08	0,76	0,52	0,74	0,78	97,08
		(0,0001)		(0,0001)	(0,0004)	(0,0002)	(0,0002)	(0,0004)	(0,0003)	(0,0008)

E = Personnes ayant un emploi, C = Personnes en chômage, I = Personnes inactives.
EE = Classifié comme occupant un emploi par les indicateurs auto-évaluation et rétrospective.
EC = Classifié comme occupant un emploi par l'indicateur auto-évaluation et en chômage par l'indicateur rétrospective.
EI = Classifié comme occupant un emploi par l'indicateur auto-évaluation et en inactif par l'indicateur rétrospective.
CE = Classifié comme en chômage par l'indicateur auto-évaluation et occupant un emploi par l'indicateur rétrospective.
CC = Classifié comme en chômage par les indicateurs auto-évaluation et rétrospective.
CI = Classifié comme en chômage par l'indicateur auto-évaluation et inactif par l'indicateur rétrospective.
IE = Classifié comme inactif par l'indicateur auto-évaluation et occupant un emploi par l'indicateur rétrospective.
IC = Classifié comme inactif par l'indicateur auto-évaluation et en chômage par l'indicateur rétrospective.
II = Classifié comme inactif par les indicateurs auto-évaluation et rétrospective.

La composition estimée du marché du travail au premier trimestre, comparativement à la composition observée (tableau 3.2), démontre un pourcentage de chômage légèrement inférieur à celui obtenu au moyen des deux indicateurs auto-évaluation, et plus élevé que celui de l'indicateur du BIT.

Les transitions estimées décrivent un marché du travail plus stable que celui qui a été observé avec les trois indicateurs, sauf deux transitions (voir le tableau 3.4). Les mouvements bruts estimés sont beaucoup plus semblables à ceux observés avec les questions reposant sur l'auto-évaluation et la rétrospective que ceux observés avec l'indicateur du BIT. Ce phénomène se manifeste également dans le cas de l'erreur de mesure estimée (tableau 5.3). Une objection immédiate à ce résultat serait que nous avons utilisé deux indicateurs très semblables (les deux auto-évaluations), ainsi qu'un troisième très différent (BIT). En fait, un résultat semblable – des erreurs de mesure plus faibles pour l'auto-évaluation que pour l'indicateur du BIT – a été obtenu en estimant un MMCL avec seulement deux indicateurs par variable latente : le BIT et l'auto-évaluation.

6 Conclusion

Le présent document propose une approche basée sur les classes latentes pour corriger les mouvements bruts à partir des erreurs corrélées. On met l'accent sur la capacité de tenir compte des erreurs de classification à l'étendue des données sur les panels, en raison du plan par renouvellement de l'enquête, qui entraîne des tendances aux données manquantes et à une hétérogénéité non observée.

L'approche des classes latentes a été appliquée aux transitions sur le marché du travail italien en fonction des trois états habituels (occupant un emploi, en chômage et inactif). Les données portent sur les années 2005 à 2009 et ont été recueillies dans le cadre de l'enquête italienne continue sur la population active auprès d'un échantillon de ménages italiens avec un plan par renouvellement de type 2-2-2 d'un trimestre à un autre. L'information sur l'état de la population active au cours d'un trimestre de référence a été recueillie à trois reprises : (i) les répondants ont été classifiés comme occupant un emploi, en chômage ou inactifs, conformément à la définition du Bureau international du travail en fonction des réponses à un groupe de questions en particulier; (ii) on a demandé aux répondants de se classifier comme occupant un emploi, en chômage ou inactifs (c'est-à-dire l'état auto-évalué); (iii) une question rétrospective visait à déterminer l'état sur le marché du travail un an plus tôt. Autrement dit, trois indicateurs de l'état sur le marché du travail étaient disponibles. Les trois indicateurs donnaient des descriptions très différentes du marché du travail italien, révélant un degré d'incohérence significatif. Ce phénomène indique une erreur de mesure dans les données.

Le modèle convenant le mieux était un MMCL reposant sur les personnes mobiles-stables, où les transitions latentes sur le marché du travail suivent une chaîne markovienne de premier ordre, les personnes stables déclarent toujours correctement leur état sur le marché du travail; pour les personnes mobiles, les erreurs de mesure étaient constantes au fil du temps et corrélées aux deux indicateurs d'auto-évaluation; le sexe et l'âge des répondants étaient inclus comme covariables; le plan par renouvellement de l'enquête était traité comme de l'information manquante sur une base aléatoire. Le modèle corrige les mouvements bruts observés vers un marché du travail plus stable et estime que l'indicateur de l'état sur le marché du travail basé sur la définition du BIT est touché par le plus grand degré d'erreur de mesure.

Deuxième conclusion : en cas d'hétérogénéité non observée, un MMCL mixte convient mieux aux données que le MMCL standard. Cette conclusion cadre avec d'autres rapports (par exemple Magidson et coll. 2007). Cependant, dans notre cas, les deux modèles estiment la même quantité d'erreur de mesure, la différence d'ajustement étant attribuable aux mouvements estimés. Au lieu de cela, les auteurs susmentionnés ont découvert une surestimation de l'erreur de mesure lorsque l'hétérogénéité non observée n'était pas prise en compte.

Un dernier facteur pris en considération a trait au plan d'échantillonnage de l'enquête, qui comporte deux degrés, comme indiqué à la section 3. Dans nos analyses, nous n'avons pas tenu compte du plan d'échantillonnage complexe, mais avons estimé les mouvements bruts de la population longitudinale fournie par l'*Italian Institute of Statistics*. Dans les recherches à venir, il serait intéressant de comparer comment les résultats peuvent être touchés par l'intégration de méthodes d'enquêtes à des échantillons complexes au moyen de notre stratégie d'estimation. Lu et Lohr (2010) ont déjà abordé le sujet d'une perspective intéressante.

Bibliographie

Bartolucci, F., Lupparelli, M. et Montanari, A. (2009). Latent Markov model for longitudinal binary data: An application to the performance evaluation of nursing homes. *Annals of Applied Statistics,* 3, 611-636.

Bassi, F., et Trivellato, U. (2009). A latent class approach for estimating gross flows in the presence of correlated classification errors. Dans *Methodology of Longitudinal Surveys,* (Éd., P. Lynn), Chichester: Wiley, 367-380.

Bassi, F., Padoan, A. et Trivellato, U. (2012). Inconsistencies in reported characteristics among employed stayers. *Statistica,* 1, 93-109.

Bassi, F., Torelli, N. et Trivellato, U. (1998). Stratégies de collecte de données et de modélisation dans l'estimation de flux bruts relatifs à la population active entachés d'erreurs de classification. *Techniques d'enquête,* 24, 2, 117-132. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1998002/article/4348-fra.pdf.

Bassi, F., Croon, M., Hagenaars, J.A. et Vermunt, J.K. (2000). Estimating true changes when categorical panel data are affected by correlated and uncorrelated classification errors. An application to unemployment data. *Sociological Methods and Research,* 29, 230-268.

Biemer, P.P., et Bushery, J.M. (2000). Validité de l'analyse markovienne de structure latente pour l'estimation de l'erreur de classification des données sur la population active. *Techniques d'enquête,* 26, 2, 157-171. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2000002/article/5534-fra.pdf.

Boschetto, B., Discenza, A.R., Lucarelli, C., Rosati, S. et Fiori, F. (2009). Longitudinal data for the analysis of Italian labor market flows. *Italian Journal of Applied Statistics,* 22, 129-150.

Bound, M., Brown C. et Mathiowetz N.A. (2001). Measurement error in survey data. Dans *Handbook of Econometrics,* (Éds., J.J. Heckman et E. Leamer), Amsterdam: Elsevier, 3705-3843.

Clark, K., et Summers, L.H. (1979). Labour market dynamics and unemployment: A reconsideration. *Brooking Papers on Economic Activity,* 1, 13-69.

Dayton, C.M., et McReady, G.B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association,* 83, 173-178.

Flinn, C.J., et Heckman, J.J. (1983). Are unemployment and out of the labour force behaviourally distinct market states? *Journal of Labour Economics,* 1, 28-42.

Gonul, F. (1992). New evidence on whether unemployment and out of the labour force are two distinct states. *Journal of Human Resources,* 27, 329-361.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika,* 61, 215-231.

Hagenaars, J.A. (1994). Latent variables in log-linear models of repeated observations. Dans *Latent Variable Analysis. Applications for Developmental Research,* (Éds., A. von Eye et C. Clogg), Thousand Oaks (CA): Sage, 329-352.

Langeheine, R. (1994). Latent variable Markov models. Dans *Latent Variable Analysis. Applications for Developmental Research,* (Éds., A. von Eye et C. Clogg), Thousand Oaks (CA): Sage, 373-395.

Lu, Y., et Lohr, S. (2010). L'estimation des flux bruts dans les enquêtes à base de sondage double. *Techniques d'enquête,* 36, 1, 13-24. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2010001/article/11248-fra.pdf.

Magidson, J., Vermunt, J.K. et Tran B. (2007). Using a mixture of latent Markov model to analyze longitudinal U.S. employment data involving measurement error. Dans *New Trends in Psychometrics,* (Éds., K. Shigemasu, A. Okada, T. Imaizumi et T. Hoshino), Tokyo: Universal Academy Press, 235-242.

Manzoni, A., Vermunt, J.K., Luijkx, R. et Muffels, R. (2010). Memory bias in retrospectively collected employment careers: A model-based approach to correct for measurement errors. *Sociological Methodology,* 40, 39-73.

Organisation internationale du travail (OIT) (2008). Résolution sur la mise à jour de la classification internationale type des professions, Genève. Article accessible à l'adresse http://www.ilo.org/public/french/bureau/stat/isco/docs/resol08.pdf.

Paas, L.J., Vermunt, J.K. et Bijmolt, T.H. (2007). Discrete-time discrete-state latent Markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society, Series A,* 170, 955-974.

Pavlopoulos, D., et Vermunt, J.K. (2015). Mesure de l'emploi temporaire. Les données d'enquête ou de registre disent-elles la vérité ? *Techniques d'enquête,* 41, 1, 205-224. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2015001/article/14151-fra.pdf.

Pavlopoulos, D., Muffles, R. et Vermunt, J.K. (2012). How real is mobility between low pay, high pay and non-employment? *Journal of the Royal Statistical Society, Series A,* 170, 749-773.

Poterba, J.M., et Summers, L.S. (1986). Reporting errors and labour market dynamics. *Econometrica,* 54, 1319-1338.

Richiardi, M. (2002). What does the ECHP tell us about labour status misperception? A journey in less known regions of labour discomfort. *LABORatorio Revelli,* document de travail nº 69.

van de Pol, F., et Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology,* 33, 231-247.

Vermunt, J.K. (1997). *Log-Linear Models for Event History*. Thousand Oaks (CA): Sage.

Vermunt, J.K. (2010). Longitudinal research using mixture models. Dans *Longitudinal Research with Latent Variables,* (Éds., K. van Montfort, J.H.L. Oud et A. Satorra), Heidelberg: Springer, 119-152.

Vermunt, J.K., et Magidson, J. (2008). *LG-Syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K., et Magidson, J. (2013). *Technical Guide for Latent Gold 5.0. Basic, Advanced, and Syntax.* Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K., Tran, B. et Magidson, J. (2008). Latent class models in longitudinal research. Dans *Handbook of Longitudinal Research: Design, Measurement, and Analysis,* (Éd., S. Menard), Burlington, MA: Elsevier, 375-385.

Wiggins, L.M. (1973). *Panel Analysis: Latent Probability for Attitude and Behavior Processes*. New York: Elsevier Scientific.

Techniques d'enquête, juin 2017
Vol. 43, nº 1, pp. 139-154
Statistique Canada, nº 12-001-X au catalogue

Estimation de la variance dans le calage à plusieurs phases

Noam Cohen, Dan Ben-Hur et Luisa Burck[1]

Résumé

L'obtention d'estimateurs dans un processus de calage à plusieurs phases requiert le calcul séquentiel des estimateurs et des poids calés des phases antérieures afin d'obtenir ceux de phases ultérieures. Déjà après deux phases de calage, les estimateurs et leurs variances comprennent des facteurs de calage provenant des deux phases, et les formules deviennent lourdes et non informatives. Par conséquent, les études publiées jusqu'à présent traitent principalement du calage à deux phases, tandis que le calage à trois phases ou plus est rarement envisagé. Dans certains cas, l'analyse s'applique à un plan de sondage particulier et aucune méthodologie complète n'est élaborée pour la construction d'estimateurs calés ni, tâche plus difficile, pour l'estimation de leur variance en trois phases ou plus. Nous fournissons une expression explicite pour calculer la variance d'estimateurs calés en plusieurs phases qui tient pour n'importe quel nombre de phases. En spécifiant une nouvelle représentation des poids calés en plusieurs phases, il est possible de construire des estimateurs calés qui ont la forme d'estimateurs par la régression multivariée, ce qui permet de calculer un estimateur convergent de leur variance. Ce nouvel estimateur de variance est non seulement général pour tout nombre de phases, mais possède aussi certaines caractéristiques favorables. Nous présentons une comparaison à d'autres estimateurs dans le cas particulier du calage à deux phases, ainsi qu'une étude indépendante pour le cas à trois phases.

Mots-clés : Calage; échantillonnage à plusieurs phases; régression généralisée.

1 Introduction

La statistique des sondages fait appel à l'information auxiliaire disponible sur les totaux de population connus pour améliorer les estimations. Un estimateur par calage utilise des poids calés qui, selon une mesure de distance donnée, sont aussi proches que possible des poids de sondage initiaux, tout en satisfaisant un ensemble de contraintes induites par l'information auxiliaire. Des plans d'échantillonnage arbitraires sont permis à toutes les phases de l'échantillonnage, et l'information auxiliaire peut être utilisée à toute phase et est intégrée dans le processus d'estimation.

L'échantillonnage à plusieurs phases assorti du calage sur des données auxiliaires connues est une technique puissante et rentable. Le processus de calage a été étudié abondamment et, parmi les plans à plusieurs phases, le cas particulier de l'échantillonnage à deux phases est une exception qui a fait l'objet de recherches minutieuses. Rao (1973) et Cochran (1977, chapitre 12) ont donné les résultats fondamentaux pour la stratification et la non-réponse sous échantillonnage à deux phases. Un cadre détaillé de l'approche de pondération linéaire sous échantillonnage à deux phases est présenté dans Särndal, Swensson et Wretman (1992, chapitre 9). D'autres procédures d'estimation ont été étudiées pour des plans d'échantillonnage importants, dont le cas où l'échantillon de deuxième phase a été restratifié en utilisant l'information recueillie auprès de l'échantillon de première phase (Binder, Babyak, Brodeur, Hidiroglou et Jocelyn 2000). L'estimation de la variance a été le sujet principal de travaux de recherche dynamiques faisant appel à différentes approches, telles la méthode de linéarisation présentée dans Binder (1996), l'utilisation du jackknife (Kott et Stukel 1997) ou d'autres procédures de rééchantillonnage (Rao et Shao 1992; Fuller 1998; Kim, Navarro et Fuller 2006). Davantage en rapport avec nos travaux, Breidt et Fuller (1993) ont donné des

1. Noam Cohen, Dan Ben-Hur et Luisa Burck, Statistical Methodology Department, The Central Bureau of Statistics, 95464 Jérusalem, Israël. Courriel : avinoam.cohen@mail.huji.ac.il.

procédures d'estimation efficaces pour l'échantillonnage à trois phases en présence d'information auxiliaire, et Hidiroglou et Särndal (1998) ont étudié l'utilisation d'information auxiliaire pour l'échantillonnage à deux phases tout en permettant une légère modification de la fonction de distance qui aboutit à des facteurs de calage additifs (également appelés *facteurs* g) plutôt que multiplicatifs. Une caractéristique commune de ces résultats est la représentation des poids calés de dernière phase au moyen des poids calés des phases antérieures. Il s'agit d'un inconvénient important, car cela requiert le calcul des poids de toutes les phases antérieures pour obtenir ceux des dernières phases, ce qui rend difficile la présentation d'une méthodologie bien établie montrant comment estimer la variance des estimateurs calés sous des plans comptant plus de deux phases.

Afin de résoudre ce problème, nous utilisons la modification de la fonction de distance des moindres carrés généralisée (MCG), introduite par Hidiroglou et Särndal (1998), pour obtenir une représentation du vecteur des poids calés en plusieurs phases qui ne contient que des poids exprimés au moyen des poids de sondage initiaux et n'inclut pas les facteurs g. Partant de cette représentation, nous pouvons construire des estimateurs calés en plusieurs phases possédant la forme d'estimateurs par la régression multivariée, ce qui à son tour permet d'établir une formule générale pour un estimateur convergent de la variance des estimateurs calés en plusieurs phases qui est vérifiée pour tout nombre de phases de calage. Dans le cas relativement simple du calage à deux phases, pour lequel une autre formule d'un estimateur de variance existe dans la littérature, une comparaison montre que les deux estimateurs diffèrent fondamentalement en forme et en interprétation. Il importe de souligner que, dans ce cas particulier, le nouvel estimateur de variance proposé n'apparaît pas supérieur (ni inférieur) en ce qui concerne le biais ou la variance, mais qu'il manifeste certaines autres caractéristiques favorables qui seront discutées à la section 3.2. Cependant, l'objectif principal de l'article n'est pas de prouver la supériorité dans le cas à deux phases, mais de présenter l'approche de rechange sous laquelle la nouvelle représentation des poids calés peut produire une formule explicite pour un estimateur de la variance des estimateurs calés en plusieurs phases qui est vérifiée pour tout nombre de phases.

La présentation de l'article est la suivante. À la section 2, nous donnons la notation, qui est très semblable à celle utilisée par Hidiroglou et Särndal (1998). À la section 3, nous exposons la méthodologie et présentons plus en détail, à la sous-section 3.2, les cas particuliers du calage à deux et à trois phases. À la section 4, nous présentons une étude en simulation pour illustrer certaines caractéristiques de la nouvelle approche. Enfin, nos conclusions sont présentées à la section 5 avec des propositions de domaines à explorer dans des études ultérieures.

2 Notation

La notation que nous utilisons est similaire à celle donnée dans Särndal et coll. (1992) et dans Hidiroglou et Särndal (1998). Considérons une population finie $U = \{1, \ldots, k, \ldots, N\}$. Un échantillon probabiliste de première phase $s_1 \left(s_1 \subseteq U\right)$ est tiré de la population U en utilisant un plan d'échantillonnage qui génère la probabilité de sélection π_{1k} pour la k^{e} unité de la population. Sachant que s_{i-1} a été tiré, l'échantillon de la i^{e} phase $s_i \left(s_i \subseteq s_{i-1}\right)$ est sélectionné à partir de s_{i-1} selon un plan d'échantillonnage ayant les

probabilités de sélection $\pi_{ik|s_{i-1}} \equiv \Pr\left(k \in s_i \,\middle|\, k \in s_{i-1}\right)$. Soulignons la nature conditionnelle des probabilités de sélection de la phase résultante. À partir de ce point, nous travaillons uniquement avec les poids dans le processus d'estimation. Le poids d'échantillonnage de l'unité $k \in s_i$ à la i^{e} phase conditionnée et son poids d'échantillonnage global seront désignés par $w_{ik} = 1/\pi_{ik|s_{i-1}}$ et $w_{ik}^* = \prod_{j=1}^{i} w_{jk}$, respectivement.

Soit y_k la valeur de la variable cible pour la k^{e} unité de la population à laquelle un vecteur auxiliaire $\mathbf{x}_k = \left(x_{1k},\ldots,x_{jk},\ldots,x_{Jk}\right)$ est associé. Désignons par y le vecteur d'éléments de la variable cible obtenu à la dernière phase d'échantillonnage, p. Comme il est décrit dans Särndal et coll. (1992, chapitre 9), nous partitionnons le vecteur $\mathbf{x}$ comme $\mathbf{x} = \left(\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_p'\right)'$ avec $p \le J$, de sorte que nous pourrions obtenir plus d'une variable auxiliaire à certaines phases. Le total de population de $\mathbf{x}$, $t_{\mathbf{x}} = \sum_U \mathbf{x}_k$ est supposé inconnu. Cependant, certains totaux démographiques peuvent être connus en s'appuyant sur des sources relativement exactes, comme les données de recensement ou d'autres types de fichiers administratifs. Sans perte de généralité, désignons par $\mathbf{x}_1$ le vecteur des variables connues pour toutes les unités dans la population U. Désignons par $\mathbf{x}_2$ le vecteur des variables obtenues dans l'échantillon de première phase s_1, et ainsi de suite. Pour les éléments contenus dans s_r, $r \le p$, l'information complète est alors résumée dans le vecteur $\mathbf{x} = \left(\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_r'\right)'$. Écrivons aussi $t_i = t_{\mathbf{x}_i}$.

Soit X_r la matrice de plan comprenant n_r lignes représentant n_r unités échantillonnées, et un nombre de colonnes correspondant au nombre de variables auxiliaires dans le vecteur $\mathbf{x}_r$. Notons que X_r est obtenue dans l'échantillon s_{r-1} à la $r-1^{\text{e}}$ phase de l'échantillonnage, si bien que nous pouvons concevoir U comme un échantillon s_0. Dans les conditions qui figurent par exemple dans Särndal et coll. (1992) et dans Hidiroglou et Särndal (1998), la matrice de plan X_r englobe toutes les variables auxiliaires $\mathbf{x}_1,\ldots,\mathbf{x}_r$, plutôt que simplement $\mathbf{x}_r$, et est appelée le *vecteur complet*. Néanmoins, l'analyse est la même dans les deux cas.

L'information auxiliaire disponible à chaque phase de l'échantillonnage peut être utilisée pour obtenir des poids améliorés grâce au processus de calage qui produit des facteurs de calage à utiliser dans le processus d'estimation. Nous utilisons l'indice supérieur « * » pour désigner les poids globaux, c'est-à-dire les poids tenant compte de toutes les phases. Le symbole superposé « $\sim$ » désigne les poids calés. Les facteurs g de la i^{e} phase sont désignés par g_{ik}, ce qui donne les poids calés de la i^{e} phase $\tilde{w}_{ik} = \tilde{w}_{i-1,k} w_{ik} g_{ik}$ pour $k \in s_i$, où les $\tilde{w}_{i-1,k}$ sont les poids calés de la $i-1^{\text{e}}$ phase et $\tilde{w}_{0k} = 1$. Pour $k \in s_i$ le calage par rapport à toutes les phases produit des facteurs de calage globaux désignés par g_{ik}^*. Par conséquent, nous aurons les poids calés globaux $\tilde{w}_{ik} = w_{ik}^* g_{ik}^*$, où w_{ik}^* est le poids d'échantillonnage global. Désignons par w_i le vecteur dont les composantes sont $w_{ik}; k = 1,\ldots,n_i$, et par D_i une matrice diagonale de taille n_i avec w_i sur sa diagonale. La même notation sera utilisée avec les vecteurs $w_i^*, \tilde{w}_i$ et g_i.

3 Calage avec la distance MCG

Le calage requiert la spécification d'une fonction de distance mesurant la distance entre les poids initiaux et les nouveaux poids calés. Plusieurs fonctions de distance ont été étudiées, certaines étant résumées dans

Deville et Särndal (1992). Nous nous concentrons sur la mesure de distance par les moindres carrés généralisée (MCG). La forme classique du calage à plusieurs phases sous la fonction de distance MCG consiste à trouver les valeurs $\tilde{w}_{ik}$ pour l'ensemble $k \in s_i$ qui minimisent l'expression

$$\sum_{k \in s_i} \frac{c_{ik}\left(\tilde{w}_{ik} - \tilde{w}_{i-1,k} w_{ik}\right)^2}{\tilde{w}_{i-1,k} w_{ik}} \tag{3.1}$$

sous la contrainte

$$\sum_{k \in s_i} \tilde{w}_{ik} x_{ik} = \sum_{k \in s_{i-1}} \tilde{w}_{i-1,k} x_{ik} \tag{3.2}$$

(autrement, on peut écrire $\tilde{w}_{i-1,k} w_{ik} g_{ik}$ au lieu de $\tilde{w}_{ik}$) où les $\left\{\tilde{w}_{i-1,k} : k \in s_i\right\}$ sont les poids initiaux au début de la phase i, c'est-à-dire les poids calés obtenus à la phase $i-1$; les $\left\{\tilde{w}_{ik} : k \in s_i\right\}$ sont les poids calés de la phase i que nous voulons obtenir; et les $\left\{c_{ik} : k \in s_i\right\}$ sont les facteurs positifs spécifiés utilisés pour contrôler l'importance relative que nous voulons attribuer à chacun des éléments de la somme en fonction de l'information auxiliaire disponible pour $k \in s_{i-1}$. Pour simplifier la notation, supposons à partir de maintenant que $c_{ik} = 1$ pour tout i, k. Les poids résultant de ce scénario de calage sont $\tilde{w}_{ik} = \tilde{w}_{i-1,k} w_{ik} g_{ik}$, où $g_{ik} = 1 + \left(\sum_{l \in s_{i-1}} \tilde{w}_{i-1,l} x_{il} - \sum_{l \in s_i} \tilde{w}_{i-1,l} w_{il} x_{il}\right)' T_i^{-1} x_{ik}$ avec $T_i = \sum_{l \in s_i} w_{il}^* g_{i-1,l}^* x_{il} x_{il}'$. D'où, les facteurs de calage dans ce processus agissent multiplicativement pour donner un facteur de calage global $g_{ik}^* = \prod_{j=1}^{i} g_{jk}$ pour $k \in s_i$ à la fin de la phase i.

La mesure de distance (3.1) peut être critiquée, parce que les facteurs $1/\tilde{w}_{i-1,k} w_{ik}$ pour une phase i pourraient ne pas être forcément tous finis et positifs, car les termes $g_{i-1,k}$ qui figurent dans $\tilde{w}_{i-1,k}$ au dénominateur peuvent être nuls ou négatifs, ce qui contredit la notion de distance. Un autre choix de fonction de distance, et celui que nous utiliserons dans notre analyse, consiste à remplacer (3.1) par

$$\sum_{k \in s_i} \frac{\left(\tilde{w}_{ik} - \tilde{w}_{i-1,k} w_{ik}\right)^2}{w_{i-1,k}^* w_{ik}} \tag{3.3}$$

c'est-à-dire par des poids non calés au dénominateur. Il est facile de vérifier que les poids calés globaux résultant de la minimisation de (3.3) sous la contrainte (3.2) sont (pour $p = 2$, voir Hidiroglou et Särndal 1998)

$$\tilde{w}_{pk} = w_{pk}^* \left(g_{1k} + \ldots + g_{ik} + \ldots + g_{pk} - (p-1)\right) \tag{3.4}$$

où

$$g_{ik} = 1 + \left(\sum_{l \in s_{i-1}} \tilde{w}_{i-1,l} x_{il} - \sum_{l \in s_i} \tilde{w}_{i-1,l} w_{il} x_{il}\right)' T_i^{-1} x_{ik} \tag{3.5}$$

pour $k \in s_p$ avec $T_i = \sum_{l \in s_i} w_{il}^* x_{il} x_{il}'$. Le choix d'une mesure de distance dans la construction des estimateurs calés n'est pas critique, puisque les estimateurs résultants pour une large gamme de mesures de distance sont asymptotiquement équivalents à celui qui utilise la mesure de distance MCG (3.1), Deville et

Särndal (1992). Il en est de même de la mesure de distance (3.3). Puisque l'estimateur de Horvitz-Thompson $X_1'w_1^*$ est sans biais pour t_1 avec un écart-type d'ordre de grandeur $N \cdot O\left(n_1^{-1/2}\right)$, alors $g_{1k} = 1 + O\left(n_1^{-1/2}\right)$ pour tout $k \in s_1$ et donc $\tilde{w}_{1k} = w_{1k}^*\left(1 + O\left(n_1^{-1/2}\right)\right)$. Par induction, $g_{ik} = 1 + O\left(n_i^{-1/2}\right)$ pour tout i et découlant de (3.4), $\tilde{w}_{pk} \big/ w_{pk}^* \to 1$ en probabilité avec n_p. Suggérant de nouvelles techniques en vue d'améliorer l'estimation, Farrell et Singh (2002) ont proposé d'autres types de fonction de distance du khi carré pénalisée.

3.1 Estimation

L'analyse qui suit est motivée par la nature récursive de $\tilde{w}_{ik}$ dans (3.4), où les poids calés des phases antérieures $1,\ldots,i-1$ sont emboîtés dans chaque facteur g_{ik}, ce qui requiert le calcul séquentiel des poids calés; autrement dit, il faut calculer tous les poids calés des phases antérieures pour obtenir ceux des phases ultérieures. Soient $\hat{B}_{ij}^{+} = \left(\sum_{k\in s_i} w_{ik}^* x_{ik} x_{ik}'\right)^{-1} \sum_{k\in s_j} w_{jk}^* x_{ik} x_{jk}'$ et $\hat{B}_{ij}^{-} = \left(\sum_{k\in s_i} w_{ik}^* x_{ik} x_{ik}'\right)^{-1} \sum_{k\in s_{j-1}} w_{j-1,k}^* x_{ik} x_{jk}'$ les estimateurs de $B_{ij} = \left(\sum_{k\in U} x_{ik} x_{ik}'\right)^{-1} \sum_{k\in U} x_{ik} x_{jk}'$, le coefficient de régression de $\mathbf{x}_j$ sur $\mathbf{x}_i$. La différence entre les deux estimateurs tient au fait que, tandis que $\hat{B}_{ij}^{-}$ utilise l'ensemble complet d'unités connues pour $\mathbf{x}_j$ qui est obtenu dans s_{j-1}, $\hat{B}_{ij}^{+}$ utilise uniquement le sous-ensemble $s_j \subseteq s_{j-1}$ et, donc, plus de variables que $\hat{B}_{ij}^{-}$. Soit $\hat{Z}_{ij} = \hat{B}_{ij}^{+} - \hat{B}_{ij}^{-}$ la différence entre les deux coefficients estimés qui converge vers zéro. Notons aussi $\hat{Z}_{i_1 i_2 \ldots i_k} = \prod_{j=2}^{k} \hat{Z}_{i_{j-1} i_j}$ pour $k \geq 2$ et $\hat{Z}_{i_1} = 1$ pour $k = 1$. Soit $\hat{t}_i^{-} = \sum_{k\in s_{i-1}} w_{i-1k}^* x_{ik}$ et $\hat{t}_i^{+} = \sum_{k\in s_i} w_{ik}^* x_{ik}$ les deux estimateurs de Horvitz-Thompson pour t_i, fondés sur les unités obtenues dans les échantillons s_i et s_{i-1}, respectivement. Notons que tous les estimateurs définis dans le présent paragraphe utilisent les poids de sondage globaux w^* et non les poids calés. Dans le lemme qui suit, nous donnons une représentation de $\tilde{w}_p$, le vecteur de poids calés après p phases de calage, qui dépend uniquement des poids de sondage connus au préalable $\{w_i^*\}_{i=1}^{p}$.

Lemme 3.1 *Considérons un plan d'échantillonnage à plusieurs phases avec un scénario de calage qui produit des facteurs g additifs comme il est défini dans (3.3). Une représentation des poids calés à la phase p fondée entièrement sur les poids de sondage est*

$$\begin{aligned} \tilde{w}_p &= D_p^{*\prime} 1_{n_p} + \sum_{i_1=1}^{p} A_{i_1} - \sum_{i_1<i_2}^{p} A_{i_1 i_2} \\ &\quad + \ldots + (-1)^{k+1} \sum_{i_1<\ldots<i_k}^{p} A_{i_1 i_2 \ldots i_k} + \ldots + (-1)^{p+1} A_{i_1 i_2 \ldots i_p} \end{aligned} \tag{3.6}$$

où $A_{i_1 i_2 \ldots i_k}' = \left(\hat{t}_{i_1}^{-} - \hat{t}_{i_1}^{+}\right)' \hat{Z}_{i_1 i_2 \ldots i_k} \left(X_{i_k}' D_{i_k}^* X_{i_k}\right)^{-1} X_{i_k}' D_p^*$.

Preuve. Voir l'annexe A.

Notons la forme « inclusion-exclusion » de $\tilde{w}_p$ dans le lemme 3.1. La k'^{e} sommation comprend $\binom{p}{k}$ opérandes $A_{i_1 i_2 \ldots i_k}$, pour lesquels chaque $\hat{Z}_{i_1 i_2 \ldots i_k} = \prod_{j=2}^{k}\left(\hat{B}_{i_{j-1} i_j}^{+} - \hat{B}_{i_{j-1} i_j}^{-}\right)$ contient 2^k opérandes. Soit, un total

de $\binom{p}{k}2^k$ opérandes. Le nombre global de termes dans (3.6) est par conséquent 3^p comme il est montré dans la preuve du lemme. Notons aussi que les termes $A_{i_1 i_2 \ldots i_k}$ comprennent le produit des composantes $\hat{t}_{i_1}^- - \hat{t}_{i_1}^+$ et $\hat{Z}_{i_1 i_2 \ldots i_k}$, ayant toutes deux une espérance nulle, de sorte que le poids calé $\tilde{w}_p$ est égal à $D_p^{*\prime} 1_{n_p}$, le poids de sondage global, plus les termes de correction d'ordres de grandeur plus faibles, et maintient la caractéristique bien connue des poids calés. Jusqu'à présent, nous nous sommes limités dans notre discussion à une représentation du vecteur des poids dans un processus de calage à plusieurs phases qui fait intervenir uniquement des paramètres du plan de sondage et n'inclut pas les facteurs g. Or, partant de cette représentation de $\tilde{w}_p$, il est possible de déduire un estimateur novateur pour la variance des estimateurs calés en plusieurs phases. Soit y une variable d'intérêt pour laquelle nous voulons estimer le total de population Y. Soit $\hat{\beta}_j = \left(\sum_{k \in s_j} w_{jk}^* x_{jk} x_{jk}'\right)^{-1} \sum_{k \in s_p} w_{pk}^* x_{jk} y_k$, le coefficient de régression de y sur $\mathbf{x}_j$, et $\hat{Y}_{\mathrm{HT}_p} = 1_{n_p}' D_p^* y$, l'estimateur de Horvitz-Thompson non calé, calculé sur les éléments compris dans s_p. Le réarrangement des termes dans (3.6) produit une représentation plus classique de l'estimateur calé en plusieurs phases $\tilde{w}_p' y$ sous forme d'un estimateur par la régression multivariée

$$\tilde{w}_p' y = \hat{Y}_{\mathrm{HT}_p} + \sum_{i_1=1}^{p} \left(\hat{t}_{i_1}^- - \hat{t}_{i_1}^+\right)' \hat{\gamma}_{i_1} \tag{3.7}$$

où

$$\hat{\gamma}_{i_1} = \hat{\beta}_{i_1} - \sum_{i_1 < i_2}^{p} \hat{Z}_{i_1 i_2} \hat{\beta}_{i_2} +$$

$$\ldots + (-1)^{k+1} \sum_{i_1 < \ldots < i_k}^{p} \hat{Z}_{i_1 i_2 \ldots i_k} \hat{\beta}_{i_k} + \ldots + (-1)^{p-(i_1-1)+1} \hat{Z}_{i_1 \ldots p} \hat{\beta}_p.$$

L'établissement d'un estimateur convergent de la variance des estimateurs calés en plusieurs phases est maintenant simple en ce sens qu'il suit à peu près les étapes utilisées dans le calcul de la variance sous un scénario de calage multivarié à une phase.

Théorème 3.1 *Soit* $\hat{e}_{rk} = x_{rk}' \hat{\gamma}_r - x_{r+1,k}' \hat{\gamma}_{r+1}$ *pour* $r < p$ *et* $\hat{e}_{pk} = x_{pk}' \hat{\gamma}_p - y_k$. *Un estimateur convergent de la variance de* $\tilde{w}_p' y$ *est*

$$\sum_{1 \le r_1, r_2 \le p} \sum_{k \in s_{r_m},\, l \in s_{r_M}} \frac{w_{r_M l}^*}{w_{r_m l}^*} \left(w_{r_m k}^* w_{r_m l}^* - w_{r_m k l}^*\right) \hat{e}_{r_m k} \hat{e}_{r_M l} \tag{3.8}$$

où $r_m = \min(r_1, r_2)$ *et* $r_M = \max(r_1, r_2)$.

La preuve comprend l'évaluation des ordres de grandeur les plus élevés et l'estimation de leur variance. Une attention particulière est accordée à l'évaluation de la probabilité conjointe des événements $\{k \in s_i, l \in s_j\}$ et à l'estimation de la covariance entre les unités provenant de différentes phases d'échantillonnage.

Preuve. À la première étape, nous allons voir que le remplacement des estimateurs des coefficients $\hat{\gamma}_i; i = 1 \ldots p$ par leurs valeurs réelles γ_i affecte l'estimation de la variance d'un facteur $N^2 o\left(n_p^{-1}\right)$ et, donc, n'affecte pas la convergence de l'estimateur substitué. À cette fin, notons que $\hat{B}_{ij}^+, \hat{B}_{ij}^-$ sont tous deux convergents vers B_{ij}. Écrivons $\hat{B}_{ij}^+ = B_{ij} + \left(\hat{B}_{ij}^+ - B_{ij}\right)$ de sorte que $\hat{B}_{ij}^+ = B_{ij} + O_p\left(n_j^{-1/2}\right)$. Rappelons que $\hat{Z}_{ij} = \hat{B}_{ij}^+ - \hat{B}_{ij}^-$, où $\hat{B}_{ij}^-$ est basé sur s_{j-1}, tandis que $\hat{B}_{ij}^+$ est basé sur son sous-échantillon s_j et, donc, $\hat{Z}_{ij} = O_p\left(n_j^{-1/2}\right)$ et, par conséquent, $\hat{Z}_{i_1 i_2 \ldots i_k}$ est borné par $O_p\left(n_{i_k}^{-1/2}\right)$. De même, $\hat{\beta}_j$ est $\beta_j + O_p\left(n_p^{-1/2}\right)$, parce que y est observé uniquement à la dernière phase d'échantillonnage s_p. Donc, $\hat{\gamma}_i$ est convergent vers γ_i pour tout i, où les $\hat{\beta}_i$ dans $\hat{\gamma}_i$ sont remplacés par β_i dans γ_i. La convergence n'implique pas nécessairement la convergence des moments et, en particulier, pas de la variance. Cependant, pour une population finie, c'est-à-dire un espace de probabilité fini, les concepts coïncident. Il s'ensuit que, pour n_p suffisamment grand, $\mathrm{Var}\left(\hat{Y}_{\mathrm{HT}_p} + \sum_{i_1=1}^{p}\left(\hat{t}_{i_1}^- - \hat{t}_{i_1}^+\right)' \hat{\gamma}_{i_1}\right)$ et $\mathrm{Var}\left(\hat{Y}_{\mathrm{HT}_p} + \sum_{i_1=1}^{p}\left(\hat{t}_{i_1}^- - \hat{t}_{i_1}^+\right)' \gamma_{i_1}\right)$ sont asymptotiquement équivalents et selon la discussion qui précède, la différence peut être quantifiée par

$$\mathrm{Var}\left(\tilde{w}_p' y\right) = \mathrm{Var}\left(\hat{Y}_{\mathrm{HT}_p} + \sum_{r=1}^{p}\left(\hat{t}_r^- - \hat{t}_r^+\right)' \gamma_r\right) + N^2 o\left(n_p^{-1}\right).$$

L'estimateur $\hat{t}_r^+$ est une sommation sur les unités comprises dans s_r, tandis que $\hat{t}_r^-$ est une sommation sur s_{r-1}. En réarrangeant les termes, la variance dans le deuxième membre de l'équation peut s'écrire $\mathrm{Var}\left(\sum_{r=1}^{p}\sum_{i \in S_r} w_{ri}^* e_{ri}\right)$, ce qui est égal à

$$\sum_{1 \le r_1, r_2 \le p} \sum_{k \in U} \sum_{l \in U} w_{r_1 k}^* e_{r_1 k} w_{r_2 l}^* e_{r_2 l} \mathrm{Cov}\left(I_{k \in s_{r_1}}, I_{l \in s_{r_2}}\right)$$

de sorte qu'un estimateur basé sur l'échantillon serait

$$\sum_{1 \le r_1, r_2 \le p} \sum_{k \in s_{r_1},\, l \in s_{r_2}} w_{r_1 k}^* \hat{e}_{r_1 k} w_{r_2 l}^* \hat{e}_{r_2 l} \left[1 - \frac{P\left(k \in s_{r_1}\right) P\left(l \in s_{r_2}\right)}{P\left(k \in s_{r_1}, l \in s_{r_2}\right)}\right]. \tag{3.9}$$

Pour calculer la covariance entre les indicateurs $I_{k \in s_{r_1}}$ et $I_{l \in s_{r_2}}$, nous devons connaître la probabilité conjointe des événements $\{k \in s_i, l \in s_j\}$. Si $s_j \subset s_i$, alors $P\left(k \in s_i, l \in s_j\right)$ est égale à la probabilité conjointe que les deux unités k, l soient dans l'échantillon $s_i = s_{\min(i,j)}$, multipliée par la probabilité conditionnelle que l'unité l soit dans l'échantillon s_j, sachant qu'il appartient à s_i. Formellement, si $s_j \subset s_i$, alors $P\left(k \in s_i, l \in s_j\right) = \frac{w_{jl}^*}{w_{jl}^*} w_{i,lk}^{*-1}$, ce qui élimine la dépendance à l'égard de s_{r_2} entre les crochets dans (3.9) et le résultat s'ensuit.

Un autre moyen d'écrire (3.8) est

$$\sum_{1 \le r \le p} \sum_{k,l \in s_r} \left(w_{rk}^* w_{rl}^* - w_{rkl}^*\right) \hat{e}_{rk} \hat{e}_{rl} + 2 \sum_{1 \le r_m < r_M \le p} \sum_{k \in s_{r_m}} \sum_{l \in s_{r_M}} w_{r_m k}^* \hat{e}_{r_m k} w_{r_M l}^* \hat{e}_{r_M l} \left(1 - \frac{w_{r_m kl}^*}{w_{r_m k}^* w_{r_m l}^*}\right).$$

Quand $p = 2,$ les termes γ_i coïncident avec les unités de variation obtenues de la décomposition de l'erreur d'échantillonnage de l'*estimateur en deux étapes* de Breidt et Fuller (1993). Des estimations convergentes pour les écarts-types des estimations calées des sous-totaux de population sont calculées de façon ordinaire en multipliant la variable cible par une variable indicatrice pour la sous-population particulière.

Jusqu'à présent dans notre discussion, nous avons donné une représentation du vecteur de poids calés de laquelle nous avons dérivé un nouvel estimateur convergent pour la variance des estimateurs calés en plusieurs phases. Cependant, dans certains cas, les estimateurs peuvent être simplifiés davantage sans perte d'exactitude. Nous discuterons brièvement ici de deux scénarios qui dépendent du fait que n_j est ou non significativement plus petit que $n_{j-1},$ c'est-à-dire du fait que, pour tout $j,$ le sous-échantillon s_j est ou non significativement plus petit que $s_{j-1}.$ Un cas type du premier scénario est celui où l'on possède un ensemble de fichiers administratifs emboîtés dont les tailles diminuent significativement. Le premier ensemble peut être, par exemple, un fichier de registre de population qui contient un nombre limité de variables au sujet de l'ensemble de la population, comme l'âge, le sexe, etc. Le deuxième ensemble peut correspondre à des données d'échantillons provenant d'une enquête de portée nationale dans le cadre de laquelle des données complètes sur les ménages ont été recueillies auprès de toutes les unités échantillonnées, mais en utilisant un questionnaire supplémentaire pour un sous-groupe de ces unités (disons, une unité sur dix). Les données pour ce sous-groupe d'unités peuvent alors être calées sur celles provenant des deux sources d'information précédentes. Un exemple du second scénario est la situation où une ou deux phases de calage sont effectuées sur le même ensemble de données. Autrement dit, contrairement au processus à plusieurs phases habituel, l'élément d'échantillonnage est présent à la première phase seulement, mais non aux phases ultérieures. Un tel scénario peut avoir lieu si nous voulons caler les données d'une enquête sur de nombreuses variables pour lesquelles nous connaissons seulement les totaux de marge, mais ne possédons pas les totaux transversaux. Dans ces conditions, une série de calages sur le même échantillon, mais en utilisant un ensemble différent de variables auxiliaires à chaque phase, en attribuant habituellement aux dernières phases les variables les plus importantes, pourrait être un compromis satisfaisant. Une meilleure façon de caractériser ce scénario serait de le dire *séquentiel*. Sous ces scénarios, $\tilde{w}_p$ et sa variance peuvent être simplifiés considérablement. Ces scénarios peuvent être énoncés comme des corollaires de notre analyse, mais nous choisissons de ne pas les prendre en considération ici afin de nous concentrer sur nos résultats courants.

3.2 Exemples : Calage à deux phases et à trois phases

Calage à deux phases. Nous utiliserons le cas particulier du calage à deux phases $(p = 2)$ pour démontrer la nouvelle méthodologie et ce qui la distingue de l'autre estimateur habituellement utilisé dans la littérature. En notation matricielle, l'estimateur calé est donné, selon (3.7), par

$$\tilde{w}_2' y = \hat{Y}_{\mathrm{HT}_2} + \left(\hat{t}_1^- - \hat{t}_1^+\right)' \hat{\gamma}_1 + \left(\hat{t}_2^- - \hat{t}_2^+\right)' \hat{\gamma}_2$$

où $\hat{\gamma}_1 = \hat{\beta}_1 - \hat{Z}_{12}\hat{\beta}_2$ et $\hat{\gamma}_2 = \hat{\beta}_2.$ Explicitement, sous forme non matricielle,

$$\tilde{w}_2' y = \sum_{k \in s_2} w_{2k}^* y_k + \left(\sum_{k \in U} x_{1k} - \sum_{k \in s_1} w_{1k} x_{1k} \right) \hat{\gamma}_1 + \left(\sum_{k \in s_1} w_{1k} x_{2k} - \sum_{k \in s_2} w_{2k}^* x_{2k} \right) \hat{\gamma}_2$$

où

$$\hat{\gamma}_1 = \left(\sum_{k \in s_1} w_{1k} x_{1k} x_{1k}' \right)^{-1} \left[\sum_{k \in s_2} w_{2k}^* x_{1k} y_k - \left(\sum_{k \in s_2} w_{2k}^* x_{1k} x_{2k}' - \sum_{k \in s_1} w_{1k} x_{1k} x_{2k}' \right) \hat{\gamma}_2 \right]$$

$$\hat{\gamma}_2 = \left(\sum_{k \in s_2} w_{2k}^* x_{2k} x_{2k}' \right)^{-1} \sum_{k \in s_2} w_{2k}^* x_{2k} y_k .$$

Cet estimateur produit des estimations identiques à l'estimateur calé en deux phases utilisé dans Hidiroglou et Särndal (1998) ou dans Särndal et coll. (1992), section 9.7. Cependant, une fois que l'estimateur des paramètres γ_1, γ_2 est calculé, la représentation de $\tilde{w}_2' y$ devient simple et informative, car elle possède la structure d'un simple estimateur par la régression multivariée. Cet estimateur linéaire est fondé sur les coefficients γ qui englobent l'effet total de la variable $\mathbf{x}$ qu'ils multiplient et, donc, diffèrent légèrement des coefficients β. $\hat{\gamma}_i$ englobe l'effet global que le calage sur la variable $\mathbf{x}_i$ a sur l'estimation de Y. Dans le cas général, il tient compte de la projection de y sur $\mathbf{x}_i$, de la projection de y sur $\mathbf{x}_{i+1}$ multipliée par la projection de $\mathbf{x}_{i+1}$ sur $\mathbf{x}_i$, et ainsi de suite. En outre, comme nous allons le montrer, les estimateurs de variance diffèrent significativement en ce qui concerne tant les estimations que la représentation. Étant donné la complexité de l'évaluation de la variance des estimateurs qui comprennent des facteurs g, jusqu'à présent dans la littérature sur le calage à deux phases, il était d'usage pratique de commencer par donner aux facteurs g la valeur approximative de 1, puis d'utiliser la loi de la variation totale pour obtenir deux composantes, une pour chaque phase, conformément à

$$\hat{V}_C \left(\tilde{w}_2' y \right) = \sum_{k,\, l \in s_2} w_{2kl} \left(w_{1k} w_{1l} - w_{1kl} \right) \left(g_{1k} \breve{e}_{1k} \right) \left(g_{1l} \breve{e}_{1l} \right)$$

$$+ \sum_{k,\, l \in s_2} w_{1k} w_{1l} \left(w_{2k} w_{2l} - w_{2kl} \right) \left(g_{2k} \breve{e}_{2k} \right) \left(g_{2l} \breve{e}_{2l} \right) \qquad (3.10)$$

où les termes d'erreur $\breve{e}_{1k} = y_k - x_{1k}' \hat{\gamma}_1$ et $\breve{e}_{2k} = y_k - x_{2k}' \hat{\gamma}_2$ sont tous deux définis pour $k \in s_2$, parce que y est observé uniquement sur s_2, et on notera la représentation simple des termes d'erreur sous la notation faisant appel aux coefficients γ. Les facteurs g sont définis comme dans (3.5). La valeur approximative de 1 donnée aux facteurs g dans le calcul de (3.10) peut indubitablement aboutir à des estimations imprévisibles, car ces facteurs s'écartent de l'unité précisément dans les situations où le calage est essentiel. Par ailleurs, l'estimateur de variance proposé en (3.8) pour un estimateur calé en deux phases est donné par

$$\hat{V}_P \left(\tilde{w}_2' y \right) = \sum_{k,\, l \in s_1} \left(w_{1k} w_{1l} - w_{1kl} \right) \hat{e}_{1k} \hat{e}_{1l} + \sum_{k,\, l \in s_2} \left(w_{2k}^* w_{2l}^* - w_{2kl}^* \right) \hat{e}_{2k} \hat{e}_{2l}$$

$$+ 2 \sum_{k \in s_1,\, l \in s_2} \frac{w_{2l}^*}{w_{1l}} \left(w_{1k} w_{1l} - w_{1kl} \right) \hat{e}_{1k} \hat{e}_{2l} . \qquad (3.11)$$

La différence entre les estimateurs de variance issus des deux méthodes représentées par les équations (3.10) et (3.11) est fondamentale. Elle se manifeste sous divers aspects. Tandis que le terme d'erreur de la deuxième phase est le même dans les deux méthodes, c'est-à-dire $\hat{e}_{2k} = \breve{e}_{2k}$, le terme d'erreur de la première phase diffère. $\breve{e}_{1k}$ est fondé sur la différence entre y_k et le prédicteur de régression $x'_{1k}\hat{\gamma}_1$, tandis que $\hat{e}_{1k}$ est basé sur la différence entre deux prédicteurs de Y provenant des phases un et deux $x'_{1k}\hat{\gamma}_1 - x'_{2k}\hat{\gamma}_2$. Cette modification fait que le premier opérande dans (3.11) est calculé sur s_1 et non sur s_2 où l'échantillon est plus grand. Comme on le voit, l'estimateur (3.11) comprend un troisième opérande qui contient le produit des deux termes d'erreur provenant des deux phases et n'a pas de parallèle dans (3.10). Bien que ce produit soit souvent proche de zéro quand les termes d'erreur ne sont pas fortement corrélés, il peut être non négligeable quand y est fortement corrélé avec $\mathbf{x}_1$. Un avantage évident est l'absence des facteurs g qui rend l'estimateur plus simple à calculer, c'est-à-dire qu'une fois que nous avons calculé les estimations des paramètres $\hat{\gamma}_i; i = 1 \ldots p$, l'estimateur (3.11) peut être calculé en utilisant les paramètres du plan uniquement, sans impliquer les facteurs g provenant de toutes les phases du calage. Enfin, aspect peut-être le plus important du point de vue opérationnel, comme nous le montrerons aussi dans l'étude en simulation, l'avantage de (3.11) est que, pour une grande gamme de plans de sondage, le deuxième opérande représente la majorité absolue de la variance, tandis que dans (3.10), les opérandes sont habituellement du même ordre de grandeur. Cette caractéristique découle du fait que le terme $\left(w^*_{2k}w^*_{2l} - w^*_{2kl}\right)$, qui comprend les poids d'échantillonnage totaux, est très grand comparativement à $w_{2kl}\left(w_{1k}w_{1l} - w_{1kl}\right)$ ou $w_{1k}w_{1l}\left(w_{2k}w_{2l} - w_{2kl}\right)$. Dans l'estimateur de variance, la fonction $f(w) = w_k w_l - w_{kl}$ atteint son maximum sur la diagonale $k = l$, où elle est proportionnelle à w_k^2, et puis elle est multipliée par le carré de son reste $\hat{e}_k$, qui est un terme non négatif. D'où, quand le taux d'échantillonnage de la seconde phase est suffisamment élevé, il accroît fortement les termes qui dépendent des poids totaux w^*_2 de cette phase, comparativement à un terme parallèle provenant de la phase précédente. Donc, le deuxième opérande peut, pratiquement à lui seul, être un bon estimateur de la variance de l'estimateur calé.

Calage à trois phases. Le calage à plusieurs phases peut être mis en œuvre quand, dans une série d'échantillons de taille décroissante (non croissante), chaque paire de phases consécutives présente certaines variables communes. Il peut être effectué que les échantillons soient emboîtés, c'est-à-dire si s_i est un sous-échantillon de s_{i-1}, ou non. En pratique, le cas le plus simple et le plus fréquent est évidemment le calage à deux phases où un plus petit échantillon (emboîté ou non) est calé sur un échantillon beaucoup plus grand, comme celui d'une Enquête sur la population active, qui est à son tour fréquemment calé sur un fichier administratif contenant des variables démographiques. Cependant, étant donné la faisabilité des calculs et les progrès méthodologiques, les plans comportant un plus grand nombre de phases de calage demeurent répandus et les plans à trois phases occupent le second rang quant à la simplicité et à la mise en œuvre. Par conséquent, cela vaut la peine de s'étendre un peu plus sur l'estimateur pour ce cas.

L'approximation (3.8) contient six termes différents, trois pour les trois phases d'échantillonnage et trois autres pour la covariance entre les phases. Nous désignons ces termes par V_1, V_2, V_3 et C_{12}, C_{13}, C_{23}, respectivement. Chacun correspond à la multiplication d'un terme qui comprend les poids d'échantillonnage par les restes pour les phases pertinentes. Les formules pour le calage à trois phases sont présentées à l'annexe B. Comme nous l'avons exposé pour le cas à deux phases, quand $w_i > 1$, les V_i suivent

vraisemblablement un ordre clair $V_1 < V_2 < V_3$ et V_3 deviendra d'autant plus dominant que les taux d'échantillonnage de la troisième phase seront grands. Cette situation est représentée par le cas 3 dans le tableau 3.1, et dans notre simulation, cela se manifeste aux lignes 2 et 6 du tableau 4.2, où w_{3k} est égal à 10 et à 5, respectivement. Ce n'est manifestement pas très souvent le cas en réalité, car l'approximation dépend aussi des tailles des termes de reste, qui dépendent du choix des variables de calage et de leurs corrélations particulières qui sont parfois très fortes. Le cas échéant, les restes seront très petits et il serait préférable d'utiliser tous les termes de (3.8). Comme pour les termes de covariance, même si C_{13} comprend les poids globaux $\{w^*_{3k}\}$, il est peu probable qu'il ajoute une valeur importante à la variance totale en raison de la corrélation généralement faible entre les restes des phases 1 et 3. Par ailleurs, le terme C_{23}, même s'il est pondéré par les poids globaux de 2^e phase seulement, peut être significatif en raison de la forte corrélation entre les restes des phases 2 et 3, car ils contiennent tous deux le terme $x'_{3k}\hat{\gamma}_3$ pour $k \in s_3$. L'importance relative des termes pour certains plans généraux est spécifiée dans le tableau 3.1. Les coefficients γ, qui englobent l'effet total des variables $\mathbf{x}$ qu'ils multiplient, prennent maintenant une forme plus intéressante et compliquée. Par exemple, $\hat{\gamma}_1$ tient compte des projections de $\mathbf{x}_1$ sur $\mathbf{x}_2$ et de $\mathbf{x}_1$ sur $\mathbf{x}_3$, mais avec déduction de la projection de $\mathbf{x}_1$ sur la projection de $\mathbf{x}_2$ sur $\mathbf{x}_3$.

Tableau 3.1
Une représentation générale de l'importance relative de chacun des termes dans (3.8) pour certains scénarios. Les points noirs indiquent une forte dominance, les points gris foncé, une dominance modérée et les points gris clair, une non-dominance

Cas	Description	V_1	V_2	V_3	C_{12}	C_{13}	C_{23}
1	Pratiquement aucun échantillonnage supplémentaire aux deuxième et troisième phases : $w_2 \approx w_3 \approx 1$.	●	●	●	○	●	○
2	Les poids w_1, w_2, w_3 sont de taille modérée.	○	◉	●	○	●	○
3	n_3 nettement plus petit que n_2, indépendamment des tailles de w_1, w_2.	○	○	●	○	○	○

4 Une étude en simulation

L'objectif principal de l'analyse exposée dans le présent article est de fournir un estimateur convergent de la variance des estimateurs calés en plusieurs phases qui est vérifié pour tout nombre de phases de calage. Une étude en simulation pourrait donc être exécutée pour comparer le nouvel estimateur à d'autres décrits dans la littérature. Comme on ne trouve généralement aucun estimateur de rechange dans la littérature pour des plans de calage à trois phases ou plus $(p \geq 3)$, notre comparaison porte principalement sur le cas à deux phases qui est celui le plus étudié. Nous avons également exécuté une étude pour $p = 3$ afin d'évaluer l'écart de l'estimateur proposé par rapport à la valeur simulée réelle. Les études sont décrites ici en termes généraux. Elles visent essentiellement à démontrer la pertinence de l'estimateur proposé, sa concordance avec la « condition limite » du cas à deux phases, et son potentiel en ce qui concerne les plans comportant plus de deux phases. Une étude approfondie en vue de caractériser l'efficacité de l'estimateur proposé en tant que fonction des paramètres du plan, tels que les taux d'échantillonnage, le choix des variables de calage et leur corrélation avec y, etc., est réservée à de futurs travaux de recherche.

Un processus d'estimation sous calage à deux phases a été appliqué aux données d'une enquête récente sur la carrière et la mobilité des titulaires d'un doctorat (TD). Comme il n'existe pas de base de sondage des TD, les données sur les études supérieures ont été extraites d'un recensement de population récent. Cependant, seul un échantillon S_1 qui représente un cinquième des ménages dénombrés au recensement a reçu un questionnaire détaillé contenant des questions sur les études supérieures. Pour l'enquête sur les TD, on a tiré de S_1 un sous-échantillon S_2 dans lequel les personnes qui étaient en fait TD ont reçu un questionnaire encore plus détaillé. Donc, un scénario de calage à deux phases pour estimer les caractéristiques des TD était de mise. La première phase comprenait le calage des variables conjointes de S_1 et S_2 sur les totaux estimés calculés d'après S_1. À la deuxième phase, les données démographiques de S_1 ont été calées sur les totaux connus provenant du registre de la population complète U. Nous avons réalisé une étude en simulation sur ces données, dans laquelle les données d'enquête ont servi de population réelle. Mille échantillons (réalisations) $\{u, s_1, s_2\}$ de tailles $N = 1\,000$, $n_1 = 200$, $n_2 = 50$ ont été tirés aléatoirement de l'ensemble de données S_2 de TD. À chaque échantillon, nous avons appliqué le même processus de calage à deux phases en utilisant l'estimateur donné par (3.7) avec l'équation (3.6) comme représentation des poids calés $\tilde{w}_2$, et son estimateur de variance donné par (3.11) comme un cas particulier de (3.8). Comme nous l'avons déjà mentionné, quand $p = 2$, les estimations $\hat{Y} = \tilde{w}_2' y$ sont identiques sous la nouvelle représentation ou sous la représentation classique utilisée jusqu'à présent dans la littérature, Särndal et coll. (1992). Donc, nous nous sommes concentrés sur les estimateurs de variance (3.10) et (3.11) calculés selon les deux méthodes. Un profil type de la comparaison entre les deux estimateurs de variance dans ce cas particulier du calage à deux phases est présenté à la figure 4.1. On voit que, malgré la différence fondamentale entre les deux estimateurs de variance, dans la plupart des réalisations, la différence entre leurs estimations est assez faible. Néanmoins, pour l'une des réalisations, elle peut aller jusqu'à 20 %. Pour la variable particulière présentée dans la figure, les valeurs moyennes des deux estimateurs de la variance étaient très semblables, à savoir $54{,}17^2$ et $54{,}65^2$, tandis que la valeur réelle dans les données de simulation était de $54{,}46^2$. Même les variances de leur estimateur de l'écart-type, à savoir $5{,}73^2$ contre $5{,}93^2$, étaient presque les mêmes pour cette variable. Ces résultats sont présentés au tableau 4.1. La caractéristique favorable de l'estimateur proposé ressort dans la 5e colonne. Contrairement à l'estimateur classique dans lequel les deux termes de l'estimateur de variance sont du même ordre de grandeur, le 2e terme de (3.11) représente plus de 99 % de la variance, avec une variation de moins de 2 % sur l'ensemble des 1 000 réalisations. Nous avons donné l'explication de ce phénomène à la section 3.2. Les résultats présentés ici se sont répétés pour toutes les variables étudiées et nous avons jugé non pertinent à ce stade de présenter d'autres variables ou d'étudier plus en profondeur ces données particulières ou le cas particulier du calage à deux phases.

Tableau 4.1
Estimateur proposé (P) c. classique (C) pour l'écart-type d'un estimateur calé en deux phases

Variable	Valeur moyenne	É.-T.	Couverture de l'IC	2e terme en pourcentage de $\widehat{\text{É.-T.}}\left(\tilde{w}_2' y\right)$
$\tilde{w}_2' y$	200,43	54,46		
$\widehat{\text{É.-T.}}_C\left(\tilde{w}_2' y\right)$	54,65	5,93	95,2 %	77 % ± 7 %
$\widehat{\text{É.-T.}}_P\left(\tilde{w}_2' y\right)$	54,17	5,73	95,1 %	99 % ± 2 %

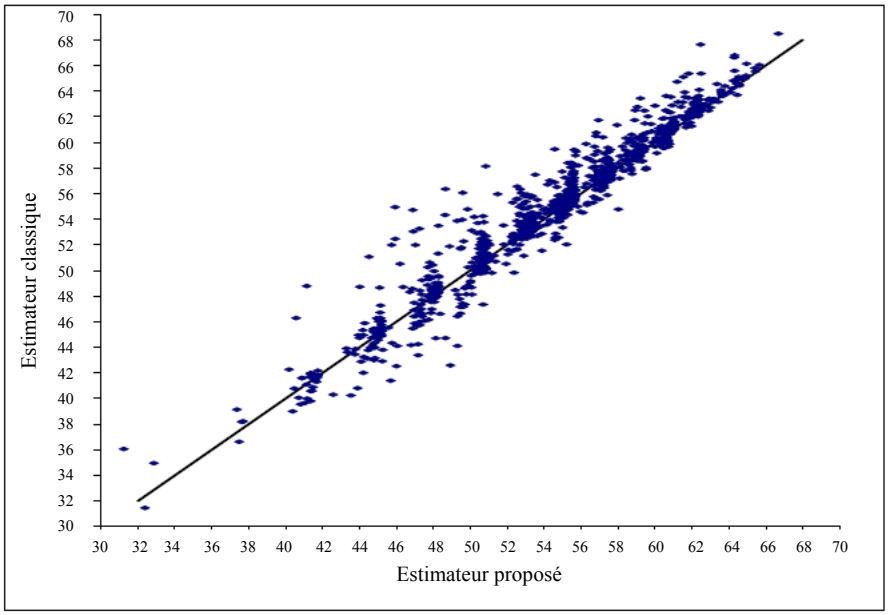

Figure 4.1 Estimations de la variance dans le calage à deux phases. Un profil type de 1 000 réalisations de l'estimateur proposé (équation 3.11) en fonction de l'estimateur classique (équation 3.10) pour la variance d'un estimateur calé de Y. La droite en trait plein est la diagonale principale.

La similarité des estimations des deux estimateurs de variance dans le cas du calage à deux phases est rassurante, mais il n'a pas été possible d'effectuer la comparaison dans le cas du calage à trois phases ou plus, parce qu'il n'existe pas d'alternative à l'estimateur proposé. Une méthode par rééchantillonnage pour l'échantillonnage à deux phases stratifié a été proposée par Kim et coll. (2006), et nous exposons brièvement une ébauche de généralisation pour un cas à trois phases, mais sans formulation explicite ni résultats de simulation. Nous avons ajouté une troisième phase de calage dans notre simulation en utilisant certaines variables en commun avec l'échantillon de deuxième phase des TD, choisies en fonction de l'expérience sur le terrain, et avons procédé de la même façon que dans le cas à deux phases. L'étude en simulation a de nouveau révélé une excellente estimation pour la variance d'un estimateur calé en trois phases pour toutes les variables Y examinées et chacun des différents ensembles de variables de calage à toutes les phases. Les taux de convergence de l'estimateur de variance sont rapides, même pour de très petites tailles d'échantillon, telles que $n = 25$ ou moins à la troisième phase. Certains résultats pour divers paramètres de plan de sondage sont présentés au tableau 4.2. Comme indiqué plus haut, la simulation a été exécutée sur une taille de population de 1 000 de manière que les trois premiers plans aient un poids global de $w^* = 40$, et les trois suivants, de $w^* = 20$. Donc, comme prévu, la variance de l'estimateur calé pour les trois premiers plans est généralement plus élevée, bien qu'elle dépende aussi des tailles d'échantillon des première et deuxième phases, comme le montre, par exemple, le cas artificiel numéro 4 qui dépeint un scénario généralement impossible en pratique. Les biais relatifs $\frac{E\left(\widehat{\text{É.-T.}}_P\right)}{\text{É.-T.}} - 1$ sont proches de zéro pour tous les plans étudiés et les couvertures des intervalles de confiance (IC) à 95 %, estimées également, se sont avérées principalement raisonnables et proches des niveaux nominaux. L'écart-type de $\widehat{\text{É.-T.}}_P$ vaut approximativement 5 % à 10 % de la valeur de l'estimateur, comme le montre la colonne 7.

Tableau 4.2
Valeurs vraie et estimée de l'écart-type d'un estimateur calé en trois phases $\tilde{w}_3' y$ pour divers paramètres de plan de sondage

Cas	n1	n2	n3	Valeur vraie	$\widehat{\text{É.-T.}}_P$	É.-T. de $\widehat{\text{É.-T.}}_P$ en %	Couverture de l'IC à 95 %
1	100	50	25	882,6	866,9	7,1 %	94,9 %
2	500	250	25	781,5	774,1	10,8 %	95,2 %
3	500	100	25	733,9	731,5	10,2 %	96,0 %
4	50	50	50	902,8	892,1	4,8 %	95,6 %
5	200	100	50	598,1	591,4	5,4 %	94,4 %
6	500	250	50	543,0	542,2	8,3 %	96,3 %
7	333	100	33	650,8	654,4	8,6 %	95,3 %

5 Conclusion

Le présent article décrit la construction d'une nouvelle représentation des poids calés en plusieurs phases qui permet de représenter un estimateur calé en plusieurs phases sous la forme d'un estimateur multivarié calé en une phase. Cette représentation rend possible le calcul d'une approximation sous forme explicite de la variance des estimateurs calés en plusieurs phases pour tout nombre de phases. Une comparaison avec une autre approximation connue dans la littérature pour le cas à deux phases montre que, même si les deux approximations sont convergentes, elles diffèrent en ce qui concerne leurs estimations, leur forme et leur interprétation. Nous avons discuté de certains avantages de la nouvelle approximation dans le cas du calage à deux phases et avons aussi montré sa convergence au moyen d'une étude en simulation du calage à trois phases où elle a donné de très bons résultats pour tous les plans étudiés. L'examen de l'efficacité de l'estimateur proposé en fonction des taux d'échantillonnage et d'autres paramètres du plan fera l'objet de futurs travaux de recherche.

Annexe A

Pour abréger la notation, nous effectuerons notre analyse sous forme matricielle. Nous utiliserons la convention selon laquelle, pour $j > i,$ la sommation dans les produits scalaires $X_i' w_j$ et $X_i' D_j$ (ou avec w_j^* ou $\tilde{w}_j$) est faite sur les unités $k \in s_j$ (et non sur s_i), c'est-à-dire sur l'échantillon indiqué par le dernier ensemble de poids dans le produit scalaire. D'où, $\hat{Z}_{ij} = \left(X_i' D_i^* X_i\right)^{-1} X_i' \left(D_j^* - D_{j-1}^*\right) X_j$ sous cette notation.

Preuve du lemme 3.1. Les poids qui satisfont l'équation de calage à la j^{e} phase avec les poids initiaux $\tilde{w}_{j-1}$ sont donnés par l'équation (3.4). Sous notre notation matricielle

$$\tilde{w}_j = D_j^* \left[g_1 + \ldots + g_j - (j-1)\right]$$

où $g_j = 1 + X_j T_j^{-1} \left(X_j' \tilde{w}_{j-1} - X_j' D_j \tilde{w}_{j-1}\right)$ (voir l'équation (3.5)). Donc

$$\begin{aligned} \tilde{w}_j &= D_{j-1}^* D_j \left[g_1 + \ldots + g_{j-1} - (j-2) + g_j - 1\right] \\ &= D_j \left[\tilde{w}_{j-1} + D_{j-1}^* \left(g_j - 1\right)\right]. \end{aligned} \tag{A.1}$$

L'insertion de g_j donne $\tilde{w}_j = D_j\left[\tilde{w}_{j-1} + D^*_{j-1}X_jT_j^{-1}\left(X_j'\tilde{w}_{j-1} - X_j'D_j\tilde{w}_{j-1}\right)\right]$ qui fait intervenir le poids $\tilde{w}_{j-1}$ provenant de la phase de calage précédente et son produit scalaire avec X_j' et $X_j'D_j$, tandis que les autres multiplicateurs sont des paramètres du plan. L'expression entre crochets contient trois opérandes et donc, après j phases de calage, nous aurions 3^j opérandes qui contiendraient uniquement des paramètres du plan. L'introduction de $\tilde{w}_{j-1}$ provenant de (A.1) par substitution dans $X_j'D_j\tilde{w}_{j-1}$ donne

$$\begin{aligned} X_j'D_j\tilde{w}_{j-1} &= X_j'D_j\left\{D_{j-1}\tilde{w}_{j-2} + D^*_{j-1}\left(g_{j-1} - 1\right)\right\} \\ &= X_j'D_jD_{j-1}\tilde{w}_{j-2} + X_j'D_jD^*_{j-1}X_{j-1}T_{j-1}^{-1}\left(X_{j-1}'\tilde{w}_{j-2} - X_{j-1}'D_{j-1}\tilde{w}_{j-2}\right) \quad \text{(A.2)} \end{aligned}$$

et donc aussi

$$X_j'\tilde{w}_{j-1} = X_j'D_{j-1}\tilde{w}_{j-2} + X_j'D^*_{j-1}X_{j-1}T_{j-1}^{-1}\left(X_{j-1}'\tilde{w}_{j-2} - X_{j-1}'D_{j-1}\tilde{w}_{j-2}\right). \quad \text{(A.3)}$$

La combinaison des termes donne une expression pour $\tilde{w}_j$ qui fait intervenir les poids calés provenant de la phase $j-2$ uniquement

$$\begin{aligned} \tilde{w}_j &= D_jD_{j-1}\tilde{w}_{j-2} \\ &\quad + D^*_jX_{j-1}T_{j-1}^{-1}\left(X_{j-1}'\tilde{w}_{j-2} - X_{j-1}'D_{j-1}\tilde{w}_{j-2}\right) \\ &\quad + D^*_jX_jT_j^{-1}\left(X_j'D_{j-1}\tilde{w}_{j-2} - X_j'D_jD_{j-1}\tilde{w}_{j-2}\right) \\ &\quad - D^*_jX_jT_j^{-1}\hat{Z}'_{j-1,\,j}\left(X_{j-1}'\tilde{w}_{j-2} - X_{j-1}'D_{j-1}\tilde{w}_{j-2}\right). \quad \text{(A.4)} \end{aligned}$$

L'insertion de (A.2) et (A.3) avec $j = p$ dans (A.1) et la récurrence $p-1$ fois sur les groupes de calage respectifs produisent le résultat souhaité.

Annexe B

Un estimateur convergent du total de population dans le calage à trois phases peut être représenté par $\hat{w}_3'y = \hat{Y}_{\mathrm{HT}_3} + \sum_{i=1}^{3}(\hat{t}_1^- - \hat{t}_1^+)'\,\hat{\gamma}_i$, où

$$\begin{aligned} \hat{\gamma}_1 &= \hat{\beta}_1 - \hat{Z}_{12}\hat{\beta}_2 - \hat{Z}_{13}\hat{\beta}_3 + \hat{Z}_{12}\hat{Z}_{23}\hat{\beta}_3 \\ \hat{\gamma}_2 &= \hat{\beta}_2 - \hat{Z}_{23}\hat{\beta}_3 \\ \hat{\gamma}_3 &= \hat{\beta}_3. \end{aligned}$$

Un estimateur convergent de la variance est donné par

$$\begin{aligned} \hat{V}_P\left(\tilde{w}_3'y\right) &= \sum_{k,\,l\in s_1}\left(w^*_{1k}w^*_{1l} - w^*_{1kl}\right)\hat{e}_{1k}\hat{e}_{1l} + \ldots + \sum_{k,\,l\in s_3}\left(w^*_{3k}w^*_{3l} - w^*_{3kl}\right)\hat{e}_{3k}\hat{e}_{3l} \\ &\quad + 2\sum_{k\in s_1,\,l\in s_2} w_{2l}\left(w_{1k}w_{1l} - w_{1kl}\right)\hat{e}_{1k}\hat{e}_{2l} + 2\sum_{k\in s_2,\,l\in s_3} w_{3l}\left(w^*_{2k}w^*_{2l} - w^*_{2kl}\right)\hat{e}_{2k}\hat{e}_{3l} \\ &\quad + 2\sum_{k\in s_1,\,l\in s_3} w_{2l}w_{3l}\left(w^*_{3k}w^*_{3l} - w^*_{3kl}\right)\hat{e}_{1k}\hat{e}_{3l}. \end{aligned}$$

où $\hat{e}_{1k} = x'_{1k}\hat{\gamma}_1 - x'_{2k}\hat{\gamma}_2$, $\hat{e}_{2k} = x'_{2k}\hat{\gamma}_2 - x'_{3k}\hat{\gamma}_3$ et $\hat{e}_{3k} = x'_{3k}\hat{\gamma}_3 - y_k$ sont définis au théorème 3.1.

Bibliographie

Binder, D.A. (1996). Méthodes de linéarisation pour les échantillons à une et deux phases : Une approche de type « recette ». *Techniques d'enquête,* 22, 1, 17-22. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1996001/article/14389-fra.pdf.

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M. et Jocelyn, W. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics,* 28, 751-764.

Breidt, J., et Fuller, W.A. (1993). Regression weighting for multiphase samples. *Sankhyā,* 55, 297-309.

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

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association,* 87, 418, 376-382.

Farell, P.J., et Singh, S. (2002). Penalized chi-square distance function in survey sampling. *Proceedings of Joint Statistical Meeting,* NY, États-Unis.

Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica,* 8, 1153-1164.

Hidiroglou, M.A., et Särndal, C.-E. (1998). Emploi des données auxiliaires dans l'échantillonnage à deux phases. *Techniques d'enquête,* 24, 1, 11-20. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1998001/article/3905-fra.pdf.

Kim, J.K., Navarro, A. et Fuller, W.A. (2006). Replicate variance estimation after multi-phase stratified sampling. *Journal of American Statistical Association,* 101, 312-320.

Kott, P.S., et Stukel, D.M. (1997). La méthode du jackknife convient-elle à un échantillon à deux phases ? *Techniques d'enquête,* 23, 2, 89-98. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1997002/article/3621-fra.pdf.

Rao, J.N.K. (1973). On double sampling for stratification and analytic surveys. *Biometrika,* 6, 125-133.

Rao, J.N.K., et Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika,* 79, 811-822.

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

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 32, No. 4, December 2016

All inquires about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 33, No. 1, March 2017

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 45, No. 1, March/mars 2017

Issue Information

Original Articles

Acknowledgement

CONTENTS TABLE DES MATIÈRES

Volume 45, No. 2, June/juin 2017

Issue Information

Original Articles

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique au rédacteur en chef (statcan.smj-rte.statcan@canada.ca). Avant de soumettre l'article, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 39, n° 1) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word et MathType pour les expressions mathématiques. Une version pdf ou papier pourrait être requise pour les formules et graphiques.

1. Présentation

1.1 Les textes doivent être écrits à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.

1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.

1.3 Le nom (écrit au long) et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.

1.4 Les remerciements doivent paraître à la fin du texte.

1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.

3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(\cdot)$ et $\log(\cdot)$, etc.

3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées par un chiffre arabe à la droite si l'auteur y fait référence plus loin. Utiliser un système de numérotation à deux niveaux selon le numéro de la section. Par exemple, l'équation (4.2) est la deuxième équation importante de la section 4.

3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.

3.5 Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0 ; l, 1).

3.6 Si possible, éviter l'emploi de caractères gras dans les formules.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux). Utiliser un système de numérotation à deux niveaux selon le numéro de la section. Par exemple, le tableau 3.1 est le premier tableau de la section 3.

4.2 Une description textuelle détaillée des figures pourrait être requise à des fins d'accessibilité si le message transmis par l'image n'est pas suffisamment expliqué dans le texte.

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple : Cochran (1977, page 164).

5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

6. Communications brèves

6.1 Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.