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

Techniques d’enquête

Techniques d’enquête
43-2

Date de diffusion : le 21 décembre 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

décembre 2017

•

Volume 43

•

Numéro 2

Statistique Canada Statistics Canada

Canada

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

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

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 2, décembre 2017

Table des matières

Article spécial

Articles réguliers

Communication brève

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, décembre 2017
Vol. 43, nº 2, pp. 159-176
Statistique Canada, nº 12-001-X au catalogue

Théorie et méthodologie des enquêtes par sondage : orientations passées, présentes et futures

J.N.K. Rao et Wayne A. Fuller[1]

Résumé

L'exposé retrace l'évolution de la théorie et de la méthodologie des enquêtes par sondage au cours des 100 dernières années. Dans un article fondamental publié en 1934, Neyman jetait les bases théoriques de l'approche axée sur l'échantillonnage probabiliste pour l'inférence à partir d'échantillons d'enquête. Les traités d'échantillonnage classiques publiés par Cochran, Deming, Hansen, Hurwitz et Madow, Sukhatme, ainsi que Yates au début des années 1950 étendaient et étoffaient la théorie de l'échantillonnage probabiliste, en mettant l'accent sur l'absence de biais, les caractéristiques exemptes de modèle, ainsi que les plans de sondage qui minimisent la variance selon un coût fixe. De 1960 à 1970, l'attention s'est portée sur les fondements théoriques de l'inférence à partir de données d'enquêtes, contexte dans lequel l'approche dépendante d'un modèle a suscité d'importantes discussions. L'apparition de logiciels statistiques d'usage général a entraîné l'utilisation de ces derniers avec des données d'enquêtes, d'où la conception de méthodes spécialement applicables aux données d'enquêtes complexes. Parallèlement, des méthodes de pondération telles que l'estimation par la régression et le calage devenaient réalisables et la convergence par rapport au plan de sondage a remplacé la contrainte d'absence de biais comme critère pour les estimateurs classiques. Un peu plus tard, les méthodes de rééchantillonnage gourmandes en ressources informatiques sont également devenues applicables à des échantillons d'enquêtes à grande échelle. L'augmentation de la puissance informatique a permis des imputations plus avancées des données manquantes, l'utilisation d'une plus grande quantité de données auxiliaires, le traitement des erreurs de mesure dans l'estimation, et l'application de procédures d'estimation plus complexes. Une utilisation marquante de modèles a eu lieu dans le domaine en expansion de l'estimation sur petits domaines. Les orientations futures de la recherche et des méthodes seront influencées par les budgets, les taux de réponse, le degré d'actualité des données, les outils améliorés de collecte des données et l'existence de données auxiliaires, dont une partie proviendra des « mégadonnées ». L'évolution des comportements culturels et de l'environnement physico-technique aura une incidence sur la façon de réaliser les enquêtes.

Mots-clés : Collecte des données; histoire de l'échantillonnage; échantillonnage probabiliste; inférence à partir d'enquêtes.

1 Introduction

Le présent article a été préparé à la demande de M. Danny Pfeffermann, Ph.D., président de l'Association internationale des statisticiens d'enquête en 2015, qui a proposé le titre ambitieux. L'article a été présenté à l'assemblée de l'Institut international de statistique à Rio de Janeiro, au Brésil, en 2015.

Le titre définit un domaine beaucoup trop vaste pour pouvoir le traiter complètement en un seul article. De surcroît, plusieurs articles de synthèse abordent les sujets qu'évoque ce titre, dont Kish (1995), Bellhouse (2000), Rao (2005), Bethlehem (2009), Brick (2011), Groves (2011) et Brewer (2013). Nous nous inspirons de ces articles dans notre discussion, mais sans chercher l'exhaustivité. Nous donnons une brève évaluation des trois sujets et projetons un certain nombre de situations courantes dans l'avenir. Notre but est de susciter de nouvelles discussions, surtout au sujet des orientations futures. Au-delà de la discussion des controverses concernant l'échantillonnage par choix raisonné, nous nous concentrerons sur l'échantillonnage probabiliste. Le domaine de l'échantillonnage étant de type appliqué, nous parlerons de certains problèmes que l'on rencontre et de certaines méthodes que l'on emploie en pratique. Notre discussion s'applique

1. J.N.K. Rao, professeur et chercheur distingué, Université Carleton, Ottawa, Canada, K15 5B6. Courriel : jrao@math.carleton.ca; Wayne A. Fuller, professeur distingué émérite, Iowa State University, Ames, IA, É.-U. 50011. Courriel : waf@iastate.edu.

surtout aux grands échantillons polyvalents, c'est-à-dire les enquêtes au sujet desquelles nous avons le plus d'expérience. De même, notre connaissance des applications est concentrée au Canada et aux États-Unis.

La présentation de l'article est la suivante. La section 2 décrit les premières contributions marquantes de 1920 à 1960. Les problèmes inférentiels sont traités à la section 3. L'article se termine par une discussion sur l'avenir à la section 4.

2 Premières contributions marquantes : 1920 à 1960

Kiaer (1897) fut sans doute le premier à promouvoir l'échantillonnage (ou ce que l'on appelait à l'époque la méthode représentative) au lieu du dénombrement complet (recensement), quoique la référence la plus ancienne remonte à l'an 1000 avant notre ère. L'objectif de la méthode représentative est d'obtenir un échantillon qui reflète la population finie parente, ce que l'on peut réaliser par échantillonnage équilibré sur des totaux auxiliaires connus, par échantillonnage par choix raisonné ou par échantillonnage aléatoire menant à des probabilités d'inclusion égales. Dès les années 1920, la méthode représentative était d'usage très répandu. L'Institut international de statistique (IIS) a joué un rôle essentiel en créant un comité d'experts chargé de faire rapport sur cette méthode. La contribution de Bowley (1926) au rapport de l'IIS comprend ses travaux fondamentaux sur l'échantillonnage aléatoire stratifié avec répartition proportionnelle menant à des probabilités d'inclusion égales. Bowley (1936) déclare que la [*Traduction*] « première application de ce principe » d'inférer pour la population à partir d'un échantillon a été l'étude dans *Reading* en 1912. Bowley a recommandé que la méthode d'échantillonnage pour cette étude soit un échantillonnage systématique d'une liste de maisons. Bowley a appelé cette procédure systématique une [*Traduction*] « méthode pure d'échantillonnage » et a déclaré [*Traduction*] « C'est littéralement la méthode d'échantillonnage stratifié ». Bowley donne des exemples où l'échantillonnage systématique a été utilisé après 1912. Bowley (1936) souligne l'importance d'une base complète et des probabilités de sélection égales. Mais c'est Neyman (1934) qui a jeté les bases de l'échantillonnage probabiliste (ou approche fondée sur le plan de sondage). Il a démontré que l'échantillonnage aléatoire stratifié était préférable à l'échantillonnage équilibré (représentatif) tel qu'il était utilisé à l'époque. Il a également présenté le concept d'efficacité et de répartition optimale de l'échantillon, appelée aujourd'hui répartition de Neyman, qui minimise la taille totale de l'échantillon pour une précision spécifiée en relâchant la condition de probabilités d'inclusion égales de Bowley. En fait, Tchuprow (1923) avait obtenu la répartition de Neyman 10 ans plus tôt, dans un article découvert après la parution de celui de Neyman. Neyman (1934) a également montré que, pour des échantillons suffisamment grands, on pouvait obtenir pour la moyenne de population d'une variable d'intérêt des intervalles de confiance tels que la fréquence des erreurs mentionnée dans l'énoncé de confiance sous échantillonnage répété ne dépasse pas la limite établie, « quelles que soient les propriétés inconnues de la population ». Plus récemment, l'échantillonnage équilibré, préconisé au départ par Gini et Galvani, a été peaufiné de manière à intégrer les caractéristiques intéressantes de l'échantillonnage probabiliste ainsi que de l'échantillonnage équilibré sur des totaux auxiliaires connus (Deville et Tillé, 2004). La nouvelle méthode d'échantillonnage équilibré est maintenant appliquée en Europe, surtout en France, pour tirer des échantillons pour les enquêtes auprès des établissements. Une deuxième méthode de

sélection probabiliste contrôlée est l'échantillonnage réjectif, introduit par Hájek (1964) comme méthode de contrôle de la taille de l'échantillon dans l'échantillonnage de Poisson. Fuller (2009a) a étendu la procédure de manière à restreindre les échantillons acceptables à l'ensemble pour lequel les estimations de la moyenne des variables auxiliaires sont proches de la moyenne de population.

Au cours des années 1930, alors que s'accélérait la demande de renseignements socioéconomiques, les avantages de l'échantillonnage probabiliste, dont la réalisation d'études de plus grande portée, à moindre coût et plus rapidement que des recensements, ont bientôt été reconnus partout dans le monde. On a donc assisté à une multiplication et à une diversification des enquêtes fondées sur l'échantillonnage probabiliste et couvrant de grandes populations. Accepté presque universellement, l'échantillonnage probabiliste (ou approche fondée sur le plan de sondage) de Neyman est devenu un outil standard pour la recherche empirique en sciences sociales et en statistique officielle. On a également reconnu que la précision d'un estimateur dépendait en grande partie de la taille de l'échantillon et non de la fraction d'échantillonnage. Les années 1940 ont vu des études portant sur les propriétés de l'échantillonnage systématique pour différentes populations, voir Madow et Madow (1944), Cochran (1946), et Yates (1948). Cochran (1977, chapitre 8) présente une excellente discussion sur l'échantillonnage systématique clarifiant pourquoi seuls les estimateurs de variance fondés sur le modèle sont possibles. Voir aussi Bellhouse (1988). Au début, l'élaboration de la théorie de l'échantillonnage était axée sur l'estimation des totaux et des moyennes, et des erreurs d'échantillonnage connexes. Les erreurs non dues à l'échantillonnage, dont la non-réponse, les erreurs de couverture et les erreurs de mesure, étaient en grande partie ignorées en recherche théorique.

Passons maintenant à quelques avancées théoriques importantes concernant l'approche fondée sur le plan de sondage réalisées après Neyman. Dès 1937, Mahalanobis a utilisé des plans d'échantillonnage à plusieurs degrés pour les enquêtes sur les récoltes en Inde. Dans son article classique publié en 1944 (Mahalanobis, 1944), il a formulé rigoureusement les fonctions de coût et de variance permettant une conception efficace des enquêtes. Il a joué un rôle essentiel dans la création du *National Sample Survey* de l'Inde, la plus grande enquête polyvalente continue effectuée par du personnel à temps plein réalisant des interviews sur place pour les enquêtes socioéconomiques et des mesures physiques pour les enquêtes sur les récoltes. Sukhatme, qui fut l'élève de Neyman, a également contribué de manière novatrice à la conception et à l'analyse des enquêtes agricoles à grande échelle en Inde, en faisant appel à l'échantillonnage à plusieurs degrés stratifié. Les traités classiques sur l'échantillonnage publiés par Cochran (1953), Deming (1950), Hansen, Hurwitz et Madow (1953), Sukhatme (1954) et Yates (1949) ont été utiles aux étudiants ainsi qu'aux praticiens.

Durant la période de 1940 à 1960, sous la direction de Morris Hansen, les statisticiens d'enquête du *U.S. Census Bureau* ont contribué de manière fondamentale à la théorie et à la méthodologie des enquêtes par sondage. Cette période est considérée comme l'âge d'or du *Census Bureau*. Hansen et Hurwitz (1943) ont élaboré la théorie de base de l'échantillonnage en grappes à deux degrés stratifié, avec tirage d'une grappe (ou unité primaire d'échantillonnage) dans chaque strate avec probabilité proportionnelle à la taille (PPT), puis sous-échantillonnage de ces grappes à un taux assurant un échantillon autopondéré (probabilités globales de sélection égales). La sélection de grappes avec probabilités inégales peut réduire considérablement la variance en contrôlant la variabilité découlant des tailles de grappe inégales. Une autre contribution importante du *U.S. Census Bureau* a été l'utilisation de l'échantillonnage rotatif avec

renouvellement partiel des ménages pour résoudre le problème du fardeau de réponse dans les enquêtes répétées au cours du temps, comme la *Current Population Survey* réalisée mensuellement aux États-Unis pour mesurer les taux de chômage. Hansen, Hurwitz, Nisselson et Steinberg (1955) ont établi des estimateurs composites simples, mais efficaces sous échantillonnage rotatif. L'usage de l'échantillonnage rotatif et de l'estimation composite est très répandu dans les enquêtes continues à grande échelle.

Avant les années 1950, l'objectif principal était l'estimation des totaux et des moyennes de population. Au *U.S. Census Bureau*, Woodruff (1952) a élaboré une approche unifiée de construction d'intervalles de confiance pour les quantiles (en particulier, la médiane) applicable aux plans d'échantillonnage généraux. La procédure demeure l'une des pierres angulaires de l'estimation des quantiles (Francisco et Fuller, 1991).

Après la consolidation de la théorie fondamentale de l'échantillonnage fondé sur le plan de sondage, Hansen, Hurwitz, Marks et Mauldin (1951) et d'autres se sont penchés sur les erreurs de mesure ou de réponse dans les données d'enquêtes. Sous des modèles d'erreur de mesure additifs avec hypothèses de modèle minimales sur les réponses observées traitées comme des variables aléatoires, la variance totale d'un estimateur peut être décomposée en une variance d'échantillonnage, une variance de réponse simple et une variance de réponse corrélée (VRC) attribuable aux intervieweurs.

Mahalanobis (1946) a établi la méthode des sous-échantillons interpénétrants pour évaluer les erreurs d'échantillonnage ainsi que d'intervieweur. En répartissant les sous-échantillons au hasard entre les intervieweurs, il est possible d'estimer la variance totale ainsi que la composante due à l'intervieweur. Cette dernière peut dominer la variance totale quand le nombre d'intervieweurs est faible. En vue d'éliminer la composante VRC due aux intervieweurs, un autodénombrement par la poste a été lancé pour le Recensement des États-Unis de 1960.

La question de la non-réponse aux enquêtes a également été abordée durant les premiers travaux sur l'échantillonnage. Hansen et Hurwitz (1946) ont proposé l'échantillonnage à deux phases qui consiste à prendre contact par la poste avec l'échantillon sélectionné à la première phase, puis à procéder à l'interview sur place d'un sous-échantillon de non-répondants, en supposant que la réponse sera complète ou que la non-réponse sera négligeable à la deuxième phase. Cette méthode a été utilisée récemment au Canada quand l'échantillon pour le questionnaire long à réponse obligatoire du recensement a été remplacé par l'Enquête nationale auprès des ménages à participation volontaire. Après le changement de gouvernement en 2015, le premier ministre du Canada a rétabli le questionnaire long du Recensement. L'échantillonnage à deux phases continuera d'être utilisé, mais dans une moindre mesure. La méthode d'échantillonnage à deux phases de Hansen-Hurwitz a aussi été utilisée dans d'autres enquêtes incluant l'*American Community Survey*.

Une attention particulière a aussi été accordée aux inférences pour des sous-populations non planifiées (appelées domaines), telles que les groupes âge-sexe à l'intérieur d'un État. Hartley (1959) et Durbin (1958) ont élaboré une théorie unifiée pour l'estimation de domaines applicable à des plans de sondage généraux et ne nécessitant que des formules existantes pour le calcul des totaux et des moyennes de population.

Durant cette première période, la théorie de l'échantillonnage a été établie en majeure partie par des praticiens de la statistique officielle, tandis que les chercheurs universitaires, surtout aux États-Unis, accordaient peu d'attention à l'échantillonnage. Faisait exception la *Iowa State University*, où le corps

professoral a joué un rôle de chef de file dès le départ sous la direction de Cochran, Jessen et Hartley. Une autre institution qui a contribué très tôt à la pratique des enquêtes et à la recherche sur ces derniers est le *Survey Research Center* de l'Université du Michigan créé en 1947 et dont Leslie Kish fut un des premiers membres.

Au cours des années 1950, des cadres théoriques formels pour l'inférence fondée sur le plan de sondage pour des totaux et des moyennes ont été proposés en considérant les données d'échantillon comme un ensemble d'étiquettes d'échantillon groupé avec les variables d'intérêt associées. Horvitz et Thompson (1952) ont dérivé l'estimateur bien connu dans lequel la pondération est inversement proportionnelle à la probabilité d'inclusion. Narain (1951) a également proposé cet estimateur. Godambe (1955) a élaboré une classe générale d'estimateurs linéaires en permettant que le poids de sondage d'une unité dépende de l'étiquette de celle-ci, ainsi que des étiquettes des autres unités dans l'échantillon. Il a ensuite montré que le meilleur estimateur linéaire sans biais n'existe pas dans cette classe générale, même sous échantillonnage aléatoire simple.

3 Problèmes d'inférence : 1950 -

3.1 Fondements théoriques

Des tentatives ont eu lieu en vue d'intégrer la théorie des enquêtes par sondage à l'inférence statistique traditionnelle en se servant de la fonction de vraisemblance. Godambe (1966) a montré que la fonction de vraisemblance issue de l'ensemble complet de données d'échantillon, y compris les étiquettes, en considérant le vecteur des valeurs de population inconnues comme étant le paramètre, ne fournit aucun renseignement sur les valeurs non échantillonnées ni, donc, sur le total ou la moyenne de population. Cet aspect non informatif de la fonction de vraisemblance est dû à l'inclusion des étiquettes dans l'ensemble de données, ce qui rend l'échantillon unique. Une autre route fondée sur le plan de sondage ne tient pas compte de certains aspects des données d'échantillon afin de rendre l'échantillon non unique et, donc, arriver à des fonctions de vraisemblance informatives (Hartley et Rao, 1968; Royall, 1968). Cette approche de la vraisemblance non paramétrique est similaire à l'approche de la vraisemblance empirique (VE) populaire à l'heure actuelle en inférence statistique classique (Owen, 1988). L'approche de la vraisemblance empirique a été appliquée récemment aux problèmes d'échantillonnage pour estimer non seulement les totaux et les moyennes, mais aussi des paramètres plus complexes. Ainsi, les efforts d'intégration à la statistique classique ont été partiellement fructueux.

L'approche dépendante d'un modèle offre un autre chemin vers l'inférence à partir de données d'enquêtes. Cette approche requiert que la structure de la population obéisse à un modèle de superpopulation spécifié. La distribution induite par le modèle supposé sert de base pour les inférences (Brewer, 1963 et Royall, 1970). Ce genre d'inférences conditionnelles (à l'échantillon) peut être intéressant. Cependant, les estimateurs résultants sont parfois non convergents par rapport au plan et, par conséquent, peuvent donner de mauvais résultats pour les grands échantillons si la spécification du modèle est incorrecte (Hansen, Madow et Tepping, 1983).

Une approche hybride, appelée approche assistée par modèle, vise à combiner les caractéristiques désirables des méthodes fondées sur le plan de sondage et de celles dépendantes d'un modèle; consulter Cassel, Särndal et Wretman (1976). Cette approche comprend habituellement l'utilisation de données ne provenant pas de la collecte des données, appelées données auxiliaires. Les procédures qui font appel à des données auxiliaires comprennent l'estimation par la régression, l'estimation par le ratio et le *raking* (ou méthode de ratissage), c'est-à-dire des méthodes avec estimateurs linéaires en la variable d'intérêt. La force des estimateurs utilisant de l'information auxiliaire, particulièrement les estimateurs par régression, a été reconnue très tôt (Cochran, 1953). Au cours des années 1970, grâce à la puissance des ordinateurs, l'estimation par la régression est devenue réalisable, mais pour être acceptable dans les enquêtes à grande échelle, il faut que les poids de régression ne soient pas négatifs. Une première définition des poids non négatifs figure dans Huang et Fuller (1978). Deville et Särndal (1992) ont donné une méthode générale de construction des poids pour des estimateurs convergents par rapport au plan. Les méthodes assistées par modèle ne concernent que les estimateurs du total convergents par rapport au plan qui sont également sans biais sous le modèle pour un modèle de travail. Cette approche est utile pour les grands échantillons et aboutit à des inférences fondées sur le plan de sondage acceptables, indépendamment de la validité du modèle de travail. Cependant, l'efficacité des estimateurs dépend du degré auquel le modèle de travail s'approche de la vraie structure de population. La catégorie la plus répandue d'estimateurs assistés par modèle est celle des estimateurs par la régression généralisée (GREG) qui sont mis en œuvre dans les progiciels pour données d'enquêtes.

Les résultats théoriques concernant l'échantillonnage probabiliste mettent l'accent sur les deux premiers moments de la statistique d'échantillon. Des théorèmes centraux limites ont été utilisés pour justifier des intervalles de confiance fondés sur la normalité. Un des premiers théorèmes centraux limites pour des échantillons aléatoires simples est celui de Madow (1948). Hájek (1960) a donné un théorème central limite pour l'échantillonnage aléatoire simple et un théorème pour l'échantillonnage réjectif dans Hájek (1964). Bickel et Freedman (1984) ont donné un théorème central limite pour l'échantillonnage aléatoire stratifié. Une publication récente prend en considération des séries de populations finies fixes ainsi que des séries de populations finies qui sont des échantillons tirés d'une superpopulation (Fuller, 2009b, section 1.3.2).

Durant les années 1930 et les années 1940, le coût de l'estimation de la variance était très élevé, presque prohibitif, et il demeure élevé aujourd'hui. Dès le départ, le rééchantillonnage a été adopté comme méthode efficace d'estimation de la variance. Ainsi que nous l'avons mentionné, Mahalanobis (1939, 1946) a proposé une forme précoce de rééchantillonnage, qu'il a appelé échantillons « interpénétrants » et que des auteurs ultérieurs ont appelé « groupes aléatoires ». La méthode des groupes aléatoires fondés sur des demi-échantillons a été utilisée par le *U.S. Census Bureau* dans les années 1950 et 1960. McCarthy (1966, 1969) a développé et décrit l'estimation de la variance pour les demi-échantillons équilibrés. Voir aussi Kish et Frankel (1970). Wolter (2007) présente une discussion élaborée sur les demi-échantillons équilibrés. Voir aussi Dippo, Fay et Morgenstein (1984), Kish et Frankel (1974), Krewski et Rao (1981), et Rao et Shao (1999). Le jackknife et le bootstrap sont les versions courantes des premières procédures de rééchantillonnage. Wolter (2007, chapitre 4) attribue à Durbin (1959) la première utilisation du jackknife dans l'estimation sur population finie. L'utilisation du bootstrap dans les conditions classiques remonte à Efron (1979), mais son application aux échantillons tirés avec probabilités inégales et aux populations finies

n'a pas été immédiate. Parmi le grand nombre d'articles traitant du jackknife et du bootstrap pour les échantillons d'enquête figurent ceux de McCarthy et Snowden (1985), qui présentent une première version de l'échantillonnage avec remise, et de Rao et Wu (1988), qui décrivent un bootstrap modifié fondé sur un « rééchelonnement » (*rescaling*) des poids pour les échantillons d'enquête. Sitter (1992) a abordé plusieurs questions et, notamment, a fait des suggestions pour obtenir des tailles d'échantillon entières. Antal et Tillé (2011) ont exposé des méthodes bootstrap appropriées pour une grande gamme de plans de sondage, y compris l'échantillonnage de Poisson. Beaumont et Patak (2012) ont proposé des procédures bootstrap générales.

3.2 Usage analytique des données d'enquêtes

Comme nous l'avons souligné, les premiers travaux sur l'échantillonnage probabiliste étaient axés sur les totaux et les moyennes, et bon nombre de procédures d'estimation ont été élaborées pour la statistique officielle. Cependant, dès le début, les spécialistes des sciences sociales ont utilisé des échantillons d'enquête pour obtenir des réponses aux questions dans ce domaine applicables au-delà de la population finie échantillonnée. Deming et Stephan (1940) et Deming (1953) ont tenu compte explicitement de la différence entre les usages « énumératif » et « analytique » des données d'enquête et de recensement; voir aussi Hartley (1959). Les estimations analytiques sont parfois appelées estimations pour une superpopulation. Les premiers analystes traitaient souvent les données d'enquête par sondage comme provenant d'un échantillon aléatoire simple et construisaient des estimations sur cette base. Puisque ne pas tenir compte du plan de sondage pouvait introduire un biais, une théorie de l'estimation a été élaborée pour les estimations analytiques. L'une de ses composantes consistait en des tests pour évaluer les effets des poids sur les estimations; voir DuMouchel et Duncan (1983), Fuller (1984), et Korn et Graubard (1995). Une deuxième composante était l'élaboration d'une théorie fondée sur le plan de sondage pour les statistiques compliquées. Voir Fuller (1975), Rao et Scott (1981, 1984), ainsi que Binder et Roberts (2003). La troisième composante consistait à intégrer le plan de sondage dans le modèle (Skinner, 1994 et Pfeffermann et Sverchkov, 1999). Il existe aujourd'hui un certain nombre de progiciels (SAS, SUDAAN, R, STATA) pour calculer les statistiques et les erreurs-types fondées sur l'approche probabiliste. Nombre des algorithmes remontent aux travaux effectués à la *Iowa State University* (Hidiroglou, Fuller et Hickman, 1976).

3.3 Données manquantes

Presque tous les échantillons (et toutes les expériences) présentent des données manquantes ou incorrectes. Dans le cas des enquêtes par sondage, les données manquantes sont réparties en deux catégories, à savoir les unités manquantes (non-réponse totale) et les questions manquantes (non-réponse partielle), l'expression unité manquante signifiant que tous les items (variables) manquent dans l'enregistrement de la réponse de l'unité. L'ensemble de monographies publié par Madow, Nisselson et Olkin (1983) témoigne de l'importance des données manquantes dans les études par sondage. L'une des méthodes de traitement des données manquantes se résumait à communiquer la nature et le nombre des items manquants et à totaliser les items restants. Au début, cette méthode était courante, mais l'hypothèse implicite d'interchangeabilité

sur laquelle elle reposait n'était pas souvent raisonnable. Une des premières méthodes de correction de la non-réponse totale consistait à utiliser un répondant de substitution, ce qui revenait souvent à interviewer une personne « proche » du non-répondant. Une modification courante à l'étape de l'analyse était, et demeure, la poststratification (Deming, 1953; Thomsen, 1973; Kalton, 1983 et Jagers, 1986). Dans la littérature sur les données manquantes, les poststrates sont souvent appelées cellules. Les estimateurs par la régression, qui sont des extensions directes des estimateurs de cellule, représentent une méthode importante de correction des données manquantes (Fuller et An, 1998). Les méthodes de pondération en vue de traiter la non-réponse totale sont passées en revue par Brick et Montaquila (2009).

Diverses formes d'imputation pour corriger la non-réponse partielle ont été appliquées au fil du temps, l'imputation étant effectuée manuellement avant l'usage des ordinateurs. L'une des premières imputations formelles fondées sur un modèle et effectuées par ordinateur était la procédure d'imputation hot deck utilisée par le *U.S. Census Bureau* pour la *Current Population Survey* de 1947; voir la description dans Andridge et Little (2009). L'accroissement de la puissance de calcul et les progrès techniques (Little, 1982; Kalton et Kish, 1984; Rubin, 1974, 1976, 1987; Little et Rubin, 1987; Kim et Fuller, 2004) ont fait de l'imputation un élément standard de l'estimation sur des échantillons d'enquête et un domaine de recherche dynamique. Parmi les ouvrages récents à ce sujet, mentionnons Kim et Shao (2013) et Little et Rubin (2014).

3.4 Estimation sur petits domaines

L'usage croissant de modèles pour l'estimation sur petits domaines résulte de la conjonction de deux facteurs. Le premier est la demande d'estimations pour de petits domaines (par exemple, régions géographiques) pour la formulation de politiques, la répartition des fonds et la planification régionale. Le second est la grande erreur-type associée à de nombreux estimateurs de domaines sous le plan de sondage. Schaible (1996) et Purcell et Kish (1979) ont donné tôt des exemples d'estimation sur petits domaines; voir aussi Gonzalez (1973) et Steinberg (1979). Dès 1947, le *U.S. Census Bureau* utilisait des méthodes fondées sur un modèle pour l'estimation sur petits domaines (Hansen et coll., 1953; Vol. I, pages 483-486). Plus récemment, les modèles linéaires mixtes incluant des effets fixes et aléatoires ont pris de l'importance. Les premiers usages de modèles mixtes pour l'estimation sur petits domaines sont décrits dans Fay et Herriot (1979) et dans Battese, Harter et Fuller (1988). Certains ensembles d'estimations sur petits domaines peuvent être considérés comme une réaffectation des estimations de domaines, en maintenant l'estimation directe du total général convergente par rapport au plan. Les méthodes bayésiennes, en particulier les méthodes bayésiennes hiérarchiques, sont de plus en plus fréquemment utilisées, en raison de la capacité à traiter des modèles complexes; voir Rao et Molina (2015, chapitre 10). En raison de la demande croissante, les publications en la matière se sont multipliées et l'estimation sur petits domaines fait aujourd'hui l'objet d'assemblées de spécialistes fréquentes, ainsi que d'un livre (Rao, 2003) dont la seconde édition est parue récemment (Rao et Molina, 2015).

3.5 Pratiques des enquêtes

Les questions relatives aux plans de sondage et à l'estimation dont nous venons de discuter sont des aspects essentiels d'une opération d'enquête, mais ne représentent qu'une petite fraction de l'ensemble. La

qualité du produit fini dépend des données qui figurent dans la base de sondage, de l'instrument de collecte, de la collecte, de la vérification et du traitement des données, ainsi que de la présentation des résultats. De nombreuses sources d'erreur sont difficiles à mesurer, mais les concepteurs des enquêtes font des estimations de coût implicites quand ils répartissent les ressources entre les différents volets de l'opération d'enquête. Groves et Lyberg (2010) passent en revue les tentatives en vue d'énumérer les composantes de la qualité d'une enquête et de les regrouper sous un chapiteau unique. Ils attribuent à Deming (1944) une première description des sources d'erreur dans les enquêtes par sondage et décrivent les contributions de Dalenius (1974), Anderson, Kasper et Frankel (1979), Groves (1989), Biemer et Lyerg (2003), entre autres. Groves et Herringa (2006) ont proposé des outils de contrôle dynamique des erreurs et des coûts d'enquête qui permettent d'aboutir à des plans de collecte adaptatifs pour les enquêtes auprès des ménages. En particulier, les paradonnées (mesures concernant le processus de collecte des données d'enquêtes) peuvent servir à surveiller le travail sur le terrain, à décider quand il faut intervenir durant la collecte des données, et à traiter l'erreur de mesure, la non-réponse et les erreurs de couverture (Kreuter, 2013).

4 L'avenir

Nous pouvons projeter un certain nombre de situations courantes dans l'avenir. Le contexte budgétaire sera strict et les demandes de produits augmenteront. Il y aura une demande de prévisions, ainsi que d'amélioration de l'accès des utilisateurs aux données. Il y aura une demande d'accélération de la production des statistiques, et ce, naturellement, sans compromettre la qualité. Une pression sera exercée afin que concordent les estimations provenant de sources différentes.

Nous nous attendons à ce que la plus grande vitesse des ordinateurs influe sur tous les aspects du domaine. Des algorithmes de vérification et d'imputation plus complexes seront établis. Le temps écoulé entre la collecte et la publication sera raccourci. Des analyses plus complexes seront effectuées sur les données d'enquêtes. Les procédures de couplage d'enregistrements seront améliorées. Les données seront mises à la disposition des utilisateurs sous différentes formes. Les bases de données consultables permettant à l'utilisateur de faire des recherches deviendront plus courantes. L'utilisation de données auxiliaires de toutes sortes, et en particulier des données administratives, augmentera. Les données administratives seront utilisées à la fois comme données auxiliaires et comme estimations directes pour certaines variables. Citro (2014) donne des exemples de variables pour lesquels des données administratives peuvent remplacer les réponses aux questions dans un questionnaire. L'utilisation de données auxiliaires quand l'appariement aux données recueillies est imparfait deviendra un domaine de recherche.

Les méthodes de communication modernes et les médias sociaux sont à l'origine de vastes quantités de données, la plupart d'entre elles générées pour répondre à des objectifs de court terme et mal spécifiés. Le terme « mégadonnées » n'est pas bien défini, mais la plupart d'entre nous seraient d'accord que les données des médias sociaux en font partie. Le rapport de l'AAPOR sur les mégadonnées (2015) contient une excellente analyse des possibilités et des défis associés à ce type de données. Tam et Clarke (2015) et Pfeffermann (2015) discutent de ces questions dans la perspective d'un organisme statistique gouvernemental. Parce qu'ils font partie de la société moderne, les médias sociaux présentent en eux-mêmes un intérêt pour les spécialistes des sciences sociales. Donc, des index et des résumés de ces données sont, et

seront, produits. À titre d'exemple, mentionnons le *Social Media Job Loss Index* de l'Université du Michigan. L'échantillonnage joue un grand rôle dans la création de produits à partir de ces données.

L'un des défis consiste à transformer certains types de mégadonnées en une forme utilisable comme données auxiliaires. Par exemple, Porter, Holan, Wikle et Cressie (2014) se sont servis des fréquences de recherche de mots espagnols fournies par l'outil Google Trends comme covariables fonctionnelles pour estimer, au niveau de l'État, les proportions de personnes parlant l'espagnol, en utilisant les estimations de l'*American Community Survey* comme variables dépendantes dans des modèles de petits domaines.

L'un des avantages souvent mentionnés des enquêtes par sondage comparativement aux recensements est le coût. La structure de coût a évolué avec l'accroissement de la puissance de calcul et semble destinée à continuer d'évoluer. Aux États-Unis, la *National Land Cover Database* est un recensement de la couverture des terres (Han, Yang, Di et Mueller, 2012). En principe, les procédures de classification s'amélioreront, si bien que les données de ce genre serviront plus souvent de données auxiliaires. Les organismes de collecte de données investiront davantage dans la construction de fichiers de données auxiliaires améliorés au niveau de la population, de sorte que certaines données recueillies aujourd'hui sur des échantillons le seront au niveau de la population. Les mêmes types de développement des données se poursuivront pour les statistiques sur les populations et les entreprises.

Nécessairement, notre discussion s'étend peu sur la collecte des données. La façon dont l'évolution de la technologie a modifié les procédures de collecte des données est peut-être plus évidente que le lien entre la technologie et la théorie. Pour les liens avec la théorie, consulter Bellhouse (2000). La collecte des données assistée par ordinateur devient la norme. Il faut s'attendre à ce que l'utilisation de la technologie de géolocalisation s'accroisse. Il est raisonnable de prévoir un plus grand usage d'outils de télédétection et de collecte des données à distance. Par exemple, il serait facile d'intégrer des données physiques recueillies par un appareil comme la montre Apple ou Fitbit dans une étude sur la santé. Des appareils de surveillance plus gros et moins attrayants sont utilisés à l'heure actuelle dans les enquêtes sur l'activité physique (van Remoortel, Giavedoni, Raste, Burtin, Louvaris, Gimeno-Santos, Langer, Glendenning, Hopkinson, Vogiatzis, Peterson, Wilson, Mann, Rabinovich, Puhan, Troosters et PROactive consortium, 2012).

Comme le révèle l'expérience récente, il devient de plus en plus difficile de recueillir des données par interview téléphonique ou par interview sur place. Les enquêtés font face à des activités plus vastes de collecte systématique de données. L'omniprésence des questionnaires sur la satisfaction concernant toute chose, des services médicaux au dentifrice, doit indubitablement avoir une incidence sur la volonté des individus à répondre. Il semble raisonnable de prévoir qu'il sera de plus en plus difficile d'obtenir leur coopération pour les méthodes habituelles de collecte des données. Cette tendance entraînera une étude plus approfondie de la nature des non-répondants et de la non-réponse. De même, des efforts seront déployés en vue d'adapter la collecte des données à l'évolution des méthodes de communication.

Les échantillons non probabilistes ont fait partie de l'activité de sondage tout au long de la période ultérieure à Neyman. En particulier, l'échantillonnage par quotas est couramment utilisé dans la recherche en marketing et ailleurs pour des raisons de coût (Sudman, 1966; 1976). Moser et Stuart (1953) et Stephan et McCarthy (1958) ont fait les premières comparaisons entre l'échantillonnage par quotas et l'échantillonnage probabiliste. Cochran (1977, page 136) affirme [*Traduction*] « La méthode des quotas

produira probablement des échantillons biaisés pour les caractéristiques telles que le revenu, l'éducation et l'occupation, même si elle s'accorde souvent avec les échantillons probabilistes pour les questions d'opinion et d'attitude ». Le recours à des procédures telles que la poststratification et l'estimation par la régression sur des échantillons non probabilistes s'est poursuivi au même rythme que l'utilisation sur des échantillons probabilistes. Le changement de nature de la communication humaine offre des possibilités en ce qui concerne tant les procédures fondées sur un modèle que celles fondées sur les probabilités. En raison des structures de coût, de nouvelles méthodes, telles les procédures axées sur le Web, seront souvent appliquées en premier lieu dans des conditions non probabilistes et à des fins non gouvernementales.

À mesure que les procédures d'appariement s'amélioreront et que la demande de données détaillées augmentera, les procédures de contrôle de la divulgation et les travaux de recherche connexes susciteront davantage d'attention.

L'échantillonnage dans le contexte des enquêtes est une discipline appliquée, dont le fonctionnement est défini par la conjoncture sociale, géographique, culturelle et technologique. Il est difficile de prévoir l'impact qu'auront les changements sociaux et culturels sur cette discipline, même à court terme. Le fait que l'on doive présumer aujourd'hui que presque toutes nos activités publiques et une grande partie de nos activités privées sont susceptibles d'être enregistrées entraînera-t-il une attitude plus détendue de réponse aux questions ? L'amélioration des outils de surveillance fera-t-elle que les enquêtés seront davantage disposés à ce que l'on surveille leurs activités physiques ? Ou bien toute cette surveillance secondaire entraînera-t-elle une réaction négative à l'égard de la collecte systématique de données ? La plus grande disponibilité de résultats fondés sur des données que l'on a recueillies aura-t-elle un effet positif ou négatif sur les efforts de collecte de données ? Quel sera l'impact des divers médias sociaux ?

Il ressort clairement de la présente discussion que des facteurs externes à notre discipline détermineront nos futures activités. Il nous faudra alors adapter la collecte, le traitement, ainsi que la présentation et la diffusion des données.

Remerciements

Nous remercions Graham Kalton pour ses remarques et suggestions ayant mené à des améliorations dans la version originale. Nous remercions les quatre participants, Graham Kalton, Sharon Lohr, Danny Pfeffermann et Chris Skinner, pour les suppléments sur l'histoire, les observations perspicaces sur le présent et les commentaires sur le futur de l'échantillonnage. Nous avonc choisi de ne pas préparer de réponse à ces discussions, car nous en avons apprécié une grande partie et avons trouvé peu de fondement à un désaccord.

Bibliographie

AAPOR Big Data Task Force (2015). *AAPOR Report on Big Data*. https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/images/BigDataTaskForceReport_FINAL_2_12_15_b.pdf.

Anderson, R., Kasper, J. et Frankel, M. (1979). *Total Survey Error: Applications to Improve Health Surveys*. San Francisco, CA: Jossey-Bass.

Andridge, R.H., et Little, R.J. (2009). The use of sample weights in hot deck imputation. *Journal of Official Statistics*, 25, 21-36.

Antal, E., et Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, 106, 534-543.

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Beaumont, J.-F., et Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review*/*Revue Internationale de Statistique*, 80, 127-148.

Bellhouse, D.R. (1988). Systematic sampling. Dans *Handbook of Statistics*, (Éds., P.R. Kreshnaiah et C.R. Rao), Elsevier, 6, 125-145.

Bellhouse, D.R. (2000). Évolution de la théorie de l'échantillonnage d'enquête au XX^{e} siècle et son rapport avec l'informatique. *Techniques d'enquête*, 26, 1, 13-23. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2000001/article/5174-fra.pdf..

Bethlehem, J. (2009). The rise of survey sampling. Discussion Paper (09015), Statistics Netherlands, La Haye.

Bickel, P.J., et Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

Biemer, P.P., et Lyberg, L. (2003). *Introduction to Survey Quality*. New York: John Wiley & Sons, Inc.

Binder, D.A., et Roberts, G.A. (2003). Design-based and model-based methods for estimating model parameters. Dans *Analysis of Survey Data*, (Éds., R.L. Chambers et C.J. Skinner), Wiley, Chichester, Royaume-Uni, 29-48.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin de l'Institut International de Statistique*, 22(1), 1-62.

Bowley, A.L. (1936). The application of sampling to economics and sociological problems. *Journal of the American Statistical Association*, 31, 474-480.

Brewer, K.R.W. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

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

Brick, M.J. (2011). The future of survey sampling. *Public Opinion Quarterly*, 75, 872-888.

Brick, M.J., et Montaquila, J.M. (2009). Non response and weights. Dans *Handbook of Statistics*, (Éds., D. Pfeffermann et C.R. Rao), Elsevier, Amsterdam, 29A, 163-185.

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

Citro, C.F. (2014). Des modes multiples pour les enquêtes à des sources de données multiples pour les estimations. *Techniques d'enquête,* 40, 2, 151-181. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2014002/article/14128-fra.pdf.

Cochran, W.G. (1953). *Sampling Techniques.* New York: John Wiley & Sons, Inc.

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

Dalenius, T. (1974). Ends and Means of Total Survey Design. Report in "Errors in Surveys", Stockholm University.

Deming, E. (1944). On errors in surveys. *American Sociological Review,* 9, 359-369.

Deming, E. (1950). *Some Theory of Sampling*. New York: John Wiley & Sons, Inc.

Deming, W.E. (1953). On a probability mechanism to attain an economic balance between the resultant error of non-response and the bias of non-response. *Journal of the American Statistical Association,* 48, 743-772.

Deming, W.E., et Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics,* 11, 4, 427-444.

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

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

Dippo, C.S., Fay, R.E. et Morgenstein, D.H. (1984). Computing variances from complex samples with replicate weights. *Proceedings of the American Statistical Association, Section on Survey Research Methods,* 489-494.

DuMouchel, W.H., et Duncan, G.J. (1983). Using sample survey weights in multiple regression analysis of stratified samples. *Journal of the American Statistical Association,* 78, 535-543.

Durbin, J. (1958). Sampling theory for estimates based on fewer individuals than the number selected. *Bulletin de l'Institut International de Statistique,* 36, 113-119.

Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika,* 46, 477-480.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics,* 7, 1-26.

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

Francisco, C.A., et Fuller, W.A. (1991). Estimation of quantiles with survey data. *Annals of Statistics,* 19, 454-469.

Fuller, W.A. (1975). Regression analysis sample survey. *Sankhyā, Series C,* 37, 117-132.

Fuller, W.A. (1984). Application de la méthode des moindres carrés et de techniques connexes aux plans de sondage complexes. *Techniques d'enquête*, 10, 1, 107-130. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1984001/article/14352-fra.pdf.

Fuller, W.A. (2009a). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 1-12.

Fuller, W.A. (2009b). *Sampling Statistics*. New York: John Wiley & Sons, Inc.

Fuller, W.A., et An, A.B. (1998). Regression adjustments for nonresponse. *Journal of Indian Society of Agricultural Statistics*, 51, 331-342.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, 17, 269-278.

Godambe, V.P. (1966). A new approach to sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, 28, 310-328.

Gonzalez, M.E. (1973). Use and evaluation of synthetic estimates. *Proceedings of the Social Statistics Section of the American Statistical Association*, 33-36.

Groves, R.M. (1989). *Survey Errors and Survey Costs.* New York: John Wiley & Sons, Inc.

Groves, R.M. (2011). Three eras of survey research. *Public Opinion Quarterly*, 73, 861-871.

Groves, R.M., et Heeringa, S.G. (2006). Responsive designs for household surveys: Tolls for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A*, 169, 439-457.

Groves, R.M., et Lyberg, L. (2010). Total survey error: Past, present and future. *Public Opinion Quarterly*, 74, 849-879.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy*, 5, 361-374.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.

Han, W., Yang, Z., Di, L. et Mueller, R. (2012). CropScape: A Web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support. *Computers and Electronics in Agriculture*, 84, 111-123.

Hansen, M.H., et Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

Hansen, M.H., et Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

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

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

Hansen, M.H., Hurwitz, W.N., Marks, E.S. et Mauldin, W.P. (1951). Response errors in surveys. *Journal of the American Statistical Association,* 46, 147-190.

Hansen, M.H., Hurwitz, W.N., Nisselson, H. et Steinberg, J. (1955). The redesign of the census current population survey. *Journal of the American Statistical Association,* 50, 701-719.

Hartley, H.O. (1959). Analytical studies of survey data. Dans le volume en l'honneur de Corrado Gini, Instituto di Statistica, Rome, 1-32.

Hartley, H.O., et Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika,* 55, 547-557.

Hidiroglou, M.A., Fuller, W.A. et Hickman, R.D. (1976). *SUPER CARP* Statistical Laboratory, Survey Section, Iowa State University, Ames, IA.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association,* 47, 663-685.

Huang, E.T., et Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section,* American Statistical Association, 300-305.

Jagers, P. (1986). Post-stratification against bias in sampling. *International Statistical Review*/*Revue Internationale de Statistique,* 54, 159-167.

Kalton, G. (1983). *Compensating for Missing Survey Data.* Survey Research Center, University of Michigan, Ann Arbor, Michigan.

Kalton, G., et Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics,* A13, 1919-1939.

Kiaer, A. (1897). The representative method of statistical surveys (1976 English translation of the original Norwegian), Oslo, Central Bureau of Statistics of Norway.

Kim, J.K., et Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika,* 91, 559-578.

Kim, J.K., et Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. CRC Press, Boca Raton, FL.

Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition*, 2, 813-830.

Kish, L., et Frankel, M.R. (1970). Balanced repeated replications for standard errors. *Journal of the American Statistical Association*, 65, 1071-1094.

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

Korn, E.L., et Graubard, B.I. (1995). Analysis of large health surveys: Accounting for the sampling designs. *Journal of the Royal Statistical Society, Series A,* 158, 263-295.

Kreuter, F. (2013). *Improving Surveys with Paradata.* Hoboken: Wiley.

Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

Little, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

Little, R.J.A., et Rubin, D.B. (1987, 2014). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc., (Deuxième édition 2014).

Madow, W.G. (1948). On the limiting distribution of estimates based on samples from finite universes. *Annals of Mathematical Statistics*, 19, 535-545.

Madow, W.G., et Madow, L.H. (1944). On the theory of systematic sampling, I. *The Annals of Mathematical Statistics*, 15, 1-24.

Madow, W.G., Nisselson, H. et Olkin, I. (Éds.) (1983). *Incomplete Data in Sample Surveys*, 1, 2, 3, New York: Academic Press.

Mahalanobis, P.C. (1939). A sample survey of the acreage under jute in Bengal. *Sankhyā*, 4, 511-531.

Mahalanobis, P.C. (1944). On large-scale sample surveys. *Philosophical Transactions of the Royal Society of London, Series B*, 231, 329-451.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.

McCarthy, P.J. (1966). Replication: An approach to the analysis of data from complex surveys. *Vital and Health Statisitcs*, Series 2, No. 14, National Center for Health Statistics, Public Health Service, Washington DC.

McCarthy, P.J. (1969). Pseudoreplication: Further evaluation and application of the balanced half-sample technique. *Vital and Health Statistics*, Series 2, No. 31, National Center for Health Statistics, Public Health Service, Washington, DC.

McCarthy, P.J. (1969). Pseudoreplication: Half-samples. *International Statistical Review*/*Revue Internationale de Statistique*, 37, 239-264.

McCarthy, P.J., et Snowden, L.B. (1985). The bootstrap and finite population sampling. *Vital Health Statistics*, 2-95, *Public Health Service Publication*, 85-1369, U.S. Government Printing Office, Washington, D.C.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

Owen, A.B. (1988). Empirical likelihood ration confidence intervals for a single functional. *Biometrika*, 75, 237-249.

Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics. *Journal of Survey Statistics and Methodology*, 3, 425-483.

Pfeffermann, D., et Sverchkov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhyā, Series B,* 61, 166-186.

Porter, A.T., Holan, S.H., Wikle, C.K. et Cressie, N. (2014). Spatial Fay-Herriot model for small area estimation with functional covariates. *Spatial Statistics,* 10, 27-42.

Purcell, N., et Kish, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.

Rao, J.N.K. (2003). *Small Area Estimation.* Hoboken: Wiley.

Rao, J.N.K. (2005). Évaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage. *Techniques d'enquête,* 31, 2, 127-151. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2005002/article/9040-fra.pdf.

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

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 multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics,* 12, 46-60.

Rao, J.N.K., et Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika,* 86, 403-415.

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association,* 83, 231-241.

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

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika,* 57, 377-387.

Rubin, D.B. (1974). Characterizing the estimation of parameters in incomplete data problems. *Journal of the American Statistical Association,* 69, 467-474.

Rubin, D.B. (1976). Inference and missing data. *Biometrika,* 63, 581-590.

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

Schaible, W.L. (Éd.) (1996). *Indirect Estimators in U.S. Federal Programs.* New York: Springer.

Sitter, R.R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association,* 87, 755-765.

Skinner, C.J. (1994). Sample models and weights. Dans *Proceedings of the Survey Research Methods Section,* American Statistical Association, 133-142.

Steinberg, J. (Éd.) (1979). *Synthetic Estimates for Small Areas: Statistical Workshop Papers and Discussion.* NIDA Research Monograph nº 24. U.S. Government Printing Office, Washington, D.C., États-Unis.

Stephan, F., et McCarthy, P.J. (1958). *Sampling Opinions*. New York: John Wiley & Sons, Inc.

Sudman, S. (1966). Probability sampling with quotas. *Journal of the American Statistical Association,* 61, 749-791.

Sudman, S. (1976). *Applied Sampling.* New York: Academic Press.

Sukhatme, P.V. (1954). *Sampling Theory of Surveys with Applications*. Iowa State College Press, Ames.

Tam, S.-M., et Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. *International Statistical Review*/*Revue Internationale de Statistique*, 83, 436-448.

Tchuprow, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, 2, 461-493, 646-683.

Thomsen, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of non-response when analyzing survey data. *Statistisk Tidskrift*, 11, 278-283.

van Remoortel, H., Giavedoni, S., Raste, Y., Burtin, C., Louvaris, Z., Gimeno-Santos, E., Langer, D., Glendenning, A., Hopkinson, N.S., Vogiatzis, I., Peterson, B.T., Wilson, F., Mann, B., Rabinovich, R., Puhan, M.A., Troosters, T. et PROactive consortium (2012). Validity of activity monitors in health and chronic disease: A systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, 9.

Wolter, K.M. (2007). *Introduction to Variance Estimation.* New York: Springer-Verlag.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association,* 47, 635-636.

Yates, F. (1948). Systematic sampling. *Philosophocal Transaction of the Royal Society of London, Series A,* A241, 345-377.

Yates, F. (1949). *Sampling Methods for Censuses and Surveys*. Griffin, Londres.

Techniques d'enquête, décembre 2017
Vol. 43, nº 2, pp. 177-182
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article de Rao et Fuller (2017)

Danny Pfeffermann[1]

Résumé

Cette note de Danny Pfeffermann présente une discussion de l'article « Théorie et méthodologie des enquêtes par sondage : orientations passées, présentes et futures » où J.N.K. Rao et Wayne A. Fuller partagent leur vision quant à l'évolution de la théorie et de la méthodologie des enquêtes par sondage au cours des 100 dernières années.

Mots-clés : Collecte des données; histoire de l'échantillonnage; échantillonnage probabiliste; inférence à partir d'enquêtes.

C'est avec plaisir que je reconnais avoir invité J.N.K. Rao et Wayne Fuller à faire un exposé lors d'une séance spéciale parrainée par l'Association internationale des statisticiens d'enquêtes (AISE) à l'assemblée de l'Institut international de statistique (IIS) qui s'est tenue à Rio de Janeiro en 2015. Toutefois, le mérite revient au professeur Vijay Nair, président de l'IIS à l'époque, qui a instauré ce genre de séance sollicitée pour toutes les sections de l'IIS. Ainsi, quand j'ai su qu'il était possible d'organiser une telle séance, j'ai tout de suite pensé qu'il serait naturel de demander à Rao et Fuller, les deux rois incontestés de la théorie et de la méthodologie modernes des sondages, de donner un exposé sur les orientations passées, présentes et futures de cette discipline, et par bonheur, ils ont accepté d'emblée. Honnêtement, je ne m'attendais pas à leur accord immédiat; en effet, avant de les inviter, j'avais dressé une liste d'arguments en vue de les convaincre, mais comme nous le savons tous, les rois ont des devoirs envers leurs concitoyens. Ce que j'ignorais à l'époque, c'est que, deux ans plus tard, je serais invité à discuter d'un article fondé sur leur exposé, une invitation que j'ai acceptée avec gratitude.

Dans la discussion qui suit, je me concentrerai principalement sur la troisième partie de l'article, l'avenir des sondages, sujet dont je me préoccupe de plus en plus, depuis le tournant pris par ma carrière il y a quatre ans lorsque je suis devenu statisticien national et directeur du bureau central des statistiques d'Israël (ICBS). Naturellement, ma discussion s'appuie sur mon expérience en Israël, mais j'ai de bonnes raisons de croire qu'elle représente en grande partie ce qui se passe également dans d'autres pays.

À la section intitulée « L'avenir », Rao et Fuller (ci-après désignés « RF », comme mon autre roi, Roger Federer) donnent une longue liste d'éléments, qui, selon eux, domineront l'avenir des sondages. Je suis entièrement d'accord avec cette liste, mais avec une réserve. Nombre des éléments de cette liste sont déjà prédominants aujourd'hui. Les budgets sont déjà restreints et les demandes de produits augmentent constamment, non seulement les demandes intérieures faites dans les pays, mais aussi celles des organisations internationales, dont les Nations Unies, l'OCDE, Eurostat, le FMI et la Banque mondiale. Les demandes de statistiques de ces organisations sont souvent similaires et parfois même chevauchantes, mais elles doivent être satisfaites sous différentes formes et à différents moments, couvrant différentes périodes, créant ainsi une surcharge de travail pour les bureaux nationaux de la statistique (BNS). L'utilisation de

1. Danny Pfeffermann, Statisticien national et Directeur du Central Bureau of Statistics, Israël, Hebrew University of Jerusalem, Israël; Southampton Statistical Sciences Research Institute, Royaume-Uni. Courriel : MsDanny@cbs.gov.il.

données administratives fait déjà l'objet de travaux dans les BNS du monde entier. Dans les pays scandinaves, les recensements de population sont fondés uniquement sur des données administratives, et de nombreux autres pays européens ainsi qu'Israël investissent beaucoup de ressources dans la création de bases de données administratives fiables qui remplaceront leurs recensements dans la prochaine génération de recensements, qui seront réalisés autour de 2030. Dans ces pays, les recensements actuels s'appuient déjà fortement sur des données administratives, mais requièrent encore des échantillons pour corriger le sous- ou le surdénombrement. L'usage de données administratives soulève une question juridique importante ayant trait à l'accès à ces données. Les organismes publics et privés refusent simplement de transférer les données. En Israël, la personne occupant le poste de statisticien national est autorisée par la loi à obtenir tout ensemble de données qu'il ou qu'elle juge important pour la production des statistiques officielles, mais certains organismes maintiennent qu'ils se sont engagés à l'égard de leurs clients à ne transférer aucune donnée privée. Je sais que d'autres pays sont confrontés à un conflit similaire. Arriverons-nous à favoriser une culture de partage des données dans l'avenir ? À l'heure actuelle, le défi semble colossal.

Comme il est mentionné dans l'article, la collecte de données par interview téléphonique ou sur place devient de plus en plus difficile, ce qui se traduit par des taux plus élevés de non-réponse. La situation est même pire pour les enquêtes auprès des entreprises, parce que, très souvent, on demande aux mêmes établissements (grands pour la plupart) de participer annuellement à de nombreuses enquêtes (parfois sept et plus). C'est ce qui se passe également en Israël, mais contrairement à d'autres pays, nous arrivons encore à maintenir des taux de réponse raisonnables (autour de 70 %), car la participation à la plupart de nos enquêtes est obligatoire. Diverses approches peuvent être adoptées pour faire face à ce problème. La première consiste à élaborer de nouvelles procédures d'échantillonnage pour alléger le fardeau de réponse. Plusieurs procédures fondées sur les nombres aléatoires permanents sont déjà appliquées, mais des méthodes plus pointues doivent être établies, même si les établissements ou entreprises uniques, ainsi que les gros établissements ou entreprises, n'en bénéficieront qu'à peine, voire jamais. Cela nous pousse à trouver de meilleures méthodes d'imputation au moyen de données antérieures et de données observées pour d'autres unités échantillonnées, quoique l'imputation pour les unités relativement grandes puisse être pratiquement impossible. Un moyen idéal d'alléger le fardeau de réponse consisterait à recueillir les microdonnées telles quelles et les métadonnées pertinentes auprès des entreprises, puis à les traiter au BNS. Cette approche nécessitera évidemment la coopération des entreprises et de plus grandes compétences en science des données dans leurs bureaux. Les chances d'une telle coopération sont-elles raisonnables ? À mon âge, j'ai le droit d'être sceptique, mais en y réfléchissant, moyennant des accords de protection des données appropriés, pourquoi pas ?

À la section 3.3, RF discutent de plusieurs méthodes d'imputation, en énumérant des références clés. Autant que je sache, une hypothèse sous-jacente à toutes ces méthodes est l'accès à des données externes ou internes (antérieures et autres données observées pour les mêmes unités échantillonnées) qui expliquent entièrement les données manquantes. Cependant, selon mon expérience à l'ICBS, cela n'est souvent pas le cas; en outre, dans de nombreuses enquêtes, la non-réponse ne correspond pas à des données manquant au hasard et est donc une non-réponse NMAR (pour *not missing at random*). Le traitement de la non-réponse NMAR est un problème très difficile pour la simple raison que les variables d'intérêt cibles ne sont pas

observées pour les non-répondants, ce qui oblige à formuler des hypothèses au sujet du mécanisme de non-réponse sous forme de modèles statistiques, en n'ayant que des possibilités limitées de les tester. La présente discussion n'est pas censée mettre en relief mes propres travaux de recherche et ceux de mes collègues, mais je voudrais mentionner une approche proposée dans Pfeffermann et Sikov (2011), Feder et Pfeffermann (2015) et Pfeffermann (2017) pour traiter la non-réponse NMAR et qui permet de tester le modèle de réponse, du moins partiellement. L'idée qui sous-tend cette approche consiste à supposer un modèle (paramétrique ou non paramétrique) pour les données de population, et un modèle paramétrique pour le mécanisme de réponse, puis à vérifier si le modèle implicite tient pour les données observées, en utilisant des procédures de vérification de modèle classiques. Voir aussi Riddles, Kim et Im (2016) pour une condition d'identifiabilité importante pour le modèle vérifié pour les données observées, et Sverchkov et Pfeffermann (2017) pour l'application de cette condition dans le contexte de l'estimation sur petits domaines sous non-réponse NMAR.

Deux autres approches pour traiter la non-réponse nécessiteront une étude plus approfondie au cours des années à venir, à savoir l'utilisation des plans de collecte de données adaptatifs et l'acceptation de modes de collecte de données de rechange. L'idée générale qui sous-tend le plan de collecte de données adaptatif est l'utilisation d'information auxiliaire provenant de données de registre et/ou des observations faites par les intervieweurs, afin de personnaliser le plan de collecte de manière à optimiser les taux de réponse et, conséquemment, à réduire le biais de non-réponse. Pour obtenir des précisions et des exemples, consulter l'ouvrage publié récemment par Schouten, Peytchev et Wagner (2017), ainsi que les nombreuses références qu'il contient. RF discutent brièvement des méthodes de collecte des données et de leurs effets sur l'inférence. Nous connaissons tous bien les modes traditionnels ainsi que plus modernes de collecte des données, à savoir l'interview sur place, l'interview par téléphone (téléphone mobile), par la poste (courrier électronique) et, aujourd'hui, par Internet, qui est le mode le moins coûteux et donc, privilégié de collecte des données, même s'il requiert encore l'envoi d'une adresse Internet et d'un mot de passe aux unités échantillonnées. Afin d'accroître les taux de réponse, les organismes d'enquêtes ont tendance à attribuer différents modes de collecte à différentes unités échantillonnées ou, plus généralement, à laisser les unités échantillonnées choisir leur mode de réponse préféré. On parle alors d'enquêtes à mode mixte. Une autre variante des enquêtes à mode mixte, souvent appliquée en pratique, consiste à offrir séquentiellement différents modes de réponse aux unités qui n'ont pas répondu par le mode précédent. Cependant, ces procédures posent un problème d'inférence en raison d'effets de mode éventuels découlant d'un effet de sélection (l'effet de différences entre les caractéristiques des répondants privilégiant différents modes de réponse et, par conséquent, de différences éventuelles dans les valeurs des variables étudiées), et un effet de mesure (l'effet de réponses différentes fournies par un même membre de l'échantillon selon le mode de réponse). Dans Pfeffermann (2015), je passe brièvement en revue les méthodes disponibles pour résoudre ce problème et propose une autre approche, mais les travaux de recherche doivent se poursuivre à ce sujet. Notons qu'un effet de mode (de mesure) peut exister même si un seul mode de réponse est offert.

Un autre aspect de la collecte des données, qui n'est pas évoqué dans l'article, mais qui est très fréquent en pratique, par exemple dans les enquêtes sur la population active, les enquêtes sur la santé et même les recensements modernes, est le recours à la réponse par personne interposée, où un membre du ménage

fournit l'information pour tous les autres membres du ménage. Ce genre d'enquête peut poser un problème éthique en ce sens que l'on ne demande pas aux membres du ménage non interrogés de consentir à ce que l'information à leur sujet soit fournie par le membre qui est interrogé. D'un point de vue statistique, l'utilisation d'enquêtes par personne interposée peut accroître la possibilité d'erreurs de mesure (corrélées). Le membre du ménage interrogé connaît-il l'état de santé de tous les autres membres du ménage, ou sait-il s'ils étaient à la recherche de travail durant la semaine qui a précédé l'interview ? Et s'il se trompe au sujet d'un membre du ménage, ne pourrait-il pas se tromper au sujet d'autres membres ? Ce problème a-t-il été étudié de manière systématique dans la littérature et des solutions ont-elles été proposées ? Pour empirer encore la situation, qu'en est-il de l'interaction entre les erreurs de mesure et la non-réponse ?

La dernière question, mais sans doute la principale lorsqu'on discute de l'avenir des enquêtes par sondage, est celle de l'utilisation éventuelle de mégadonnées en tant qu'alternative aux enquêtes traditionnelles fondées sur des échantillons probabilistes. Ces dernières années, j'ai moi-même donné des exposés sur cette question et mes principales réflexions sont déjà résumées dans Pfeffermann (2015). J'en répéterai certaines ici, dans la perspective d'un statisticien national, préoccupé par la production de statistiques officielles.

RF déclarent correctement que le terme mégadonnées n'est pas bien défini, mais mentionnent les données des médias sociaux comme exemple de ce genre de données. Je pense effectivement que c'est un bon exemple, qui, cependant, illustre aussi les problèmes que pose le traitement de ce genre de données. Ces données sont généralement diverses et non structurées, apparaissent irrégulièrement et peuvent même cesser d'exister. (Néanmoins, tous les types de mégadonnées ne se comportent pas comme cela, et d'autres types pourraient être considérés simplement comme des mégaversions de ce qui est habituellement appelé données administratives.) RF ajoutent que les données provenant des réseaux sociaux intéressent les spécialistes des sciences sociales. De nouveau, je suis d'accord avec eux, mais les BNS devraient-ils publier des estimations des indices sociaux obtenus à partir des réseaux sociaux ? Peut-être que oui, mais quelle population ces indices représenteront-ils ? RF mentionnent aussi que les mégadonnées devraient servir de données auxiliaires. Ils n'entrent pas dans les détails, mais je suppose qu'ils pensent à leur utilisation comme covariables dans l'inférence assistée par modèle ou dépendante d'un modèle, de la même manière que les données administratives. C'est une évidence et de nombreux exemples ont été publiés dans la littérature. Mais la question n'est pas aussi simple qu'elle le paraît, en raison de tous les obstacles informatiques qu'il faudra surmonter avant que les données soient prêtes à l'emploi, y compris le couplage d'enregistrements, s'il faut apparier des mégadonnées et des microdonnées d'enquêtes. Manifestement, il faudra ajuster et tester des modèles appropriés.

Néanmoins, le vrai défi, selon moi, tient à l'usage possible des mégadonnées pour remplacer des enquêtes par sondage classiques. Pour calculer l'IPC, il est indiscutable qu'il est bien plus efficace et moins coûteux d'obtenir les prix de vente (et les quantités) électroniquement, directement auprès des magasins, au lieu d'envoyer des enquêteurs pour faire le relevé des prix. Mais des indices des prix sont souvent demandés pour des sous-groupes de la population, définis selon l'âge, l'origine, etc., et les mégadonnées obtenues électroniquement ne contiennent en général pas d'information démographique. Arriverons-nous à utiliser

l'information des cartes de crédit pour relier les acheteurs aux achats ? Les sociétés de cartes de crédit fourniront-elles l'information requise ? Comment cela se fera-t-il ? Et qu'en est-il des problèmes de couverture des mégadonnées ? Les opinions exprimées sur les réseaux sociaux représentent-elles les opinions du grand public ? Les annonces d'offre d'emploi sur Internet peuvent-elles remplacer les enquêtes auprès des entreprises au sujet des emplois vacants ? Lors d'une conférence qui a eu lieu plus tôt cette année pour célébrer le 80e anniversaire de Jon Rao, le professeur Jae Kim a envisagé trois procédures possibles pour corriger le biais de couverture possible des mégadonnées. À l'assemblée de l'IIS tenue à Marrakech, j'ai proposé une quatrième procédure. Les quatre procédures paraissent raisonnables, mais il est clair que beaucoup d'autres études théoriques et appliquées seront nécessaires avant que l'on puisse effectivement recommander l'usage de n'importe laquelle d'entre elles.

Une des procédures proposées par Kim requiert l'appariement des mégadonnées à un échantillon approprié, afin d'estimer les probabilités d'inclusion dans les mégadonnées. Il s'agit là d'un très bel exemple de combinaison de mégadonnées avec des données d'enquêtes. Di Zio, Zhang et De Waal (2017) discutent d'une autre utilisation de l'échantillonnage classique, pour « assister » la construction et la validation de modèles. À cet égard, une question importante est de savoir comment intégrer les erreurs d'échantillonnage avec les erreurs de modélisation dans l'inférence subséquente. Devrait-on évaluer l'estimateur pour mégadonnées dans une perspective fondée sur un modèle uniquement, quand la construction du modèle fait appel à des données d'échantillon susceptibles de présenter des erreurs d'échantillonnage ? Mise en garde : lorsque l'on combine des mégadonnées et des données d'enquêtes, on devrait vérifier soigneusement la définition des variables qui figurent dans les deux ensembles de données, même si elles semblent mesurer le même phénomène. Le chômage dans un ensemble de mégadonnées pourrait avoir une définition très différente de celle de l'Oganisation internationale du Travail (OIT) adoptée dans les enquêtes sur la population active. Un autre aspect de la combinaison éventuelle de mégadonnées avec des données d'enquêtes, et encore davantage, de l'usage de mégadonnées seulement, est l'élaboration de nouveaux algorithmes d'échantillonnage à appliquer aux mégadonnées. L'échantillonnage des mégadonnées permet de réduire l'espace de stockage, aide à protéger les renseignements personnels et à prévenir la divulgation, et produit des ensembles de données maniables sur lesquels des algorithmes peuvent être exécutés pour ajuster des modèles et produire des estimations. Toutefois, l'échantillonnage aléatoire applicable à des mégadonnées versatiles et dynamiques est manifestement différent de l'échantillonnage pour les populations finies et requiert de nouveaux algorithmes. L'échantillonnage par réseau des mégadonnées remplacera-t-il l'échantillonnage en population finie classique ?

Pour conclure cette brève discussion, je dirai qu'à mon avis, les enquêtes par sondage classiques continueront de jouer un rôle essentiel dans l'avenir prévisible. Cependant, comme le souligne Marker (2017), l'existence de mégadonnées a modifié les attentes en matière d'actualité des données et les BNS devront trouver un moyen de réaliser les enquêtes et les recensements plus rapidement, sinon les utilisateurs s'appuieront sur les mégadonnées disponibles sans comprendre ce qu'ils y perdent.

Bibliographie

Di Zio, M., Zhang, L.C. et De Waal, T. (2017). Statistical methods for combining multiple sources of administrative and survey data. *The Survey Statistician*, 17-26.

Feder, M., et Pfeffermann, D. (2015). Statistical inference under non-ignorable sampling and nonresponse-an empirical likelihood approach. Southampton Statistical Sciences Research Institute, University of Southampton, Royaume-Uni. http://eprints.soton.ac.uk/id/eprint/378245.

Marker, D. (2017). How have National Statistical Institutes improved quality in the last 25 years? *Statistical Journal of the IAOS*, 1, 1-11.

Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics. *The Journal of Survey Statistics and Methodology* (*JSSAM*), 3, 425-483.

Pfeffermann, D. (2017). Bayes-based non-bayesian inference on finite populations from non-representative samples. A unified approach. *Calcutta Statistical Association (CSA) Bulletin*, 69, 35-63.

Pfeffermann, D., et Sikov, A. (2011). Estimation and imputation under non-ignorable non-response with missing covariate information. *Journal of Official Statistics*, 27, 181-209.

Riddles, K.M., Kim, J.K. et Im, J. (2016). A propensity-scoreadjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology*, 4, 215-245.

Schouten, B., Peytchev, A. et Wagner, J. (2017). *Adaptive Survey Design*. Chapman&Hall/CRC Statistics in the Social and Behavioral Sciences.

Sverchkov, M., et Pfeffermann, D. (2017). Small area estimation under informative sampling and not missing at random nonresponse. (À l'étude).

Techniques d'enquête, décembre 2017
Vol. 43, nº 2, pp. 183-188
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article de Rao et Fuller (2017)

Graham Kalton[1]

Résumé

Cette note de Graham Kalton présente une discussion de l'article « Théorie et méthodologie des enquêtes par sondage : orientations passées, présentes et futures » où J.N.K. Rao et Wayne A. Fuller partagent leur vision quant à l'évolution de la théorie et de la méthodologie des enquêtes par sondage au cours des 100 dernières années.

Mots-clés : Collecte des données; histoire de l'échantillonnage; échantillonnage probabiliste; inférence à partir d'enquêtes.

Le bref article rédigé par Jon Rao et Wayne Fuller est d'une très grande portée. Sa lecture m'a incité à réfléchir à ma propre expérience quant à l'histoire de la recherche par sondage au cours des 50 dernières années ou plus, principalement en tant que spécialiste de la recherche appliquée en enquêtes sociales au Royaume-Uni et aux États-Unis. Dans l'ensemble, d'étonnants progrès ont eu lieu dans toutes les disciplines de la recherche par sondage au cours de ma vie professionnelle, y compris dans les domaines de l'échantillonnage et des méthodes de collecte de données. Jon et Wayne ont, naturellement, contribué considérablement à ces progrès. Ma discussion portera sur deux grands volets des changements qui ont eu lieu.

Évolution du rôle des modèles dans l'inférence par échantillonnage

Au début de ma carrière, le mode d'inférence fondé sur le plan de sondage établi par Neyman primait, mais sa prédominance a diminué au fil du temps, surtout récemment. L'inférence fondée sur le plan de sondage doit son intérêt au fait que la convergence des estimateurs des paramètres de population fondés sur un échantillon probabiliste tiré d'une population finie ne dépend pas de modèles, contrairement à l'estimation dépendante d'un modèle où l'inférence est fonction de la validité des hypothèses de modélisation. Dès le début, l'échantillonnage probabiliste et l'inférence fondée sur le plan de sondage étaient assistés par modèle, par exemple dans le cas de la répartition de l'échantillon sous échantillonnage stratifié et estimation par la régression. Néanmoins, bien que la formulation incorrecte de ces modèles de travail ait une incidence sur la précision des estimateurs de l'enquête, la convergence des estimateurs demeure intacte.

Les conditions à satisfaire pour une inférence purement fondée sur le plan de sondage sont que chaque unité de la population cible possède une probabilité de sélection non nulle connue (ou du moins une probabilité de sélection *relative* connue) et que des données d'enquête valides soient obtenues pour chaque unité échantillonnée. Dans les enquêtes sociales, ces conditions ne sont presque jamais entièrement satisfaites en pratique, parce que des données manquent inévitablement en raison de problèmes de

1. Graham Kalton est un vice-président principal à Westat, 1600 Research Blvd., Rockville, MD, 20850, États-Unis. Courriel : grahamkalton@westat.com.

non-couverture et de non-réponse (totale ainsi que partielle). Les modèles jouent un rôle essentiel dans le traitement des données manquantes, que ce soit par des méthodes de pondération et d'imputation ou en ignorant le problème, c'est-à-dire en employant implicitement un modèle qui traite les données manquantes comme des données manquant entièrement au hasard (MCAR pour *missing completely at random*).

Si je me souviens bien de la situation au début de ma carrière, l'usage de modèles pour produire des estimations au moyen de données d'enquêtes n'était guère reconnu et, en fait, l'opposition au recours à des modèles pour l'inférence à partir d'enquêtes était forte. De nombreux chercheurs ont résisté avec acharnement à l'imputation quand celle-ci a commencé à se répandre autour des années 1980, sous prétexte qu'elle impliquait des « données fabriquées »; les analystes procédaient plutôt à une analyse de cas complète, en émettant implicitement l'hypothèse MCAR. À cette époque, les taux de non-réponse étaient faibles, de sorte que l'appui sur des hypothèses de modélisation n'était pas important; quand de simples ajustements de la pondération pour tenir compte de la non-réponse étaient effectués, peu d'attention leur était accordée. Étant donné la hausse des taux de non-réponse ces dernières années, la situation a changé considérablement et les estimations d'après des données d'enquêtes dépendent maintenant fortement de modèles. Par conséquent, une foule de travaux de recherche ont porté sur les méthodes de modélisation des données manquantes, comme l'ont souligné Rao et Fuller.

Une autre façon dont les modèles jouent un rôle dans la pratique des sondages découle de l'utilisation de méthodes d'échantillonnage non probabiliste ou de méthodes d'échantillonnage qui ne répondent pas strictement à l'exigence que les probabilités de sélection soient connues. Les considérations de coût jouent un rôle important dans l'élaboration de l'échantillon, et elles peuvent mener au choix d'un plan d'échantillonnage avec probabilités de sélection inconnues qui sont ensuite calculées approximativement en se basant sur des hypothèses de modélisation. Un exemple bien connu est l'échantillonnage par quota, une méthode d'échantillonnage non probabiliste qui a été utilisée à grande échelle dans les études de marché et qui a été appliquée, au départ, dans un certain nombre d'études sociales (voir Stephan et McCarthy, 1958). Sudman (1966) décrit un plan de répartition par quota selon lequel des intervieweurs employaient des contrôles de quota pour créer des groupes de personnes que l'on supposait posséder les mêmes probabilités de sélection. Dans un secteur donné, l'intervieweur était alors libre de sélectionner n'importe quel sujet, à la condition que l'échantillon résultant satisfasse les contrôles de quota. Un autre exemple est celui de l'échantillonnage par marche aléatoire (*random walk*) qui permet d'éviter le coût du listage des logements dans un secteur échantillonné; selon cette méthode, l'intervieweur doit commencer à un emplacement spécifié et suivre une route donnée. Bauer (2016) montre comment cette méthode ne fournit pas l'échantillon avec probabilités égales qu'elle est censée produire. Les coûts de listage sont également évités dans le cas des enquêtes d'évaluation rapide du Programme élargi de vaccination (PEV) de l'Organisation mondiale de la santé lesquelles sont conçues pour estimer la couverture de la vaccination des enfants dans une région donnée. À l'intérieur d'une grappe échantillonnée (par exemple, un village), l'intervieweur commence par un ménage « sélectionné au hasard », puis passe au ménage suivant le plus proche, et ainsi de suite, en série, jusqu'à ce que la taille d'échantillon spécifiée soit atteinte (souvent sept enfants admissibles). En plus de supposer que les enfants sont échantillonnés au hasard à l'intérieur de la grappe échantillonnée, la méthode s'appuie sur l'hypothèse incorrecte que les grappes sont échantillonnées avec des probabilités exactement

proportionnelles au nombre d'enfants admissibles dans la grappe au moment de l'enquête. Bennett (1993) et d'autres ont proposé des modifications pour éviter les biais découlant du modèle hypothétique de probabilités de sélection égales pour cette méthode d'échantillonnage du PEV utilisée à très grande échelle.

Ces dernières années, la demande d'enquêtes pour étudier des populations rares, dont certaines sont définies en fonction de caractéristiques sensibles (y compris des comportements illégaux), a augmenté considérablement. Voir, par exemple, Tourangeau, Edwards, Johnson, Wolter et Bates (2014). Le recours à des méthodes d'échantillonnage non probabilistes est nécessaire dans les situations où l'échantillonnage probabiliste est jugé infaisable. Cependant, ces méthodes n'offrent pas la sécurité de l'inférence fondée sur le plan de sondage. Les plans de sondage non probabiliste largement utilisés pour l'étude de populations rares difficiles à échantillonner comprennent l'échantillonnage boule de neige, l'échantillonnage dirigé par les répondants, l'échantillonnage de lieux (fondé sur les lieux de rencontre), et les enquêtes en ligne.

Les enquêtes en ligne sont attrayantes parce qu'elles permettent d'obtenir des réponses aux enquêtes à peu de frais et presque instantanément. Ces enquêtes prennent de nombreuses formes, dont les enquêtes en ligne auto-sélectionnées, les panels volontaires d'utilisateurs d'Internet et les panels en ligne fondés sur des échantillons probabilistes (Couper, 2000). Les tailles d'échantillon des enquêtes en ligne auto-sélectionnées et des panels de volontaires sont souvent très grandes, mais la principale préoccupation tient aux biais possibles dans les estimations d'enquête. Comme l'indique le tristement célèbre sondage de 1936 de la revue *Literacy Digest*, les grands échantillons ne protègent pas contre le biais dans les estimations d'enquête. Le sondage en question était une enquête par la poste dont le questionnaire a été envoyé à environ 10 millions de personnes sélectionnées principalement à partir d'annuaires téléphoniques et de listes d'immatriculations de véhicules, et environ 2 millions de personnes ont répondu. Le sondage prédisait une victoire écrasante d'Alf Landon à l'élection présidentielle de 1936 aux États-Unis, alors que Franklin Roosevelt l'a remportée largement (voir Converse, 1987, pages 456-457 pour des références tentant d'expliquer l'échec du sondage). Il reste à déterminer si les méthodes modernes d'ajustement des pondérations appliquées à une collecte de données en ligne non probabiliste à grande échelle permettent de contourner les problèmes du sondage du *Literary Digest* et, élément plus critique, dans quelles conditions on peut se fier en toute confiance à la qualité des estimations dépendantes d'un modèle qui sont produites. Même quand un panel en ligne est recruté en utilisant un plan d'échantillonnage probabiliste, le taux de réponse global généralement très faible met gravement en question la sécurité de l'inférence fondée sur le plan de sondage.

Après des années d'opposition, les méthodes d'estimation sur petits domaines dépendantes d'un modèle sont aujourd'hui généralement acceptées, comme le font remarquer Rao et Fuller qui ont beaucoup contribué à la littérature sur ce sujet. Cette acceptation est attribuable au fait que la grande demande d'estimations sur petits domaines manifestée par les décideurs et d'autres ne peut pas être satisfaite par des méthodes fondées sur le plan de sondage avec des tailles d'échantillon d'un coût abordable. L'estimation sur petits domaines débute certes par la collecte de données auprès d'un échantillon probabiliste, mais ensuite elle « emprunte de l'information » à des modèles qui utilisent des données administratives, des données de recensements antérieurs, et d'autres données disponibles au niveau des petits domaines. Les modèles de petits domaines sont construits prudemment et évalués dans la mesure du possible, mais les estimations sur petits domaines résultantes dépendent néanmoins du modèle.

En résumé, se fier entièrement à l'inférence fondée sur le plan de sondage n'est pas raisonnable de nos jours pour diverses raisons. Une plus grande attention doit être accordée aux moyens de communiquer l'incertitude au sujet des estimations produites à partir de données hybrides contenant des composantes fondées sur le plan de sondage et d'autres dépendantes de modèles, en tenant compte des niveaux plausibles d'erreur de spécification du modèle.

Évolution des capacités informatiques des dernières décennies

Les progrès concernant les capacités informatiques au cours des dernières décennies ont eu une influence importante sur tous les aspects de la recherche par sondage. Au début de ma carrière, l'analyse des données d'enquêtes se faisait au moyen de cartes perforées sur des compteuses-trieuses et autres appareils de ce genre. La totalisation était presque la seule forme d'analyse. Les erreurs-types reflétant les plans d'échantillonnage complexes étaient rarement calculées; on appliquait plutôt de simples règles empiriques pour modifier les erreurs-types sous échantillonnage aléatoire simple. Dans son ouvrage intitulé *Sample Design in Business Research*, Deming (1960) recommandait que les échantillons soient conçus en 10 répliques pour faciliter l'estimation de la variance, et en outre, il proposait que l'erreur-type d'une estimation soit obtenue par la simple division par 10 de la différence entre la plus grande et la plus petite répliques. Dans *Survey Sampling*, Kish (1965) insistait sur la simplicité des calculs de variance fondés sur un plan avec échantillons appariés selon lequel deux unités primaires d'échantillonnage sont sélectionnées dans chaque strate, et il a exposé la façon d'effectuer ces calculs à la main. Aujourd'hui, les estimations de variance pour des statistiques simples ou complexes fondées sur des plans d'échantillonnage complexes sont calculées facilement à l'aide de progiciels utilisant des techniques telles que les répliques répétées équilibrées, les répliques répétées jackknife, le bootstrap et la linéarisation. En outre, pour les méthodes de rééchantillonnage, il est simple de recalculer des ajustements de pondération, même complexes, pour chaque réplique, ce qui permet d'intégrer dans les estimations de variance la variabilité associée à ces ajustements.

L'impact des ordinateurs sur la statistique d'enquête ne se limite pas à l'estimation de la variance. Les ordinateurs permettent aussi d'appliquer des plans de sondage plus complexes, ce qui a mené à la multiplication des méthodes complexes d'analyse (comme l'expliquent Rao et Fuller). Considérons le cas d'une stratification poussée à titre d'exemple d'un plan plus complexe. Goodman et Kish (1950) décrivent une méthode de stratification poussée, appelée sélection contrôlée, pouvant être effectuée par de simples calculs. Les méthodes d'échantillonnage équilibré plus récentes, à savoir l'échantillonnage par la méthode du cube et la méthode connexe de l'échantillonnage réjectif mentionnées par Rao et Fuller, sont de loin plus compliquées à appliquer.

Les ordinateurs ont également eu des effets importants sur d'autres aspects du processus d'enquête. Il y a 50 ans, les données d'enquêtes étaient recueillies au moyen d'interviews utilisant papier et crayon (IPC) ou de questionnaires envoyés par la poste. La méthode IPC a été en grande partie remplacée par la collecte de données assistée par ordinateur (Couper, Baker, Bethlehem, Clark, Martin, Nicholls et O'Reilly, 1998). Les interviews sur place assistées par ordinateur (IPAO) sont réalisées en utilisant des ordinateurs portables ou, plus fréquemment aujourd'hui, des tablettes. Certaines données – surtout les données sensibles – peuvent être recueillies par auto-interviews assistées par ordinateur avec interface audio (audio-AIAO).

Dans le cas d'une collecte de données par IPAO, la totalité ou des parties d'une interview peuvent être enregistrées (interview enregistrée assistée par ordinateur – IEAO); l'IEAO peut être utile pour les prétests et pour vérifier la performance des intervieweurs tout au long de la période de collecte des données. Les ordinateurs permettent aussi de recueillir les emplacements GPS des interviews, ce qui rend possible la surveillance des falsifications des intervieweurs et fournit des données pour diverses analyses fondées sur la localisation. Plus récemment, la collecte de données en ligne est devenue un mode de collecte rentable intéressant, mais à une époque où les taux de réponse s'effondrent, elle doit souvent être complétée par des interviews sur place ou d'autres méthodes. Les enquêtes à mode de collecte mixte deviennent de plus en plus populaires, et leur utilisation augmentera vraisemblablement dans l'avenir (en accordant l'attention nécessaire aux différences possibles de réponse à certaines questions selon le mode de collecte). Voir Dillman (2017) pour une revue de problèmes liés au fait d'inciter initialement les participants aux enquêtes à mode mixte à répondre sur le Web.

Conclusion

L'histoire de la recherche par sondage est caractérisée par une augmentation rapide du nombre et de la complexité des demandes de données d'enquêtes. Ces demandes ont abouti à l'utilisation de fichiers à grande diffusion (FGD) pour les analyses individuelles et à des préoccupations concernant la protection des données confidentielles des répondants. La diffusion de nombreux tableaux et d'autres analyses suscite des préoccupations comparables. Pour répondre à ces préoccupations, des méthodes de contrôle de la divulgation statistique sont élaborées pour les FGD (par exemple, voir Hundepool, Domingo-Ferrer, Franconi, Giessing, Schulte Nordholt, Spicer et de Wolf, 2012), des fichiers de données à usage restreint sont utilisés et des enclaves statistiques sont créées pour permettre aux analystes d'effectuer leurs analyses sous supervision. En outre, des archives ont été établies pour stocker et gérer les ensembles de données d'enquêtes.

La croissance rapide de la demande de données d'enquêtes constatée au cours des dernières décennies est une tendance qui se poursuivra vraisemblablement, ce qui semble réserver un bel avenir à la recherche par sondage. Toutefois, comme l'a dit Paul Valery, « Le problème avec notre temps, c'est que l'avenir n'est plus ce qu'il était », remarque qui semble particulièrement pertinente dans le cas de la recherche par sondage en ce moment. D'aucuns voient dans les mégadonnées et les dossiers administratifs de sérieux concurrents pour les enquêtes, mais je ne suis pas aussi convaincu. Ces deux types de données peuvent satisfaire certains besoins, mais la nature multivariée des enquêtes et, souvent, la nécessité de recueillir certains éléments de données qui ne peuvent être obtenus qu'auprès des répondants (par exemple, opinions, niveau de littératie des adultes, dépenses des ménages, diabète) signifient que les enquêtes continueront de jouer un rôle important. Les données administratives peuvent produire des estimations pour certaines statistiques officielles (surtout les statistiques économiques), particulièrement quand il est permis de fusionner des ensembles de fichiers administratifs. Cependant, selon moi, dans les enquêtes sociales officielles, les dossiers administratifs seront utilisés principalement comme un supplément susceptible de réduire le fardeau de réponse en remplaçant les questions de l'enquête par des données enregistrées et pouvant fournir des

données longitudinales pour la période qui précède ainsi que celle qui suit la collecte des données de l'enquête. Naturellement, la qualité des données enregistrées et des couplages d'enregistrements doit être évaluée. À mon avis, la menace la plus importante qui pèse sur la recherche par sondage tient à la réticence croissante des membres du public à répondre aux enquêtes. Jusqu'à présent, aucune bonne solution contre cette menace n'a été découverte.

Bibliographie

Bauer, J.J. (2016). Biases in random route surveys. *Journal of Survey Statistics and Methodology*, 4, 263-287.

Bennett, S. (1993). Cluster sampling to assess immunization: A critical appraisal. *Bulletin of the International Statistical Institute*, 49e Session, 55(2), 21-35.

Converse, J.M. (1987). *Survey Research in the United States: Roots and Emergence 1890-1960*. Berkeley: University of California Press.

Couper, M.P. (2000). Web surveys. *Public Opinion Quarterly*, 64, 464-494.

Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nicholls, W.L. et O'Reilly, J.M. (Éds.) (1998). *Computer Assisted Survey Information Collection.* New York: John Wiley & Sons, Inc.

Deming, W.E. (1960). *Sample Design in Business Research*. New York: John Wiley & Sons, Inc.

Dillman, D.A. (2017). Inciter les participants aux enquêtes à mode mixte à répondre sur le Web : les promesses et les défis. *Techniques d'enquête*, 43, 1, 3-34. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2017001/article/14836-fra.pdf.

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

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K. et de Wolf, P.-P. (2012). *Statistical Disclosure Control*. Chichester, Royaume-Uni: Wiley.

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

Stephan, F.F., et McCarthy, P.J. (1958). *Sampling Opinions*. New York: John Wiley & Sons, Inc.

Sudman, S. (1966). Probability sampling with quotas. *Journal of the American Statistical Association*, 61, 749-771.

Tourangeau, R., Edwards, B., Johnson, T.P., Wolter, K.M. et Bates, N. (Éds.) (2014). *Hard-to-survey populations*. Cambridge, Royaume-Uni: Cambridge University Press.

Techniques d'enquête, décembre 2017
Vol. 43, nº 2, pp. 189-195
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article de Rao et Fuller (2017)

Sharon L. Lohr[1]

Résumé

Cette note de Sharon L. Lohr présente une discussion de l'article « Théorie et méthodologie des enquêtes par sondage : orientations passées, présentes et futures » où J.N.K. Rao et Wayne A. Fuller partagent leur vision quant à l'évolution de la théorie et de la méthodologie des enquêtes par sondage au cours des 100 dernières années.

Mots-clés : Collecte des données; histoire de l'échantillonnage; échantillonnage probabiliste; inférence à partir d'enquêtes.

Rao et Fuller méritent des remerciements pour leur tour d'horizon succinct d'un domaine auquel ils ont tous deux tellement contribué. Ce n'est pas une mince affaire de résumer l'histoire de l'échantillonnage probabiliste et de dégager les futures orientations en 18 pages!

Il est toujours risqué de prédire l'avenir. Cependant, faire l'historique des sondages nous permet de voir comment les pionniers du domaine ont résolu les défis qui se présentaient à leur époque et quel est le lien entre ces défis et leurs solutions et les problèmes qui se posent aujourd'hui.

Pour commencer, comparons les avantages et les inconvénients de l'échantillonnage probabiliste aujourd'hui aux avantages et inconvénients qui étaient perçus au milieu du XX^{e} siècle, quand l'usage d'échantillons probabilistes commençait à se généraliser. Les listes qui suivent sont tirées de Parten (1950, chapitre 4); les premiers traités d'échantillonnage publiés par Deming (1950) et par Hansen, Hurwitz et Madow (1953a) donnent des descriptions similaires. Selon Parten, les avantages des échantillons probabilistes, comparativement à la réalisation d'un recensement de la population ou au tirage d'un échantillon de commodité, sont de quatre types :

A1. Les estimations peuvent être obtenues plus rapidement à partir d'un échantillon que d'un recensement. Moins d'interviews sont nécessaires et, durant les années 1950, le traitement et la totalisation des données pouvaient être effectués plus rapidement pour un petit ensemble de données que pour un grand. Parten a écrit : [*Traduction*] « Cette économie de temps est un avantage particulièrement important dans les études de notre société moderne dynamique. Les conditions évoluent si rapidement qu'à moins de concevoir des méthodes offrant un raccourci pour mesurer les situations sociales, la mesure n'est déjà plus à jour avant que l'enquête ou le sondage soit achevé. » (Parten, 1950, page 109).

A2. Les estimations à partir d'un échantillon sont moins coûteuses qu'un recensement, parce que moins d'interviews sont nécessaires. Cela se traduit par des coûts plus faibles en dotation et en formation du personnel de terrain.

A3. L'enquête peut être taillée sur mesure en fonction des estimations d'intérêt. L'échantillonneur peut être plus prudent dans la collecte des données, en posant exactement les questions souhaitées

1. Sharon L. Lohr, Arizona State University, Tempe, AZ. Courriel : sharon.lohr@asu.edu.

et en prenant des mesures pour réduire au minimum le biais dû à la non-réponse et à d'autres sources. Au contraire, un recensement peut ne contenir que quelques questions et offrir peu de possibilités de suivi.

A4. L'échantillonnage probabiliste permet à l'échantillonneur de concevoir l'enquête de manière à obtenir une précision souhaitée et, plus tard, à communiquer la précision réalisée, sans s'appuyer sur des hypothèses de modélisation. Deming (1950, page 10) insistait sur le fait que non seulement les erreurs d'échantillonnage peuvent être calculées à partir d'échantillons probabilistes, mais aussi que les biais de sélection, de non-réponse et d'estimation sont pour ainsi dire éliminés ou contenus dans des limites connues. Hansen et coll. (1953a, page 10) déclaraient : [*Traduction*] « Les méthodes d'échantillonnage probabiliste permettent d'éviter entièrement de dépendre du jugement pour déterminer la précision. Dans ces circonstances, et avec des échantillons raisonnablement grands, la précision des résultats obtenus à partir de l'échantillon peut être mesurée d'après l'échantillon proprement dit. »

Parten a également examiné les inconvénients du tirage d'un échantillon au lieu de la réalisation d'un recensement :

D1. Il est difficile de bien faire l'échantillonnage et d'obtenir des échantillons représentatifs. Le fait de ne pas suivre correctement le protocole d'échantillonnage peut introduire des erreurs dans les estimations, et les résultats peuvent être trompeurs si un échantillon est mal conçu ou mal analysé. En outre, étant donné la pénurie de statisticiens d'enquête chevronnés, il est difficile pour la personne qui se propose de réaliser un sondage d'obtenir une aide technique.

D2. La petite taille de l'échantillon limite l'information que l'on peut obtenir. Dans un échantillon, les observations pour les sous-populations rares sont peu nombreuses. De plus, le nombre de tableaux croisés est limité, parce que le nombre de cas est trop faible pour certaines sous-classifications d'intérêt.

Qu'en est-il de nos jours des avantages et des inconvénients des échantillons probabilistes énumérés par Parten ? Les inconvénients existent toujours. En particulier, la demande d'information plus détaillée, plus à jour et plus complète augmente chaque année (D2). Néanmoins, alors que les enquêtes présentent toujours l'avantage (A3), à savoir qu'elles peuvent être taillées sur mesure pour répondre aux questions d'intérêt, les avantages (A1) et (A2) ont diminué. Durant les années 1950, la collecte de *n'importe quel* type de données était souvent onéreuse. Même les données provenant d'un petit échantillon de commodité pouvaient nécessiter des interviews coûteuses ou une transcription des documents papier demandant beaucoup de travail manuel. En revanche, aujourd'hui, d'énormes échantillons de commodité peuvent souvent être obtenus à un coût beaucoup plus faible, tandis que des échantillons probabilistes, comme ceux de l'*American Community Survey* ou de la *National Crime Victimization Survey*, deviennent de plus en plus coûteux à mesure que les taux de réponse diminuent. L'information provenant d'échantillons de commodité peut également être disponible plus rapidement que les données issues d'un échantillon probabiliste de haute qualité, pour lequel la pondération des données, le calcul des estimations et les vérifications de la qualité prennent des mois.

L'avantage (A4) consistant à pouvoir planifier une précision souhaitée a également diminué. La plupart des grandes enquêtes utilisent encore des méthodes fondées sur le plan de sondage pour communiquer la précision de l'enquête. Toutefois, la marge d'erreur fondée sur le plan de sondage qui est communiquée ne comprend généralement que l'erreur d'échantillonnage et s'appuie sur l'hypothèse implicite que la pondération de l'échantillon a éliminé le biais de non-réponse dans les estimations. À mesure que diminuent les taux de réponse, on se fie de plus en plus au jugement, en utilisant des hypothèses de modélisation, pour déterminer la précision.

Le contexte de la collecte des données est donc différent de ce qu'il était dans les années 1930, les années 1940 et les années 1950, durant lesquelles les techniques d'échantillonnage probabiliste ont été élaborées et mises en œuvre. À l'époque aux États-Unis, l'échantillonnage probabiliste répondait à un besoin urgent de fournir des renseignements plus rapidement et à moindre coût sur l'agriculture, l'activité commerciale, la production manufacturière, les caractéristiques de la main-d'œuvre et d'autres indicateurs sociaux et économiques. Les pionniers des méthodes de sondage ont révolutionné la collecte des données durant cette période. Duncan et Shelton (1978) soutenaient que cette révolution était rendue possible par des progrès parallèles dans les domaines de la théorie statistique, des comptes nationaux du revenu et de la production, de la capacité de calcul et de l'organisation du système statistique.

Bien que les sources de données, l'infrastructure, la technologie et les méthodes disponibles aient changé, le principal problème qui se pose à nouveau aujourd'hui est le même qu'en 1950 : quel est le meilleur moyen de recueillir des données et de faire des inférences à partir de celles-ci pour faciliter la réponse aux questions stratégiques et de recherche ? Si le cadre actuel des échantillons probabilistes n'existait pas, et qu'on nous demandait de construire un système de collecte de données, que ferions-nous ? Pour de nombreux problèmes, nous voudrions construire un système de collecte des données modulaire, pouvant s'adapter à de nouvelles sources de données et de nouvelles technologies de collecte des données. La plupart des méthodes que Rao et Fuller ont passées en revue seraient utiles pour ce système, mais une nouvelle infrastructure et de nouvelles méthodes – et peut-être une autre révolution – sont nécessaires.

À titre d'exemple, considérons le *National Automotive Sampling System* des États-Unis (NHTSA, 2017a). Le système est composé de deux enquêtes. La première porte sur un échantillon probabiliste stratifié à plusieurs degrés de 50 000 à 60 000 rapports d'accident par la police (RAP) tiré de l'univers d'environ 6 millions de RAP annuels, où les RAP concernant des accidents graves sont échantillonnés avec des probabilités plus élevées que ceux concernant des collisions ne comprenant que des dommages matériels mineurs. Les éléments de données extraits des RAP échantillonnés sont codés dans la base de données électronique; aucune information externe aux RAP n'est obtenue. La deuxième enquête est réalisée sur un plus petit échantillon probabiliste d'environ 5 000 RAP avec une collecte des données beaucoup plus intensive, dans le cadre de laquelle des enquêteurs ayant reçu une formation spéciale visitent le lieu de l'accident, inspectent le ou les véhicules accidentés, obtiennent la permission de consulter les dossiers médicaux, interviewent les témoins, et obtiennent d'autres renseignements au sujet de l'accident. Les données de ces deux enquêtes sont utilisées pour étudier les tendances temporelles des accidents automobiles et les effets des caractéristiques des véhicules sur la sécurité routière (voir, par exemple, NHTSA, 2017b), ainsi que pour produire des milliers de documents de recherche.

Mais supposons qu'on nous demande de concevoir ce système de collecte des données en reprenant tout à zéro. Je tiens à souligner que ces suggestions émanent purement de mon imagination et n'ont aucun lien avec quelque projet que ce soit concernant les enquêtes, qui, en raison de considérations pratiques et contraintes budgétaires courantes, doivent avoir une structure d'échantillonnage à plusieurs degrés. Si le système actuel n'existait pas, ne serait-il pas souhaitable de concevoir la première enquête comme un recensement des RAP au lieu du tirage d'un échantillon ? Cette tâche ne serait pas forcément facile. Hetzel (1997) a décrit le processus long et laborieux d'établissement du système américain de statistiques de l'État civil, qui requiert la coopération des organismes d'État et des gouvernements locaux, des procédures de collecte des données uniformes, et des études approfondies pour valider l'exactitude et la couverture des dossiers. La réalisation d'un recensement des RAP nécessiterait pareillement d'énormes investissements initiaux pour élaborer l'infrastructure et obtenir la collaboration des États et des services de police. Après cet investissement, toutefois, la collecte des données serait établie et les RAP pourraient être transmis électroniquement à mesure qu'ils sont recueillis ou mis à jour.

Les avantages de l'obtention d'un recensement de RAP plutôt qu'un échantillon seraient nombreux. Ainsi, les statistiques seraient disponibles beaucoup plus rapidement, puisqu'il ne faudrait pas attendre jusqu'à la fin de l'année de collecte des données pour pondérer et publier les données d'enquête : les statistiques pourraient être mises à jour à mesure que les données sont recueillies. Néanmoins, l'avantage le plus important d'un recensement tiendrait à l'information supplémentaire sur les sous-populations. Cela permettrait de mieux suivre les données pour déceler les risques d'accident possibles. Dans un échantillon de 50 000 unités, un véhicule correspondant à une combinaison particulière de marque/modèle/année pourrait n'être représenté que par une poignée d'observations (si tant est qu'il y en ait); la taille de l'échantillon serait beaucoup plus grande dans le cas du recensement. Dans certaines enquêtes, comme le soulignent Rao et Fuller, des méthodes d'estimation sur petits domaines peuvent être utilisées pour modéliser les résultats pour des sous-populations dont les tailles d'échantillon sont petites. Cependant, dans le cas des données sur les accidents, il est fréquent que l'on s'intéresse à une sous-population soupçonnée de représenter un point aberrant, parce que l'on suspecte que le nombre d'accidents pour une certaine marque d'automobile ou une certaine caractéristique de véhicule est plus élevé que ne le prédirait un modèle. Ces points aberrants ne peuvent pas être décelés par un modèle pour petits domaines. Le seul moyen d'obtenir l'information sur les sous-populations éventuellement aberrantes consiste à recueillir plus de données à leur sujet.

Mais si l'on procède à un recensement des RAP, quand les méthodes de recherche par sondage entreront-elles en jeu ? Il y aurait inévitablement des données manquantes qu'il faudrait examiner et modéliser, et un plan d'échantillonnage à deux phases pourrait être utilisé pour obtenir des renseignements auprès des États ou des services de police non répondants. Cependant, le principal problème de conception de l'enquête serait double : premièrement, l'échantillonnage pourrait être utilisé pour une vérification du recensement des rapports d'accident, et deuxièmement, l'échantillonnage pourrait être nécessaire pour la partie du système correspondant aux enquêtes sur les accidents nécessitant un travail intensif. Le recensement des RAP fournirait une riche base de sondage pour le système d'enquêtes sur les accidents et d'autres enquêtes. L'information de cette riche base de sondage pourrait être exploitée dans la conception de l'échantillon,

éventuellement en faisant appel à l'échantillonnage équilibré ou à un plan d'échantillonnage qui peut être adapté dynamiquement aux besoins de données et à la base de sondage mise à jour continuellement.

Naturellement, même un recensement des RAP pourrait être obsolète ou insuffisant pour répondre aux besoins de données dans l'avenir. À mesure qu'augmente le nombre de véhicules équipés de caméras et de détecteurs, ou à mesure que les véhicules automoteurs et les systèmes de surveillance deviennent plus fréquents, un échantillon ou un recensement de RAP pourrait être remplacé ou complété par des données recueillies passivement. L'usage accru de sources de données passives à grande échelle soulève d'importantes questions au sujet de la protection de la vie privée et de la propriété des données, qui devront faire l'objet de grands débats et de nombreuses études, et ces questions dépassent le cadre de la présente discussion. Cependant, au-delà des questions sociétales quant à l'éthique de la collecte des données, quelles nouvelles méthodes statistiques sont nécessaires pour répondre à la révolution concernant la disponibilité des données ?

Selon moi, dans un avenir proche, la recherche devra porter sur trois grands domaines interdépendants, qui sont apparentés aux problèmes de recherche auxquels ont fait face Parten, Deming et Hansen au milieu du siècle dernier.

- De meilleures mesures de l'incertitude pour les estimations d'enquêtes. Quand Hansen et Hurwitz (1949, page 365) ont parlé de la supériorité des échantillons probabilistes par rapport aux échantillons au jugé, ils ont souligné que le caractère exempt d'hypothèses de l'inférence à partir d'échantillons probabilistes dépend de l'obtention de taux de réponse élevés : [*Traduction*] « Au *Census Bureau*, il est habituellement supposé que, si l'information requise est obtenue auprès de plus de 95 % des ménages désignés, on peut se sentir à l'aise de supposer que l'échantillon a été tiré conformément à la théorie de l'échantillonnage, même si des hypothèses peuvent être nécessaires pour les 5 % restants. D'aucuns ont constaté que, dans certaines situations, des problèmes surviennent même si l'on ne formule des hypothèses que pour 5 %. » Deming (1950, page 13) a également utilisé 95 % comme borne inférieure pour la validité de l'inférence à partir d'échantillons probabilistes : [*Traduction*] « Un échantillon qui est probabiliste à 95 % ou 98 % et qui, pour les 5 % ou 2 % restants, est sélectionné au jugé ou ajusté au jugé pour tenir compte des refus, des personnes absentes du domicile, etc., peut encore être un excellent échantillon, quoiqu'il soit important d'étudier les 5 % ou même les 2 % restants aussitôt que possible. »

  Au fil des ans, le seuil de 95 % appliqué à l'utilisation de méthodes d'échantillonnage probabiliste pour l'inférence a baissé, au point où, aujourd'hui, les mêmes méthodes de pondération sont appliquées pour un échantillon avec un taux de réponse de 10 % que pour un échantillon avec un taux de réponse de 95 %. À mesure que les taux de réponse ont diminué, des hypothèses de modélisation de plus en plus fortes ont été formulées au sujet des mécanismes de réponse et des populations sous-représentées et non répondantes, mais l'incertitude au sujet de ces hypothèses n'est généralement pas reflétée dans les intervalles de confiance communiqués, ceux-ci étant encore fondés principalement sur l'erreur d'échantillonnage. Lohr et Brick (2017) ont attribué les échecs récents des sondages aux systèmes utilisés pour calculer les intervalles de confiance ou de prédiction a posteriori, et ont soutenu que de nouvelles méthodes statistiques, reflétant mieux

l'incertitude au sujet des estimations, sont nécessaires pour communiquer les estimations des intervalles. Comme l'a déclaré Parten (1950, page 403) : [*Traduction*] « Il n'est pas inhabituel de constater que des techniques statistiques très perfectionnées utilisées pour mesurer les erreurs aléatoires dans les données sont si biaisées que tous les moyens de correction connus ne permettraient pas à l'enquêteur de déterminer ce que devraient être les résultats corrects. »

- Combiner de multiples sources de données. Des méthodes telles que le couplage d'enregistrements, les enquêtes à bases de sondage multiples et les modèles hiérarchiques peuvent faciliter la combinaison de données, et peuvent aussi être utilisées pour évaluer la qualité des données provenant de différentes sources. Lohr et Raghunathan (2017) ont examiné les méthodes statistiques pour combiner des données provenant de sources différentes, et soutenu que l'utilisation de sources multiples d'information pourrait également aider à résoudre le problème de l'évaluation et de l'intégration des erreurs systématiques dans les estimations de l'incertitude, quoique les méthodes statistiques examinées ne résolvent pas le problème de l'obtention d'estimations pour des sous-populations qui pourraient être absentes de toutes les sources.

  Alors que les méthodes d'échantillonnage probabiliste ont bien servi la société au cours des 70 dernières années, elles pourraient jouer un rôle plus limité à l'avenir, peut-être en étant utilisées dans certaines collectes des données pour valider et vérifier d'autres sources de données, au lieu de servir elles-mêmes de sources principales de données. Hansen, Hurwitz et Pritzker (1953b) voyaient sous cet angle les enquêtes postcensitaires durant lesquelles des efforts intensifs étaient faits en vue d'obtenir l'information la plus exacte possible à partir d'un échantillon aréolaire. Leur texte parle aussi de la nécessité de comparer les coûts de l'obtention de statistiques exactes à l'option d'obtenir de plus grandes tailles d'échantillon ou un plus grand nombre de variables. Selon Hansen, Madow et Tepping (1983), c'est précisément quand des changements sociétaux ont lieu que l'absence de biais garantie par l'échantillonnage probabiliste est essentielle et que l'usage ciblé d'échantillons probabilistes de haute qualité peut fournir des évaluations de la couverture et de l'exactitude de l'information provenant d'autres sources de données.

- Enfin, il faut se concentrer à nouveau sur la conception de l'enquête, qui était le sujet des études publiées entre 1920 et 1960. Un système de collecte des données modulaire, s'appuyant sur différentes sources de données pour répondre à différents besoins d'information, pourrait s'adapter à l'évolution des besoins sociétaux et aux nouvelles technologies de collecte de données. Nous avons besoin de plans d'enquête qui sont robustes aux erreurs dans les sources individuelles et permettent d'évaluer ces erreurs, et qui sont également robustes aux changements possibles dans les sources de données.

Comme l'ont souligné Rao et Fuller, le monde de la recherche statistique a répondu à maintes reprises aux besoins d'information de la société grâce à de nouvelles innovations. Les défis que pose le traitement des nouvelles sources de données et des données manquantes sont importants, au même titre que l'étaient les problèmes passés dont la résolution a mené à l'élaboration de l'échantillonnage probabiliste, l'estimation

sur petits domaines, l'estimation de la variance par rééchantillonnage et la théorie de l'imputation. La prochaine révolution en échantillonnage pourrait se profiler juste au tournant.

Remerciements

Une partie de la présente discussion a pour source une conférence magistrale de l'auteure intitulée « The Essential Survey Statistician » donnée en 2016 dans le cadre du JPSM et qui peut être consultée à l'adresse https://www.jpsmclasses.umd.edu/ Mediasite/Catalog/catalogs/default.

Bibliographie

Deming, W.E. (1950). *Some Theory of Sampling.* New York: Dover.

Duncan, J.W., et Shelton, W.C. (1978). *Revolution in United States Government Statistics 1926-1976.* Washington, D.C.: U.S. Department of Commerce.

Hansen, M.H., et Hurwitz, W.N. (1949). Dependable samples for market surveys. *Journal of Marketing,* 14, 363-372.

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

Hansen, M.H., Hurwitz, W.N. et Pritzker, L. (1953b). The accuracy of census results. *American Sociological Review,* 18, 416-423.

Hansen, M.H., Madow, W.G. et Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 384, 776-793.

Hetzel, A.M. (1997). *U.S. Vital Statistics System: Major Activities and Developments, 1950-95.* Hyattsville, MD: National Center for Health Statistics. Disponible à https://www.cdc.gov/nchs/data/misc/usvss.pdf, dernière consultation le 5 mai 2017.

Lohr, S.L., et Brick, J.M. (2017). Roosevelt predicted to win: Revisiting the *Literary Digest* poll of 1936. *Statistics, Politics, and Policy*, 8, 65-84.

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

National Highway Transportation Safety Administration (NHTSA, 2017a). *National Automotive Sampling System (NASS)*. Disponible à https://www.nhtsa.gov/research-data/national-automotive-sampling-system-nass, dernière consultation le 5 mai 2017.

National Highway Transportation Safety Administration (NHTSA, 2017b). *Traffic Safety Facts, 2015*. Disponible à https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812384, dernière consultation le 17 mai 2017.

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

Techniques d'enquête, décembre 2017
Vol. 43, nº 2, pp. 197-199
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article de Rao et Fuller (2017)

Chris Skinner[1]

Résumé

Cette note de Chris Skinner présente une discussion de l'article « Théorie et méthodologie des enquêtes par sondage : orientations passées, présentes et futures » où J.N.K. Rao et Wayne A. Fuller partagent leur vision quant à l'évolution de la théorie et de la méthodologie des enquêtes par sondage au cours des 100 dernières années.

Mots-clés : Collecte des données; histoire de l'échantillonnage; échantillonnage probabiliste; inférence à partir d'enquêtes.

Cet article donne un excellent compte rendu de la théorie et des méthodes de sondage, en distillant de manière concise et élégante une énorme quantité de savoir sur la théorie et la pratique de la discipline. Je ne commenterai ni le passé ni le présent, mais présenterai, comme j'y ai été invité, certaines réflexions sur l'avenir. Je suis entièrement d'accord avec la dernière partie de l'article au sujet de l'avenir et considère que mes idées se superposent à celles des auteurs.

La présente discussion mettra en relief la perspective d'un Institut national de statistique (INS), et je prévois (et j'espère) que les INS joueront un rôle déterminant dans l'évolution de la méthodologie, même si l'environnement statistique connaît des changements, comme la gamme croissante d'organismes qui fournissent des données aux INS et/ou produisent eux-mêmes des statistiques.

Cibles inférentielles : À mon avis, les mêmes types de populations finies cibles descriptives (vues sous l'angle méthodologique) continueront de revêtir un intérêt fondamental. Les besoins analytiques retiendront également leur importance, mais la façon dont ils seront satisfaits dépendra de l'évolution des modalités d'accès aux données, dans le contexte, par exemple, des préoccupations concernant la confidentialité, ainsi que de l'impact de l'évolution de la pratique de la science des données, telle une plus grande importance accordée à la modélisation prédictive.

Sondages et autres sources de données : La nature et la portée des sources de données pertinentes représenteront un domaine de développement critique. Je ne pense pas que les enquêtes disparaîtront, car il existera toujours un très grand nombre de variables d'intérêt nécessitant une collecte directe de données. Mais je m'attends à ce que le sondage fasse de plus en plus partie intégrante d'un ensemble plus vaste de sources de données qui incluent les recensements, les données administratives et les « mégadonnées » (par exemple, Lohr et Raghunathan, 2017; Zhang, 2012). Le défi méthodologique consistera à intégrer efficacement une telle gamme de sources. Les différentes sources peuvent avoir plusieurs propriétaires et les modalités d'accès influeront considérablement sur la façon dont ces sources peuvent être intégrées. Je ne pense pas non plus que l'échantillonnage disparaîtra, car il sera nécessaire non seulement pour la collecte principale de données, mais aussi pour les enquêtes supplémentaires (voir plus bas) et la gestion des sources de mégadonnées.

1. Chris Skinner, London School of Economics and Political Science. Courriel : C.J.Skinner@lse.ac.uk.

Enquêtes supplémentaires : Le besoin d'échantillons d'enquête supplémentaires pour vérifier la validité ou améliorer l'inférence augmentera vraisemblablement. Les « enquêtes de référence » pourraient accroître le nombre d'échantillons non probabilistes (Elliot et Valliant, 2017); des sondages de couverture pourraient être nécessaires pour vérifier la présence à la fois de sous-dénombrement ou de surdénombrement, par exemple, dans les sources de données administratives, et pour corriger ce genre d'erreurs (Zhang, 2015); des enquêtes appariées au niveau de l'unité pourraient être nécessaires pour vérifier la présence d'une erreur de mesure dans les sources de données.

Non-réponse et échantillonnage : La non-réponse totale posera davantage problème et l'inférence devra inévitablement tenir compte de l'erreur de non-réponse, ainsi que de l'erreur d'échantillonnage. La principale difficulté consistera à éviter (réduire) le biais de sélection. Le recours à la randomisation en échantillonnage pour atteindre cet objectif et pour justifier certaines hypothèses de modélisation pourrait devenir une propriété au moins aussi importante de l'échantillonnage probabiliste que son utilisation pour l'inférence fondée sur le plan de sondage. Il pourrait devenir sensé de considérer des protocoles d'échantillonnage et de gestion de la non-réponse d'une manière plus intégrée, et l'étude de ce genre d'options pourrait permettre d'opter pour des protocoles d'échantillonnage qui comprennent des caractéristiques non probabilistes, à condition que la réduction du biais de sélection demeure l'objectif central. En ce qui concerne la réponse, les options multimode flexibles semblent vraisemblablement être les options naturelles à prendre en considération. La nature des sources de données auxiliaires et les considérations concernant l'estimation doivent aussi, évidemment, être prises en compte sérieusement lorsqu'on évalue simultanément les options d'échantillonnage et de gestion de la non-réponse.

Méthodes et théorie d'estimation : Les méthodes d'estimation évolueront afin de tirer parti de nouvelles sortes de relations statistiques dans les sources de données et entre celles-ci, à la fois pour tenir compte des effets de biais de sélection possibles et pour accroître l'efficacité. De nombreux problèmes d'estimation peuvent être formulés en fonction des variables étudiées Y, qui ne peuvent être obtenues que sur des échantillons sélectifs, et des variables explicatives X, pour lesquelles peuvent être construites de grandes bases de données dont la couverture est proche de 100 %. La construction de ce genre de bases de données pourrait être un objectif clé des INS dans le contexte tant des statistiques sur les entreprises que des statistiques sociales. Dans le second cas, cet objectif pourrait être aligné avec les développements concernant le recensement de la population, lesquels comprennent, par exemple, l'utilisation de sources de données administratives (Skinner, 2017). Dans de telles conditions, une approche générale de l'estimation pourrait combiner les distributions de X au niveau de la population et les distributions conditionnelles de Y sachant X tirées des sources d'échantillons sélectifs sous des hypothèses similaires à celles des données manquant au hasard. L'importance des considérations temporelles, dont les avantages de l'emprunt d'information à diverses périodes, augmentera vraisemblablement, et il sera possible d'exploiter les possibilités qu'offrent les sources de données administratives qui sont habituellement longitudinales. Les méthodes de calage existantes et l'estimation sur petits domaines fondée sur le modèle de Fay-Herriot continueront d'être utilisées en tant que sources appariées à un niveau agrégé. L'appariement au niveau de l'individu ou de l'unité fondée sur les coordonnées GPS (par exemple, immeuble ou adresse) pourrait aboutir à d'autres méthodes (par exemple, Lohr et Raghunathan, 2017). La théorie des sondages, y compris les méthodes de prédiction fondées sur un modèle et les méthodes d'estimation sur petits domaines, continuera de jouer un

rôle essentiel. La théorie des données manquantes offre un cadre naturel pour le traitement des sources de données intégrées, et je m'attends à une confluence croissante entre la théorie de l'échantillonnage et celle des données manquantes. Le traitement des erreurs d'appariement et des erreurs de mesure, par exemple attribuables à des différences de mesure entre les sources et les modes de collecte des données, sera également important.

Évaluation de la qualité et estimation de l'exactitude : Vu que les contraintes budgétaires persisteront, il sera essentiel de faire valoir l'importance de normes élevées de qualité auprès des utilisateurs des produits statistiques si l'on veut éviter que les sondages de haute qualité soient remplacés par des solutions bon marché et non fiables. À cet égard, il serait utile de renforcer le rôle d'évaluateurs de la qualité des organes nationaux établis pour surveiller et accroître la confiance du public dans les produits statistiques, surtout si le nombre et la diversité des fournisseurs de ces produits augmentent. Plus précisément, l'évaluation de l'exactitude sera essentielle. Les méthodes classiques d'estimation de la variance pourraient jouer un rôle et être étendues afin de saisir des sources plus vastes de variation, par exemple, en élargissant la définition des répliques dans les méthodes de rééchantillonnage. Cependant, étant donné l'usage croissant de l'inférence fondée sur un modèle, l'évaluation de l'exactitude devra aussi englober l'évaluation de l'effet des écarts par rapport aux hypothèses dans les méthodes d'estimation. Des approches telles que la vérification du modèle, l'application de diagnostics et l'analyse de sensibilité prendront vraisemblablement de l'importance.

Bibliographie

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

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

Skinner, C.J. (2018). Issues and challenges in census taking. *Annual Review of Statistics and its Application*, Volume 5.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66, 41-63.

Zhang, L.-C. (2015). On modelling register coverage errors. *Journal of Official Statistics*, 31, 381-396.

Techniques d'enquête, décembre 2017
Vol. 43, nº 2, pp. 201-230
Statistique Canada, nº 12-001-X au catalogue

Les médias sociaux comme source de données pour les statistiques officielles; l'Indice de confiance des consommateurs des Pays-Bas

Jan van den Brakel, Emily Söhler, Piet Daas et Bart Buelens[1]

Résumé

L'article aborde la question de savoir comment utiliser des sources de données de rechange, telles que les données administratives et les données des médias sociaux, pour produire les statistiques officielles. Puisque la plupart des enquêtes réalisées par les instituts nationaux de statistique sont répétées au cours du temps, nous proposons une approche de modélisation de séries chronologiques structurelle multivariée en vue de modéliser les séries observées au moyen d'une enquête répétée avec les séries correspondantes obtenues à partir de ces sources de données de rechange. En général, cette approche améliore la précision des estimations directes issues de l'enquête grâce à l'utilisation de données d'enquête observées aux périodes précédentes et de données provenant de séries auxiliaires connexes. Ce modèle permet aussi de profiter de la plus grande fréquence des données des médias sociaux pour produire des estimations plus précises en temps réel pour l'enquête par sondage, au moment où les statistiques pour les médias sociaux deviennent disponibles alors que les données d'enquête ne le sont pas encore. Le recours au concept de cointégration permet d'examiner dans quelle mesure la série de rechange représente les mêmes phénomènes que la série observée au moyen de l'enquête répétée. La méthodologie est appliquée à l'Enquête sur la confiance des consommateurs des Pays-Bas et à un indice de sentiments dérivé des médias sociaux.

Mots-clés : Mégadonnées; inférence sous le plan de sondage; inférence sous le modèle; prédiction immédiate; modélisation de séries chronologiques structurelle; cointégration.

1 Introduction

Habituellement, les instituts nationaux de statistique font appel à l'échantillonnage probabiliste conjugué à l'inférence sous le plan de sondage ou assistée par modèle pour produire les statistiques officielles. Le concept d'échantillonnage probabiliste aléatoire a été élaboré principalement en se fondant sur les travaux de Bowley (1926), Neyman (1934), ainsi que Hansen et Hurwitz (1943). Consulter, par exemple, Cochran (1977) ou Särndal, Swensson et Wretman (1992) pour une introduction détaillée à la théorie de l'échantillonnage. Il s'agit d'une approche généralement reconnue, puisqu'elle repose sur une théorie mathématique solide qui montre comment, moyennant la combinaison appropriée d'un plan d'échantillonnage aléatoire et d'un estimateur, des inférences statistiques valides peuvent être faites au sujet de grandes populations finies en se basant sur des échantillons relativement petits. En outre, le degré d'incertitude découlant de l'utilisation de petits échantillons peut être quantifié au moyen de la variance des estimateurs.

Une pression constante s'exerce sur les instituts nationaux de statistique afin qu'ils réduisent les coûts administratifs et le fardeau de réponse. De surcroît, la baisse des taux de réponse suscite la recherche de sources d'information statistique de rechange. Cela pourrait se faire en utilisant des données administratives, comme celles des registres de l'impôt, ou d'autres grands ensembles de données – ce qu'il est convenu

1. Jan van den Brakel, Statistics Netherlands, Methodology Department, Heerlen, Pays-Bas et Maastricht University School of Business and Economics, Department of Quantitative Economics, Pays-Bas. Courriel : ja.vandenbrakel@cbs.nl; Emily Söhler, étudiante en économétrie, Maastricht University; Piet Daas et Bart Buelens, Statistics Netherlands, Methodology Department, Heerlen, Pays-Bas.

d'appeler les mégadonnées – qui sont générés en tant que sous-produit des processus non reliés directement à la production de statistiques. Les renseignements sur l'heure et le lieu de l'activité de réseau fournis par les compagnies de téléphonie mobile, les messages sur les médias sociaux provenant de Twitter et de Facebook, ainsi que les comportements de recherche sur Internet provenant de Google Trends en sont des exemples. Un problème commun à ces sources de données est que le processus de génération des données est inconnu et vraisemblablement sélectif en ce qui a trait à la population cible recherchée. Par conséquent, l'utilisation de ces données pour la production de statistiques officielles représentatives de la population cible pose un problème difficile. Il n'existe aucun plan d'échantillonnage aléatoire facilitant la généralisation des conclusions et des résultats obtenus avec les données disponibles à une population cible plus grande. Donc, l'extraction d'information statistiquement pertinente à partir de ces sources est une tâche compliquée (Daas et Puts, 2014a).

Baker, Brick, Bates, Battaglia, Couper, Dever, Gile et Tourangeau (2013) étudient le problème de l'utilisation d'échantillons non probabilistes et mentionnent que des procédures d'inférence sous le plan de sondage peuvent être appliquées pour corriger le biais de sélection. Buelens, Burger et van den Brakel (2015) explorent la possibilité d'utiliser des algorithmes statistiques d'apprentissage automatique pour corriger le biais de sélection. Au lieu de remplacer les données d'enquête par des données administratives ou des mégadonnées, on peut se servir de ces sources pour améliorer l'exactitude des données d'enquête dans les procédures d'inférence sous un modèle. Marchetti, Giusti, Pratesi, Salvati, Giannotti, Perdreschi, Rinzivillo, Pappalardo et Gabrielli (2015), ainsi que Blumenstock, Cadamuro et On (2015) ont utilisé des mégadonnées comme source d'information auxiliaire pour des modèles transversaux d'estimation sur petits domaines.

De nombreuses enquêtes réalisées par les instituts nationaux de statistique sont des enquêtes répétées. Dans le présent article, nous appliquons une approche de modélisation de séries chronologiques structurelle multivariée pour combiner les séries obtenues au moyen d'une enquête répétée avec des séries provenant d'autres sources de données. Cet exercice répond à plusieurs objectifs. Premièrement, une procédure d'estimation fondée sur un modèle de séries chronologiques augmente la précision des estimations directes en tirant parti de la corrélation temporelle entre les estimations directes issues des diverses éditions de l'enquête. Le recours à la modélisation de séries chronologiques dans le but d'améliorer la précision des données d'enquête a été envisagé par de nombreux auteurs en remontant jusqu'à Blight et Scott (1973). Deuxièmement, l'extension du modèle de séries chronologiques au moyen d'une série auxiliaire permet de modéliser la corrélation entre les composantes non observées des modèles de séries chronologiques structurels, par exemple, les composantes de tendance et saisonnière. Harvey et Chung (2000) proposent un modèle de séries chronologiques pour l'Enquête sur la population active au Royaume-Uni étendu par une série sur les nombres de bénéficiaires de prestations. Si un tel modèle révèle des corrélations fortement positives entre les composantes, cela pourrait accroître encore davantage la précision des estimations des séries chronologiques de l'enquête. Les indicateurs dérivés des médias sociaux sont généralement disponibles plus fréquemment que la série reliée obtenue au moyen d'enquêtes périodiques. L'approche de modélisation de séries chronologiques susmentionnée peut donc être utilisée pour faire des prédictions précoces en temps réel quant aux résultats de l'enquête, au moment où les données des médias sociaux sont

disponibles, tandis que celles de l'enquête ne le sont pas encore. Dans ce cas, les données des médias sociaux constituent une forme de prédiction immédiate. Troisièmement, on peut appliquer le concept de cointégration dans le contexte des modèles espace-état multivariés pour déterminer dans quelle mesure les deux séries sont identiques. Si les composantes tendance des deux séries observées sont cointégrées, ces séries ont pour moteur une tendance commune sous-jacente. On peut soutenir que si une série auxiliaire est cointégrée à la série de l'enquête, les deux séries représentent le même processus stochastique sous-jacent. Cet argument pourrait servir à motiver qu'une statistique mesurée au moyen d'une source de mégadonnées est représentative d'une population cible. Toutefois, cet argument est plutôt empirique et moins solide que la théorie de l'échantillonnage probabiliste, qui prouve que l'échantillonnage aléatoire combiné à un estimateur (approximativement) sans biais sous le plan produit des statistiques représentatives.

L'Enquête sur la confiance des consommateurs (ECC) des Pays-Bas est une enquête réalisée mensuellement auprès d'environ 1 000 personnes en vue de mesurer les sentiments de la population néerlandaise au sujet du climat économique au moyen de ce que l'on appelle l'Indice de confiance des consommateurs (ICC). Daas et Puts (2014b) ont élaboré, à partir des plateformes de médias sociaux, indépendamment de l'ECC, un indice de sentiments qui s'est avéré très bien reproduire l'ICC. Cet indice est nommé Indice basé sur les médias sociaux (IMS). Dans le présent article, nous appliquons l'approche de modélisation de séries chronologiques structurelle multivariée susmentionnée aux deux séries pour tenter d'améliorer la précision de l'ICC. Nous illustrons aussi comment l'IMS peut être utilisé dans ce modèle de séries chronologiques pour faire des prédictions précoces ou prédictions immédiates de l'ICC.

À la section 2, nous décrivons le plan de sondage de l'ECC et la procédure d'estimation utilisée pour produire l'ICC. L'approche suivie par Daas et Puts (2014b) pour construire un indice de sentiments à partir des plateformes de médias sociaux est également décrite. À la section 3, nous proposons un modèle de séries chronologiques structurel pour la série de l'ICC et la série de l'IMS. À la section 4, nous présentons les résultats obtenus au moyen de ce modèle. Enfin, à la section 5, nous concluons l'article par une discussion.

2 Données

2.1 Enquête sur la confiance des consommateurs des Pays-Bas

L'Indice de confiance des consommateurs (ICC), qui est basé sur une enquête mensuelle appelée Enquête sur la confiance des consommateurs (ECC), mesure l'opinion des ménages résidant aux Pays-Bas au sujet du climat économique en général et de leur situation financière en particulier. L'ECC est une enquête continue. Chaque mois, un échantillon autopondéré d'environ 2 500 ménages est tiré selon un plan d'échantillonnage à deux degrés stratifié d'une base de sondage dérivée du Registre municipal des Pays-Bas. Des intervieweurs prennent contact avec les ménages dont le numéro de téléphone est connu pour remplir le questionnaire par interview téléphonique assistée par ordinateur durant les dix premiers jours ouvrables du mois. En moyenne, on obtient un échantillon net d'environ 1 000 ménages répondants, ce qui représente un taux de réponse d'environ 40 %. Une part importante de la non-réponse correspond aux

ménages pour lesquels aucun numéro de téléphone fixe n'est disponible. Le taux de réponse des ménages dont le numéro de téléphone est connu est de l'ordre de 60 %.

L'ICC est fondé sur cinq questions auxquelles peut être donnée une réponse positive, neutre ou négative. Les questions font référence à la situation économique ou financière au cours des 12 mois précédents ou aux attentes des enquêtés au cours des 12 mois à venir. Soit $P_{1,t}^{q}$, $P_{2,t}^{q}$, et $P_{3,t}^{q}$, les pourcentages de personnes qui ont répondu à la question $q = 1, \ldots, 5,$ durant le mois t de manière positive, neutre ou négative, respectivement. L'ICC est défini comme étant la différence entre les pourcentages de réponses positives et négatives, en prenant la moyenne sur les cinq questions :

$$I_t = \frac{1}{Q} \sum_{q=1}^{Q} \left(P_{1,t}^{q} - P_{3,t}^{q} \right). \tag{2.1}$$

Puisque l'échantillon est autopondéré et qu'aucune information auxiliaire n'est utilisée dans la procédure d'estimation, les pourcentages sont estimés en prenant la moyenne d'échantillon, c'est-à-dire

$$P_{j,t}^{q} = \frac{100}{n_t} \sum_{i=1}^{n} \delta_{i,j,t}^{q}, \tag{2.2}$$

pour la question $q = 1, \ldots, 5,$ et la catégorie de réponse $j = 1, 2, 3$. Dans (2.2), n_t est la taille d'échantillon nette durant le mois t et $\delta_{i,j,t}^{q}$ est une variable indicatrice binaire qui est égale à un si le répondant i a choisi la catégorie j pour la question q. En supposant que les ménages sont sélectionnés selon un plan d'échantillonnage aléatoire simple sans remise, on peut prouver que la variance de (2.1) peut être estimée par

$$\begin{aligned} \text{Var}\left(I_t\right) &= \frac{1}{Q^2} \sum_{q=1}^{Q} \left[\text{Var}\left(P_{1,t}^{q}\right) + \text{Var}\left(P_{3,t}^{q}\right) \right] - \frac{2}{Q^2} \sum_{q=1}^{Q} \sum_{q'=1}^{Q} \text{Cov}\left(P_{1,t}^{q}, P_{3,t}^{q'}\right) \\ &+ \frac{1}{Q^2} \sum_{q=1}^{Q} \sum_{q' \neq q}^{Q} \left[\text{Cov}\left(P_{1,t}^{q}, P_{1,t}^{q'}\right) + \text{Cov}\left(P_{3,t}^{q}, P_{3,t}^{q'}\right) \right], \end{aligned} \tag{2.3}$$

avec

$$\text{Var}\left(P_{j,t}^{q}\right) = \frac{1}{n_t} P_{j,t}^{q} \left(100 - P_{j,t}^{q}\right), \qquad \text{Cov}\left(P_{j,t}^{q}, P_{j,t}^{q'}\right) = \frac{1}{n_t} \left(P_{jj,t}^{qq'} - P_{j,t}^{q} P_{j,t}^{q'}\right),$$

$$\text{Cov}\left(P_{j,t}^{q}, P_{j',t}^{q'}\right) = \frac{1}{n_t} \left(P_{jj',t}^{qq'} - P_{j,t}^{q} P_{j',t}^{q'}\right), \quad \text{Cov}\left(P_{j,t}^{q}, P_{j',t}^{q}\right) = -\frac{1}{n_t} P_{j,t}^{q} P_{j',t}^{q},$$

$$P_{jj',t}^{qq'} = \frac{100}{n_t} \sum_{i=1}^{n} \delta_{i,j,t}^{q} \, \delta_{i,j',t}^{q'}.$$

La figure 2.1 montre l'ICC avec un intervalle de confiance (IC) à 95 % calculé en utilisant l'approche décrite à la présente section, observé durant la période de décembre 2000 à mars 2015. La publication officielle de l'ICC est manquante pour octobre 2013.

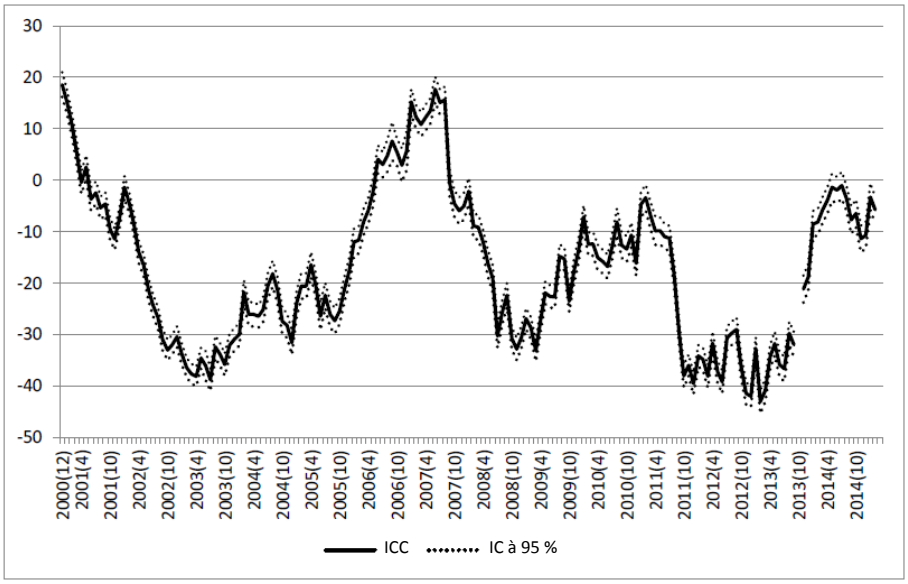

Figure 2.1 Indice de confiance des consommateurs (ICC) avec un intervalle de confiance à 95 %.

2.2 Sentiments sur les médias sociaux

Afin d'essayer de réduire les coûts administratifs et le fardeau de réponse, Daas et Puts (2014b) ont élaboré, en prenant pour sources les médias sociaux, un indice de sentiments pouvant servir d'indicateur de remplacement de l'ICC. Ils sont partis des messages affichés sur les plateformes de médias sociaux les plus populaires aux Pays-Bas et rédigés en néerlandais. Ils les ont classés comme étant des messages positifs, neutres ou négatifs en se servant d'une variante de la classification fondée sur une analyse au niveau de la phrase (Pang et Lee, 2008). L'indice est calculé en prenant la différence entre les pourcentages de messages positifs et négatifs.

Diverses combinaisons de tous les messages affichés sur Facebook et sur Twitter avec et sans application de certains filtres sur les phrases ont été comparées à l'ICC. La combinaison de tous les messages faisant partie du domaine public sur Facebook et de messages sur Twitter filtrés afin qu'ils contiennent des pronoms personnels est celle dont la corrélation avec l'ICC était la plus élevée. Il a fallu filtrer les messages sur Twitter parce qu'un grand nombre de ceux-ci ne sont pas très informatifs. Voir Daas et Puts (2014b) pour des renseignements plus détaillés. Dans le cadre de leur étude, Daas et Puts (2014b) ont également constaté que d'importants changements de comportement du public sur les médias sociaux, tels ceux causés par les grands événements et les variations du nombre de messages affichés sur chaque plateforme, ont un effet perturbateur sur la série. L'indicateur final qu'ils proposent correspond à la moyenne du sentiment dans les messages sur Facebook et sur Twitter durant chaque période.

À la figure 2.2, l'Indice basé sur les médias sociaux (IMS) est comparé à l'ICC pour la période allant de juin 2010 à mars 2015. Les deux séries se situent clairement à des niveaux différents, mais présentent une évolution plus ou moins similaire. Au cours de la période présentée, l'ICC est constamment négatif, tandis que l'IMS est constamment positif. La taille ou amplitude des mouvements est également considérablement

plus grande pour l'ICC que pour l'IMS. De nombreux facteurs sont à l'origine de cette différence, puisque l'ICC est fondé sur une enquête dont la collecte des données est effectuée par téléphone et l'IMS est fondé sur la classification de messages affichés sur Twitter et sur Facebook. La question intéressante est celle de savoir dans quelle mesure l'évolution des deux séries est comparable.

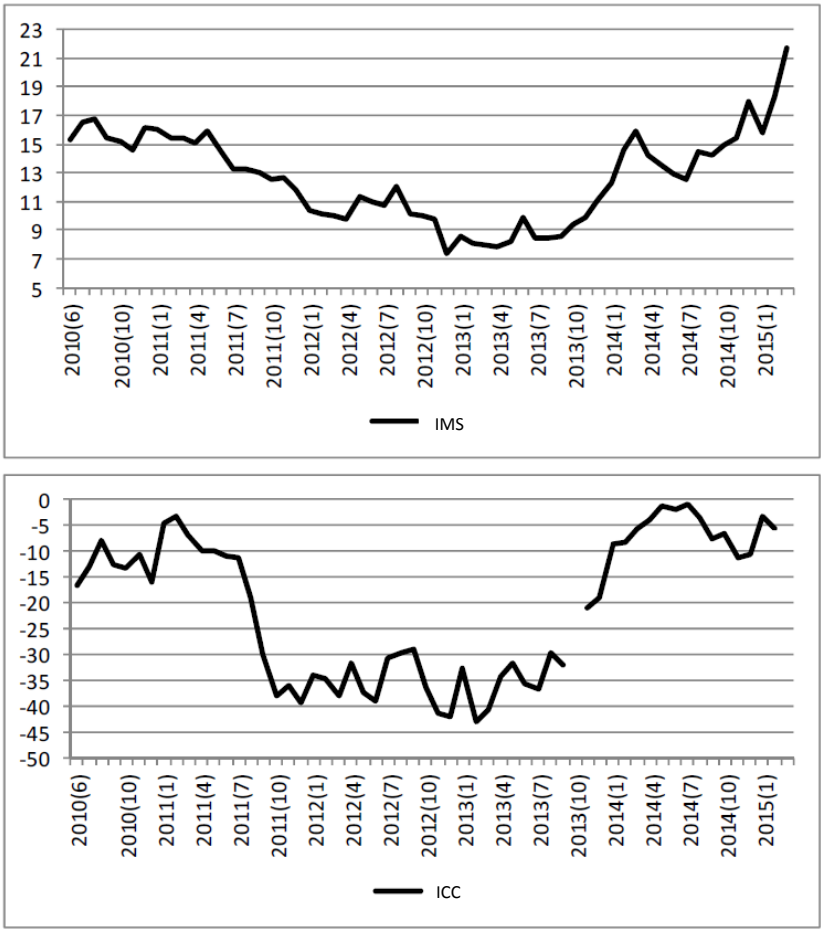

Figure 2.2 Comparaison de l'Indice basé sur les médias sociaux (IMS, graphique supérieur) à l'Indice de confiance des consommateurs (ICC, graphique inférieur).

2.3 Aspects qualitatifs de l'ICC et de l'IMS

L'exactitude d'une statistique est mesurée par sa variance et son biais. Pour simplifier, nous faisons uniquement la distinction entre le biais de sélection et le biais de mesure. La variance d'une statistique issue d'un sondage, comme l'ICC, dépend de la taille de l'échantillon et représente habituellement une part importante de l'incertitude de la statistique d'échantillon. Dans le cas des sources de mégadonnées, le concept de variance d'échantillonnage n'a pas de sens puisque le processus de génération des données n'est pas un échantillonnage probabiliste dans une population cible donnée. En lieu et place, on pourrait se servir des composantes de la variance du modèle utilisé pour décrire le processus supposé de génération des données comme mesure d'exactitude. La variance sous le modèle des statistiques obtenues par application de modèles de séries chronologiques à une série issue de données recueillies sur Internet ou sur les médias sociaux sera systématiquement positive selon la volatilité de la série, laquelle dépend principalement de la

fréquence d'observation de la série et de la dynamique du processus stochastique plutôt que du volume des données.

Le biais de sélection d'une statistique issue d'un sondage est approximativement nul dans des conditions de réponse complète. Toutefois, en pratique, il existe un biais de sélection qui résulte de la non-réponse sélective, de la couverture incomplète de la base de sondage et de la mesure dans laquelle la stratégie de travail sur le terrain permet d'atteindre avec succès la population cible. Dans le cas de l'ICC, un contact est pris uniquement avec les membres de la population dont on connaît le numéro d'appel d'une ligne téléphonique fixe et le taux de réponse de cette sous-population est d'environ 60 %. Dans le cas des sources de mégadonnées, le biais de sélection est généralement inconnu. Dans le présent article, nous appliquons le concept de cointégration pour évaluer le degré auquel l'IMS et l'ICC mesurent le même concept. Notons, cependant, que dans le cas de la cointégration, l'IMS pourrait refléter un biais de sélection relatif à la non-réponse et à la couverture similaire à celui de l'ICC. Baker et coll. (2013) ont fait remarquer qu'il existe des similarités entre le biais de sélection observé pour les échantillons probabilistes et pour l'approche non probabiliste utilisée par les sources de données telles que les médias sociaux.

Le biais de mesure de statistiques issues d'un sondage dépend habituellement de la mesure dans laquelle les variables conceptuelles que l'on veut mesurer sont concrétisées dans le questionnaire, mais aussi du mode de collecte des données et de la qualité des intervieweurs. Dans le cas des enquêtes, le problème de biais de mesure se pose parce que les variables d'intérêt sont mesurées indirectement en ce sens qu'on demande aux participants de faire des déclarations au sujet de leur comportement, ce qui introduit toutes sortes d'erreurs de mesure. Dans le cas de l'ICC, on peut se demander dans quelle mesure les enquêtés sont capables d'exprimer leur confiance à long terme dans l'économie et dans quelle mesure leur réponse est influencée par des émotions de court terme. Ces problèmes ne se posent pas dans le cas des mégadonnées si celles-ci contiennent des mesures directes du comportement des gens. Dans le cas d'un indice basé sur les médias sociaux tel que l'IMS, on peut se demander à quel point il mesure un concept similaire à celui de l'ICC. À la sous-section 2.2, nous avons déjà mentionné que les changements importants de comportement du public sur les médias sociaux perturbent la série. Tout spécialement à la fin de la série, un changement soudain de comportement sur les médias sociaux sera très difficile à distinguer d'un vrai point de retournement. Par exemple, une série issue de Google Trend sur les recherches liées aux emplois vacants pourrait suivre la courbe d'une série officielle sur le chômage. Cette série mesure le chômage, cependant, le comportement de recherche avant le début de la crise financière de 2009 pourrait être entièrement différent de celui de la période suivant directement la crise financière, ce qui invalide le concept que l'on veut mesurer.

3 Modélisation de séries chronologiques structurelle de l'ICC et de l'IMS

La présente section décrit l'élaboration de modèles de séries chronologiques structurels univarié et bivarié pour l'ICC et l'IMS. Dans un modèle de séries chronologiques structurel, la série est décomposée en une composante tendance, une composante saisonnière, d'autres composantes cycliques, une composante de régression et une composante irrégulière. Chaque composante est supposée suivre un modèle stochastique, ce qui permet que les composantes tendance, saisonnière et cyclique, mais aussi les

coefficients de régression dépendent du temps. Au besoin, des composantes autorégressives-moyennes mobiles (ARMA) peuvent être ajoutées pour tenir compte de l'autocorrélation dans la série au-delà de ces composantes structurelles. Consulter Harvey (1989) ou Durbin et Koopman (2012) pour des précisions au sujet de la modélisation de séries chronologiques structurelle.

La question abordée dans le présent article est celle de savoir dans quelle mesure l'IMS suit une courbe semblable à l'ICC, de sorte que l'IMS puisse être utilisé dans la procédure d'estimation de l'ICC, voire, dans le cas le plus extrême, le remplacer. Pour traiter cette question, nous élaborons un modèle de séries chronologiques structurel bivarié pour l'ICC et pour l'IMS, et modélisons la corrélation entre les termes de perturbation des différentes composantes du modèle structurel pour les deux séries. Nous appliquons le concept de cointégration pour déterminer si les composantes non observées des deux séries sont sous-tendues par des facteurs communs. Si, par exemple, les tendances des deux séries sont sous-tendues par une tendance commune, on pourrait soutenir que l'IMS représente une évolution des sentiments comparable à l'ICC. L'IMS pourrait aussi être utilisé comme une série auxiliaire dans une procédure d'estimation de l'ICC fondée sur un modèle ou dans une procédure de prédiction immédiate pour obtenir des estimations en temps réel plus précises.

3.1 Modèle univarié de la série de l'ICC

En guise de première étape, nous proposons un modèle de séries chronologiques univarié pour la série de l'ICC. Selon l'approche fondée sur le plan de sondage décrite à la section 2.1, l'information observée chaque mois sur l'échantillon sert à calculer une estimation de l'ICC pour le mois en question. Un inconvénient de cette approche est que l'information observée lors des périodes précédentes n'est pas utilisée pour obtenir des estimations plus précises de l'ICC. Dans le domaine des techniques d'enquête, il est fréquent d'appliquer des modèles de séries chronologiques pour obtenir des estimations pour des enquêtes périodiques. Blight et Scott (1973) et Scott et Smith (1974) ont proposé de considérer les paramètres de population inconnus comme une réalisation d'un processus stochastique qui peut être décrit au moyen d'un modèle de séries chronologiques. Cela introduit des relations entre les paramètres de population estimés à différents points dans le temps dans le cas d'échantillons non chevauchants ainsi que chevauchants. La modélisation explicite de cette relation entre les estimations issues de l'enquête au moyen d'un modèle de séries chronologiques peut servir à combiner l'information observée sur l'échantillon dans le passé pour améliorer la précision des estimations obtenues au moyen d'enquêtes périodiques. Parmi les références clés à des auteurs qui ont appliqué l'approche des séries chronologiques aux données d'enquêtes répétées pour améliorer l'efficacité des estimations par sondage, nous mentionnerons Scott, Smith et Jones (1977), Tam (1987), Binder et Dick (1989, 1990), Bell et Hillmer (1990), Tiller (1992), Rao et Yu (1994), Pfeffermann et Burck (1990), Pfeffermann (1991), Pfeffermann et Rubin-Bleuer (1993), Pfeffermann, Feder et Signorelli (1998), Pfeffermann et Tiller (2006), Harvey et Chung (2000), Feder (2001), Lind (2005) et van den Brakel et Krieg (2009, 2015).

L'élaboration d'un modèle de séries chronologiques pour les estimations par sondage observées au moyen d'une enquête périodique débute par un modèle énonçant que l'estimation par sondage peut être décomposée en la valeur de la variable dans la population et une erreur d'échantillonnage :

$$I_t = \theta_t + e_t, \tag{3.1}$$

où θ_t désigne l'ICC réel au mois t sous un dénombrement complet de la population cible et e_t, l'erreur d'échantillonnage.

L'ICC est observé mensuellement. Par conséquent, à titre de première étape, la série du paramètre de population finie peut être décomposée en une tendance stochastique, une composante saisonnière pour modéliser les écarts systématiques par rapport à la tendance durant une année, et une composante de bruit blanc pour les variations restantes, inexpliquées. Ces considérations mènent au modèle qui suit pour la série du paramètre de population finie :

$$\theta_t = L_t + S_t + \xi_t, \tag{3.2}$$

où L_t désigne une tendance stochastique, S_t, une composante saisonnière stochastique et ξ_t, la variation inexpliquée du paramètre de population finie. L'insertion de (3.2) dans le modèle de mesure (3.1) donne

$$I_t = L_t + S_t + \xi_t + e_t. \tag{3.3}$$

Dans une enquête transversale, il est difficile de séparer l'erreur d'échantillonnage du bruit blanc du paramètre de population. Donc, les deux composantes sont combinées en un terme de perturbation

$$\upsilon_t = \xi_t + e_t. \tag{3.4}$$

Nous supposons que $E(\upsilon_t) = 0$ et $\text{Var}(\upsilon_t) = \sigma_\upsilon^2$. Pour tenir compte de la variance non homogène dans les erreurs d'échantillonnage, Binder et Dick (1990) ont proposé une erreur de mesure où les termes de perturbation υ_t sont proportionnels aux erreurs-types de I_t, c'est-à-dire

$$\upsilon_t = \sqrt{\text{Var}(I_t)}\ \tilde{\upsilon}_t, \tag{3.5}$$

avec $E(\tilde{\upsilon}_t) = 0$, $\text{Var}(\tilde{\upsilon}_t) = \sigma_\upsilon^2$ et où $\text{Var}(I_t)$ est définie par (2.3) et est utilisée comme une information a priori dans le modèle de séries chronologiques. Un tel modèle serait utile si l'erreur d'échantillonnage est plus importante que le bruit blanc dans le paramètre de population. Pour la présente application, les premières analyses indiquent que la variance du bruit blanc de la population est importante, ce qui invalide (3.5). En outre, toujours dans cette application, la variance de l'erreur d'échantillonnage est constante au cours du temps. Nous avons donc décidé de combiner l'erreur d'échantillonnage et le bruit blanc de la population, et avons supposé que la variance était constante au cours du temps. Savoir comment tenir compte de la variance d'échantillonnage est une question qui se pose également dans le cas des variances de désaisonnalisation (Pfeffermann et Sverchkov, 2014). Bell (2005) a étudié la contribution de la variance d'échantillonnage à la variance de l'erreur d'estimation des séries désaisonnalisées et à la composante non saisonnière. Dans le cas de panels (rotatifs), l'erreur d'échantillonnage peut être isolée du bruit blanc de la population. Dans le cas des enquêtes transversales répétées, il est difficile d'identifier les composantes distinctes et les deux termes sont donc combinés en un terme de perturbation qui inclut à la fois la variance d'échantillonnage et la variation inexpliquée du paramètre de population.

Un exercice approfondi de sélection du modèle a montré qu'un modèle de tendance lissé est le plus approprié pour représenter la tendance et le cycle économique dans la série de l'ICC. Le modèle de tendance lissé est défini comme étant (Durbin et Koopman, 2012) :

$$L_t = L_{t-1} + R_t,$$

$$R_t = R_{t-1} + \eta_t, \quad \mathrm{E}(\eta_t) = 0, \tag{3.6}$$
$$\mathrm{Cov}(\eta_t, \eta_{t'}) = \begin{cases} \sigma_\eta^2 & \text{si} \quad t = t' \\ 0 & \text{si} \quad t \neq t' \end{cases}.$$

L'ajout d'une composante aléatoire pour le niveau dans (3.6) améliore la log-vraisemblance de cinq unités, mais aboutit à un surajustement des données en ce sens que le signal lissé suit presque exactement la série observée, avec une très petite variance de l'erreur de mesure. Un modèle au niveau local (niveau aléatoire sans une pente) améliore la log-vraisemblance de trois unités, mais a également tendance à surajuster les données.

La composante saisonnière est modélisée par un modèle trigonométrique, qui est défini comme étant (Durbin et Koopman, 2012) :

$$S_t = \sum_{j=1}^{6} \gamma_{jt}, \tag{3.7}$$

avec

$$\begin{aligned} \gamma_{jt} &= \gamma_{jt-1} \cos(\lambda_j) + \tilde{\gamma}_{jt-1} \sin(\lambda_j) + \omega_{jt}, \\ \tilde{\gamma}_{jt} &= -\gamma_{jt-1} \sin(\lambda_j) + \tilde{\gamma}_{jt-1} \cos(\lambda_j) + \tilde{\omega}_{jt}. \end{aligned}$$

Ici, λ_j désigne la fréquence des différents cycles exprimée en radians et définie comme étant

$$\lambda_j = \frac{2\pi j}{12}, \qquad \text{pour} \qquad j = 1, \ldots, 6.$$

Pour les termes de perturbation, il est supposé que

$$E(\omega_{jt}) = 0, \qquad E(\tilde{\omega}_{jt}) = 0,$$
$$\mathrm{Cov}(\omega_t, \omega_{t'}) = \begin{cases} \sigma_\omega^2 & \text{si} \quad t = t' \\ 0 & \text{si} \quad t \neq t' \end{cases}.$$

Par souci de parcimonie, nous supposons que la structure de variance est la même avec le même hyperparamètre pour $\tilde{\omega}_{jt}$. Qui plus est, nous supposons que ω_t et $\tilde{\omega}_t$ ne sont pas corrélés.

Après l'introduction de la composante de tendance stochastique (3.6) et de la composante saisonnière (3.7), aucune autre composante cyclique n'est nécessaire. La procédure de sélection du modèle a montré que deux changements de niveau sont nécessaires pour modéliser des sauts soudains dans la série. Le premier est dû à la crise financière de septembre 2008, et le second, au ralentissement économique de septembre 2011. Enfin, une valeur aberrante ponctuelle est nécessaire pour septembre 2007. L'ajout de ces trois composantes accroît la log-vraisemblance de 15 unités. Nous arrivons ainsi au modèle qui suit pour la série observée de l'ICC

$$I_t = L_t + S_t + \beta^{07}\delta_t^{07} + \beta^{08}\delta_t^{08} + \beta^{11}\delta_t^{11} + \upsilon_t, \tag{3.8}$$

avec

$$\delta_t^{07} = \begin{cases} 1 & \text{si} \quad t = 2007(9) \\ 0 & \text{si} \quad t \neq 2007(9) \end{cases}, \qquad \delta_t^{08} = \begin{cases} 1 & \text{si} \quad t \geq 2008(9) \\ 0 & \text{si} \quad t < 2008(9) \end{cases}, \qquad \delta_t^{11} = \begin{cases} 1 & \text{si} \quad t \geq 2011(9) \\ 0 & \text{si} \quad t < 2011(9) \end{cases},$$

et β^x représente les coefficients de régression correspondants.

Enfin, des composantes autorégressives (AR) et de moyennes mobiles (MA pour *moving average*) peuvent être ajoutées au modèle de séries chronologiques structurel (3.8). Dans la présente application, rien n'indique que de telles composantes soient nécessaires, puisqu'il n'y a aucun signe évident d'une corrélation sériale résiduelle entre les innovations standardisées. L'ajout d'un processus AR(1) ou MA(1) à (3.8) augmente la log-vraisemblance de 5 et de 4,5 unités, respectivement. L'ajout de modèles AR ou MA d'ordre deux n'améliore pas davantage la log-vraisemblance. L'ajout d'un processus ARMA(1,1) n'accroît pas non plus davantage la log-vraisemblance. Un processus AR(1) ou MA(1) améliore légèrement le corrélogramme, mais augmente aussi l'erreur-type des signaux lissés filtrés. Donc, nous avons finalement choisi le modèle (3.8) pour la série de l'ICC.

Les modèles espace-état supposent que les termes de perturbation suivent des lois normales indépendantes. Ces hypothèses se traduisent en l'hypothèse que les innovations suivent des lois normales indépendantes. Le tableau A.1 en annexe donne un aperçu des statistiques de qualité de l'ajustement appliquées aux innovations standardisées. Les valeurs obtenues pour l'asymétrie, l'aplatissement et le test de Bowman-Shenton ne révèlent pas d'écarts par rapport à la loi normale pour les innovations standardisées. Les valeurs pour le test de Ljung-Box et le test de Durban-Watson n'indiquent aucune corrélation sériale entre les innovations standardisées. Ces observations sont également confirmées par un corrélogramme (non présenté). En conclusion, selon ces diagnostics, le modèle (3.8) est raisonnablement bien ajusté à la série de l'ICC.

3.2 Modèle bivarié des séries de l'ICC et de l'IMS

L'étape suivante consiste à combiner le modèle univarié pour l'ICC avec la série pour l'IMS. Avant de combiner l'ICC et l'IMS dans un modèle bivarié, nous élaborons un modèle univarié pour l'IMS afin de mieux comprendre le comportement de cette série. Une procédure de sélection de modèle similaire à celle effectuée pour la série de l'ICC à la sous-section 3.1 indique que la série observée pour l'IMS peut être modélisée avec un modèle de tendance lisse et une composante de bruit blanc pour la variation inexpliquée. Aucune présence significative d'une composante saisonnière ou d'un cycle économique n'est établie. Il n'existe aucun signe de valeur aberrante ni de changements de niveau. Nous n'avons pas inclus de composante AR(1) et MA(1) puisqu'il n'existe aucune corrélation sériale entre les innovations standardisées. Ces observations ont mené à un modèle bivarié pour l'ICC et l'IMS dans lequel l'ICC comprend une tendance et une composante saisonnière, tandis que l'IMS comprend une composante tendance.

Les tableaux A.2 et A.3 en annexe donnent un aperçu des statistiques de qualité de l'ajustement pour les innovations standardisées de l'ICC et de l'IMS, respectivement. Rien n'indique que les innovations standardisées s'écartent d'une loi normale dans l'une ou l'autre des deux séries. L'hypothèse nulle d'absence de corrélation sériale entre les innovations standardisées n'a pas pu être rejetée. Le corrélogramme des innovations pour l'IMS montre toutefois un patron saisonnier non significatif (données non présentées). Les innovations de l'IMS présentent aussi une hétéroscédasticité.

Les termes de perturbation des tendances des deux séries sont corrélés. Puisque la série pour l'IMS est disponible à partir de juin 2010, le modèle pour l'ICC contient aussi le dernier changement de niveau en septembre 2011, mais non la valeur aberrante ponctuelle en septembre 2007 et le changement de niveau en septembre 2008. Par conséquent, nous obtenons le modèle bivarié suivant :

$$\begin{pmatrix} I_t \\ X_t \end{pmatrix} = \begin{pmatrix} L_t^I \\ L_t^X \end{pmatrix} + \begin{pmatrix} S_t^I \\ 0 \end{pmatrix} + \begin{pmatrix} \beta^{11}\delta_t^{11} \\ 0 \end{pmatrix} + \begin{pmatrix} \upsilon_t^I \\ \upsilon_t^X \end{pmatrix}, \tag{3.9}$$

dans lequel L_t^I et L_t^X désignent le modèle de tendance lissé défini en (3.6) avec la structure de covariance

$$\begin{aligned}
\text{Cov}\left(\eta_t^I, \eta_{t'}^I\right) &= \begin{cases} \sigma_{\eta I}^2 & \text{si } t = t' \\ 0 & \text{si } t \neq t' \end{cases}, \\
\text{Cov}\left(\eta_t^X, \eta_{t'}^X\right) &= \begin{cases} \sigma_{\eta X}^2 & \text{si } t = t' \\ 0 & \text{si } t \neq t' \end{cases}, \\
\text{Cov}\left(\eta_t^I, \eta_{t'}^X\right) &= \begin{cases} \sigma_{\eta I}\sigma_{\eta X}\rho_\eta & \text{si } t = t' \\ 0 & \text{si } t \neq t' \end{cases}.
\end{aligned}$$

Dans la dernière expression, ρ_η désigne la corrélation entre les perturbations de la pente de l'ICC et de l'IMS. De surcroît, S_t^I représente l'effet saisonnier défini par (3.7) et δ_t^{11}, le changement de niveau en septembre 2011 avec le coefficient de régression correspondant β^{11}. Enfin, υ_t^I et υ_t^X sont les termes de perturbation pour les séries de l'ICC et de l'IMS, qui sont définis comme il suit :

$$\begin{aligned}
E(\upsilon_t^I) &= E(\upsilon_t^X) = 0, \\
\text{Cov}\left(\upsilon_t^I, \upsilon_{t'}^I\right) &= \begin{cases} \sigma_{\upsilon I}^2 & \text{si } t = t' \\ 0 & \text{si } t \neq t' \end{cases}, \\
\text{Cov}\left(\upsilon_t^X, \upsilon_{t'}^X\right) &= \begin{cases} \sigma_{\upsilon X}^2 & \text{si } t = t' \\ 0 & \text{si } t \neq t' \end{cases}, \\
\text{Cov}\left(\upsilon_t^I, \upsilon_{t'}^X\right) &= 0 \text{ pour tout } t \text{ et } t'.
\end{aligned}$$

Si le modèle détecte une forte corrélation entre les tendances de l'ICC et de l'IMS, alors les tendances des deux séries se développeront dans la même direction plus ou moins simultanément. Dans ces conditions, l'information supplémentaire provenant de la série de l'IMS aboutira à une plus grande précision des estimations des chiffres de l'ICC. Dans le cas d'une forte corrélation entre les perturbations des tendances, c'est-à-dire si $\rho_\eta \to 1$, les tendances sont dites cointégrées. Dans ces conditions, il existe une tendance commune sous-jacente qui dicte l'évolution des tendances des deux séries observées. Pour le voir, nous notons que la matrice de covariance des perturbations de la pente est obtenue sous forme d'une décomposition en valeurs singulières :

$$\text{cov}\begin{pmatrix} \eta_t^I \\ \eta_t^X \end{pmatrix} = \begin{pmatrix} \sigma_{\eta I}^2 & \sigma_{\eta I}\sigma_{\eta X}\rho_\eta \\ \sigma_{\eta I}\sigma_{\eta X}\rho_\eta & \sigma_{\eta X}^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix}\begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix}\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}. \tag{3.10}$$

Au lieu de $\sigma^2_{\eta I}$, $\sigma^2_{\eta X}$ et ρ_η, ce sont les paramètres d_1, d_2 et a qui sont estimés. Si $d_2 \to 0$, il s'ensuit que $\rho_\eta \to 1$. Dans ces conditions, la matrice de covariance des perturbations de la pente est de rang réduit et les deux tendances sont sous-tendues par une tendance commune. Cela implique que les perturbations des pentes des deux séries montent ou descendent simultanément et que les perturbations de la pente de l'IMS peuvent être prédites parfaitement à partir des perturbations de la pente de l'ICC au moyen de $\eta_t^X = a\eta_t^I$. En outre, la pente de la série de l'IMS peut s'exprimer sous forme d'une combinaison linéaire de la pente de la série de l'ICC par l'expression $R_t^X = aR_t^I + \bar{R}$. De même, la tendance de la série de l'IMS peut être exprimée sous forme d'une combinaison linéaire de la tendance pour la série de l'ICC par l'expression $L_t^X = aL_t^I + \bar{L} + \bar{R}t$. Notons que $\bar{R}$ et $\bar{L}$ sont des constantes qui sont calculées à partir des états estimés aux deux dernières périodes de la série.

La cointégration accroît la précision des estimations de la tendance et du signal de la série de l'ICC, permet de formuler des modèles plus parcimonieux, mais pourrait aussi être considérée comme un argument en vue de remplacer la série de l'ICC par celle de l'IMS, puisque les deux séries sont sous-tendues par une même tendance commune et représentent toutes deux cette tendance. Pour une discussion plus détaillée de la cointégration dans le contexte de la modélisation espace-état, consulter Koopman, Harvey, Shephard et Doornik (2009, sections 6.4 et 9.1).

3.3 Estimation des modèles de séries chronologiques structurels

Le moyen général d'analyser un modèle de séries chronologiques structurel consiste à l'exprimer dans la représentation dite espace-état et à appliquer le filtre de Kalman pour obtenir des estimations optimales pour les variables d'état (voir par exemple, Durbin et Koopman (2012)). Le logiciel pour l'analyse et l'estimation des modèles de séries chronologiques est développé en Ox en combinaison avec les sous-routines de SsfPack 3.0; voir Doornik (2009) et Koopman, Shephard et Doornik (2008).

Toutes les variables d'état sont non stationnaires et initialisées au moyen d'un prior diffus, c'est-à-dire que les espérances des états initiaux sont nulles et que la matrice de covariance initiale des états est diagonale avec de grands éléments diagonaux. Dans Ssfpack 3.0, une fonction de log-vraisemblance diffuse exacte s'obtient à l'aide de la procédure proposée par Koopman (1997). Les estimations du maximum de vraisemblance (MV) pour les hyperparamètres, c'est-à-dire les composantes de variance des processus stochastiques pour les variables d'état, sont obtenues avec une procédure d'optimisation numérique (algorithme de Broyden-Fletcher-Goldfarb-Shanno (BFGS), Doornik, 2009). Pour éviter d'obtenir des estimations de variance négatives, on estime les variances log-transformées. Le lecteur trouvera d'autres renseignements techniques sur l'analyse des modèles espace-état dans Harvey (1989) ou dans Durbin et Koopman (2012).

Sous l'hypothèse que les termes de perturbation suivent une loi normale, on peut appliquer le filtre de Kalman pour obtenir des estimations optimales des variables d'état, voir par exemple, Durbin et Koopman (2012). Le filtre de Kalman suppose que les termes de variance et de covariance sont connus d'avance et ces termes sont souvent appelés hyperparamètres. En pratique, ces hyperparamètres sont inconnus et, par conséquent, remplacés par l'estimation de leur MV. Les estimations pour les variables d'état pour la période t fondées sur l'information disponible jusqu'à la période t inclusivement sont appelées *estimations filtrées*.

Elles s'obtiennent à l'aide du filtre de Kalman où les estimations du MV des hyperparamètres sont fondées sur la série chronologique complète. Les estimations filtrées des vecteurs d'état antérieurs peuvent être mises à jour si de nouvelles données deviennent disponibles. Cette procédure, appelée lissage, donne des *estimations lissées* qui sont fondées sur la série chronologique complète.

Les erreurs-types des estimations obtenues avec le filtre de Kalman ne reflètent pas l'incertitude supplémentaire due à l'utilisation des estimations du MV pour les hyperparamètres inconnus. Donc, les estimations des erreurs-types sont trop optimistes.

4 Résultats

4.1 Modèle univarié de la série de l'ICC

L'analyse univariée est fondée sur le modèle (3.8) décrit à la section 3.1 et appliqué à la série de l'ICC obtenue de décembre 2000 à mars 2015. Le tableau 4.1 donne les estimations du MV pour les hyperparamètres du modèle.

Tableau 4.1
Estimations du maximum de vraisemblance des hyperparamètres du modèle univarié de l'ICC

Écart-type	Estimation du MV
Tendance (σ_η)	1,18
Composante saisonnière (σ_ω)	0,0025
Équation de mesure (σ_υ)	2,46

La moyenne des erreurs-types des estimations directes pour l'ICC est égale à 1,21. Il découle du tableau 4.1 que l'écart-type des termes de perturbation de l'équation de mesure est égale à 2,46. Cela illustre le fait que le bruit blanc de la population est plus grand que la variance des termes de perturbation de la mesure, tel qu'il est indiqué par le choix de la structure de variance pour (3.4) à la section 3.1.

Dans le graphique supérieur de la figure 4.1, la tendance lissée plus les valeurs aberrantes sont comparées aux estimations directes pour l'ICC. Dans le graphique inférieur de la figure 4.1, le signal lissé, défini comme étant la tendance plus les valeurs aberrantes plus la composante saisonnière, est comparé aux estimations directes pour l'ICC. Dans la série contenant la tendance lissée et les valeurs aberrantes, l'effet saisonnier, le bruit blanc du paramètre de population et l'erreur d'échantillonnage sont supprimés de la série originale. Il découle de la figure 4.1 que le modèle de séries chronologiques permet d'obtenir une estimation plus stable de l'ICC. La série de la tendance filtrée plus les valeurs aberrantes est comparée aux estimations lissées à la figure 4.2. Cette série filtrée est une approximation de ce que l'on obtiendrait dans la production des statistiques officielles si aucune révision n'était publiée. Il s'ensuit que, même dans ce cas, il est possible d'éliminer une part considérable de la variation à haute fréquence et des fluctuations saisonnières. Les deux figures montrent que le filtre de Kalman fournit des imputations lissées, mais aussi filtrées, plausibles pour l'observation manquante en octobre 2013.

La figure 4.3 donne la courbe saisonnière lissée de la série de l'ICC. Puisque les effets saisonniers sont presque invariants dans le temps, les effets sont présentés pour les 12 mois d'une année seulement. Il existe clairement des effets négatifs significatifs en octobre, novembre et décembre, et des effets positifs en janvier et en août. L'objectif de l'ICC est de mesurer la confiance à long terme des enquêtés, puisque toutes les questions font référence à leur situation financière et économique au cours des 12 derniers mois ou à leurs attentes pour les 12 mois à venir. Cependant, le patron saisonnier clair et significatif indique que les réponses des enquêtés sont manifestement dictées par des émotions à beaucoup plus court terme, qui sont, entre autres, sujettes à des fluctuations saisonnières.

À la figure 4.4, l'erreur-type des estimations directes pour l'ICC est comparée aux erreurs-types de la série de la tendance filtrée et lissée plus les valeurs aberrantes. Les pics de la courbe de l'erreur-type des estimations filtrées et lissées sont le résultat des variables de valeurs aberrantes et de l'observation manquante en 2013. Si une variable de valeur aberrante est activée à un certain point dans le temps, il faut estimer un nouveau coefficient de régression. Cela produit une incertitude supplémentaire dans les estimations du modèle, ce qui se manifeste par un pic soudain de l'erreur-type de la tendance filtrée et lissée. En 2013, une observation est manquante, ce qui donne aussi lieu à une incertitude supplémentaire, puisque le modèle espace-état produit une prédiction pour la valeur manquante.

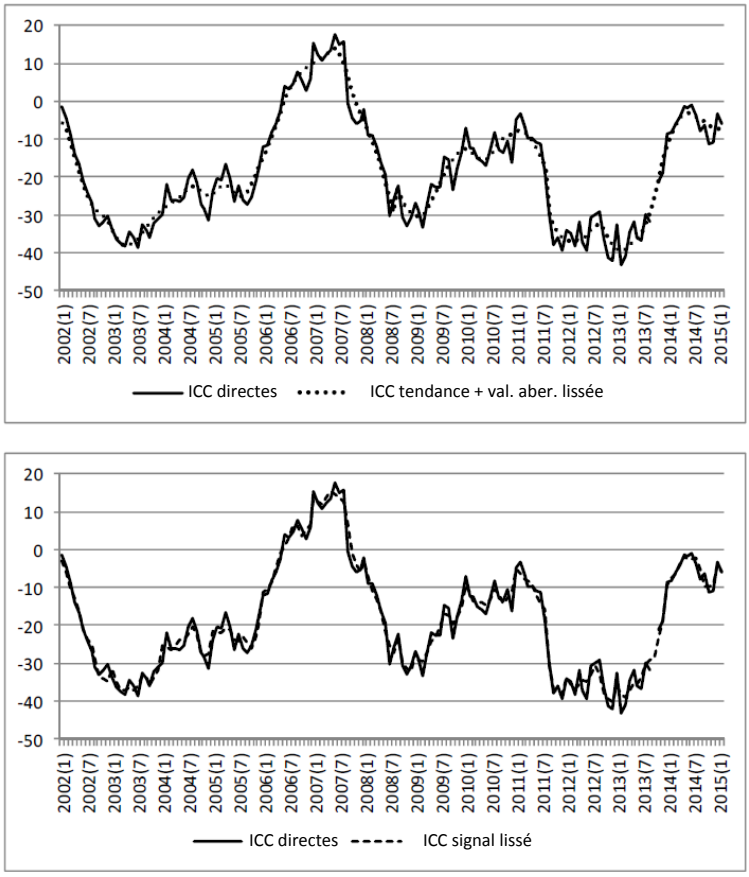

Figure 4.1 Tendance plus valeurs aberrantes lissée comparées aux estimations directes de l'ICC (graphique supérieur) et signal lissé (tendance plus valeurs aberrantes plus saisonnière) comparé aux estimations directes de l'ICC (graphique inférieur).

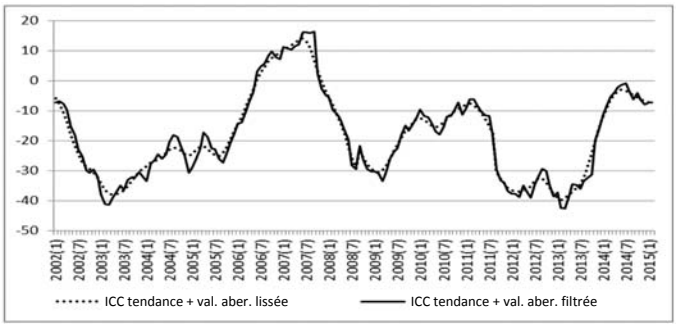

Figure 4.2 Tendance plus valeurs aberrantes filtrée comparées à la tendance plus valeurs aberrantes lissée pour l'ICC.

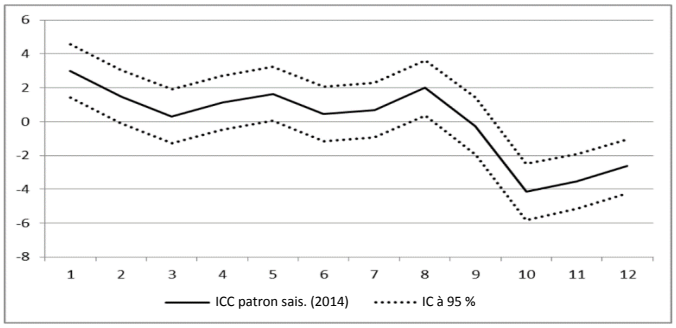

Figure 4.3 Patron saisonnier lissé de l'ICC pour 2014.

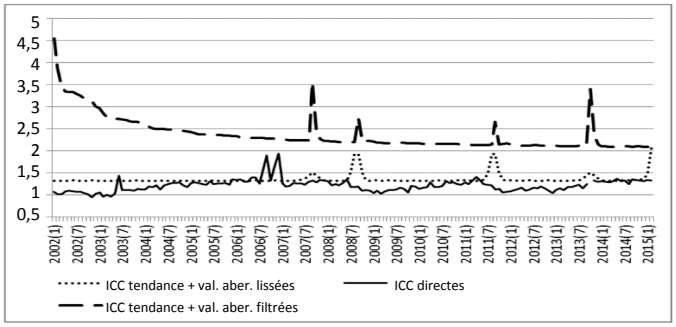

Figure 4.4 Erreur-type des tendance plus valeurs aberrantes lissées et filtrées comparée aux estimations directes de l'ICC.

Les erreurs-types des estimations lissées sont un peu plus grandes que celles des estimations directes. Les erreurs-types des estimations filtrées sont considérablement plus grandes que celles des estimations directes. Ce résultat est remarquable. Les estimations filtrées et lissées fournies par le modèle de séries chronologiques s'appuient sur un beaucoup plus grand ensemble de données puisque les données d'échantillon des périodes précédentes (dans le cas des estimations filtrées) ou la série de données complète (dans le cas des estimations lissées) sont utilisées pour obtenir une estimation optimale pour l'ICC mensuel. Par contre, les estimations directes sont fondées sur l'échantillon observé durant le mois en question seulement. La plupart des applications dans lesquelles les modèles de séries chronologiques structurels sont utilisés comme une forme d'estimation sur petits domaines donnent lieu à d'importantes réductions de

l'erreur-type comparativement aux estimations directes; voir, par exemple, van den Brakel et Krieg (2009, 2015) et Bollineni-Balabay, van den Brakel et Palm (2015, 2017).

La raison pour laquelle, dans la présente application, une approche de modélisation de séries chronologiques produit des erreurs-types pour les estimations filtrées et lissées issues du modèle de séries chronologiques plus grandes que les erreurs-types des estimations directes tient à une grande composante de bruit blanc dans la valeur de population réelle de l'ICC. Rappelons, comme il est mentionné à la section 3.1, que le terme de perturbation de (3.8) contient deux composantes, à savoir l'erreur d'échantillonnage et la variation de haute fréquence inexpliquée de la valeur de population réelle, exprimées par (3.4). Rappelons du tableau 4.1 que σ_{υ} est égale à 2,46 et est deux fois plus grande que la valeur moyenne des erreurs-types des estimations directes. Il s'agit d'un fort indice que la variance de la composante de bruit blanc dans la variable de population réelle est du même ordre que celle de l'erreur d'échantillonnage. L'estimateur direct de l'ICC obtenu à la section 2 traite l'ICC pour chaque mois particulier comme une variable fixe, mais inconnue. La variance de l'estimateur direct mesure uniquement l'incertitude, puisqu'un petit échantillon est observé au lieu de la population complète pour estimer l'ICC. Elle ne mesure pas la variation de haute fréquence de la valeur de population au cours du temps. Cela explique pourquoi l'approche de modélisation de séries chronologiques n'aboutit pas à une réduction de l'erreur-type de l'ICC estimé.

Même si le gain de précision des estimations de niveau obtenues au moyen du modèle de séries chronologiques est limité, les estimations de la tendance sont plus stables, comme l'illustrent les figures 4.1 et 4.2. Un modèle de séries chronologiques demeurera donc utile pour isoler une tendance à long terme plus stable de la variation de haute fréquence dans le paramètre de population et de l'erreur d'échantillonnage. Comme les variables d'état de la composante tendance des périodes subséquentes présenteront une forte corrélation positive, on peut s'attendre à ce que l'approche de modélisation de séries chronologiques offre davantage de gain si l'on se concentre sur les mouvements mois-à-mois; consulter, par exemple, Harvey et Chung (2000). Les estimations filtrées du mouvement mois-à-mois de l'ICC sont définies comme étant

$$\Delta_{t|t} = L_{t|t} - L_{t-1|t} + \beta^{08}_{t|t}\ \delta^{08}_{t} - \beta^{08}_{t-1|t}\ \delta^{08}_{t-1} + \beta^{11}_{t|t}\ \delta^{11}_{t} - \beta^{11}_{t-1|t}\ \delta^{11}_{t-1}, \tag{4.1}$$

où la notation $\Theta_{t|t'}$ désigne l'estimation de la variable d'état Θ pour la période t sachant les données observées jusqu'à la période t'. La valeur aberrante ponctuelle en 2007(9) est, évidemment, supprimée du signal. Qui plus est, les coefficients de régression sont invariants dans le temps. Par conséquent, $\beta^{x}_{t|t} = \beta^{x}_{t-1|t}$ pour $x =$ 08 et 11. Puisque $t =$ 2008(9) et $t =$ 2011(9) sont les mois où δ^{08}_{t} et δ^{11}_{t} changent de valeur, l'expression (4.1) peut être simplifiée en

$$\Delta_{t|t} = L_{t|t} - L_{t-1|t} + \beta^{08}_{t|t}\ \tilde{\delta}^{08}_{t} + \beta^{11}_{t|t}\ \tilde{\delta}^{11}_{t}, \tag{4.2}$$

avec $\tilde{\delta}^{08}_{t} = 1$ si $t =$ 2008(9) et $\tilde{\delta}^{08}_{t} = 0$ pour toutes les autres périodes et $\tilde{\delta}^{11}_{t} = 1$ si $t =$ 2011(9) et $\tilde{\delta}^{11}_{t} = 0$ pour toutes les autres périodes. Les estimations lissées pour le mouvement mois-à-mois de l'ICC sont définies comme étant

$$\Delta_{t|T} = L_{t|T} - L_{t-1|T} + \beta^{08}_{t|T}\ \tilde{\delta}^{08}_{t} + \beta^{11}_{t|T}\ \tilde{\delta}^{11}_{t}. \tag{4.3}$$

Pour comparer les mouvements mois-à-mois fondés sur (4.2) et (4.3) aux estimations directes, les effets saisonniers lissés en (3.8) sont soustraits des estimations directes. Les erreurs-types des estimations directes ne sont pas corrigées pour cet ajustement.

La figure 4.5 compare les estimations directes du mouvement mois-à-mois aux estimations lissées (graphique supérieur) et aux estimations filtrées (graphique du milieu) obtenues au moyen du modèle de séries chronologiques. Le graphique inférieur compare les erreurs-types des estimations lissées, filtrées et directes. Les estimations filtrées, et en particulier les estimations lissées du mouvement mois-à-mois, ont un patron plus stable que les estimations directes, ce que reflètent également les erreurs-types. Les fortes corrélations positives des états de la composante tendance entre périodes subséquentes produisent des erreurs-types pour les estimations filtrées et lissées du mouvement mois-à-mois qui sont clairement plus petites que pour l'estimateur direct. Font exception les deux périodes où un changement de niveau est nécessaire. L'introduction d'un changement de niveau accroît le niveau d'incertitude pendant une courte période.

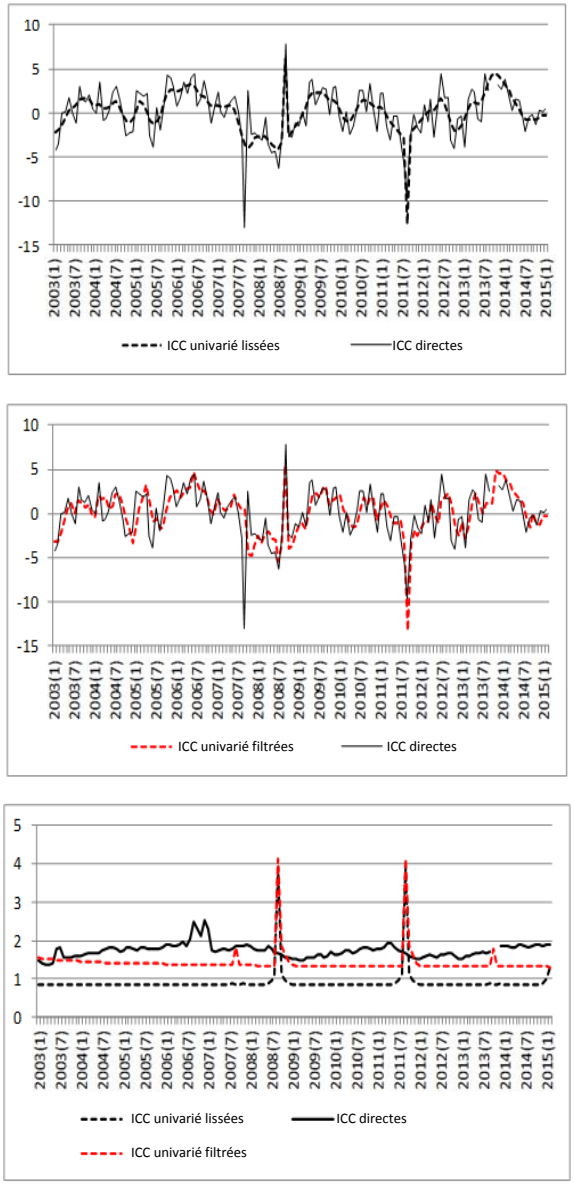

Figure 4.5 Comparaison des estimations du modèle univarié et des estimations directes du mouvement mois-à-mois. Graphique supérieur : estimations lissées, graphique du milieu : estimations filtrées, graphique inférieur : erreurs-types.

La réduction de l'erreur-type est mesurée par l'écart moyen relatif des erreurs-types (EMRET) et, pour les estimations filtrées, est définie comme étant $\text{EMRET} = 100/(T-t')*\sum_{t=t'}^{T}\left[\text{et}\left(\hat{\Delta}_t\right)-\text{et}\left(\Delta_{t|t}\right)\right]\Big/\text{et}\left(\hat{\Delta}_t\right)$, avec et $\left(\hat{\Delta}_t\right)$ l'erreur-type pour l'estimation directe du mouvement mois-à-mois. Pour les estimations lissées, EMRET s'obtient en remplaçant et $\left(\Delta_{t|t}\right)$ par et $\left(\Delta_{t|T}\right)$. Durant la période observée à partir de 2003(1), l'EMRET pour les estimations lissées égale 47 % et pour les estimations filtrées, 17 %.

4.2 Modèle bivarié pour les séries de l'ICC et de l'IMS

À la présente section, nous appliquons le modèle bivarié (3.9) proposé à la section 3.2 aux séries de l'ICC et de l'IMS, qui sont disponibles pour la période allant de juin 2010 à mars 2015. Notons que les composantes de série chronologique pour l'ICC sont réestimées en utilisant la série plus courte. Le tableau 4.2 donne les estimations du maximum de vraisemblance des hyperparamètres. Le modèle décèle une forte corrélation positive de l'ordre de 0,92 entre les perturbations de pente de l'ICC et de l'IMS. Cependant, rien n'indique que les deux tendances sont cointégrées et partagent une tendance commune. Un test du rapport de vraisemblance est appliqué pour déterminer le degré de signification de la corrélation entre les perturbations de pente dans le modèle bivarié. Quand le paramètre de corrélation est fixé à zéro, la log-vraisemblance diminue, passant de -229,9 à -233,9. La valeur p du test du rapport de vraisemblance correspondant égale 0,0047, ce qui indique que la corrélation entre les tendances des deux séries diffère de manière nettement significative de zéro et ne doit pas être supprimée du modèle bivarié. Quand le paramètre de corrélation est fixé à un (en choisissant d_2 dans (3.10) égal à zéro), la log-vraisemblance passe de -229,9 à -242,1. La valeur p du test du rapport de vraisemblance correspondant avec un degré de liberté égale zéro, ce qui indique que les tendances ne sont pas cointégrées.

Tableau 4.2
Estimations du maximum de vraisemblance des hyperparamètres du modèle bivarié de l'ICC et de l'IMS

Écart-type	Estimation du MV
Tendance de l'ICC $(\sigma_{\eta I})$	1,25
Composante saisonnière de l'ICC (σ_{ω})	7,5E-6
Tendance de l'IMS $(\sigma_{\eta X})$	0,25
Équation de mesure de l'ICC $(\sigma_{\upsilon I})$	2,68
Équation de mesure de l'IMS $(\sigma_{\upsilon X})$	0,84
Corrélation des tendances de l'ICC et de l'IMS (ρ_{η})	0,92

À la figure 4.6, nous comparons les estimations lissées de la pente de l'ICC (axe des x) et de l'IMS (axe des y) sous le modèle sans corrélation, le modèle avec une estimation du MV pour la corrélation $\left(\rho_{\eta}=0{,}92\right)$ et le modèle à tendance commune avec $\rho_{\eta}=1{,}0$. Le modèle avec pentes non corrélées révèle une corrélation nettement positive entre les pentes si les deux séries sont estimées indépendamment (graphique de gauche de la figure 4.6). Cette corrélation apparaît dans le modèle qui tient compte de la

corrélation (graphique du milieu de la figure 4.6). Cependant, il existe manifestement un écart entre les pentes des deux séries, ce que l'on peut constater en comparant le graphique du modèle avec une corrélation estimée par le MV (graphique du milieu de la figure 4.6) au graphique du modèle avec un facteur commun (graphique de droite de la figure 4.6).

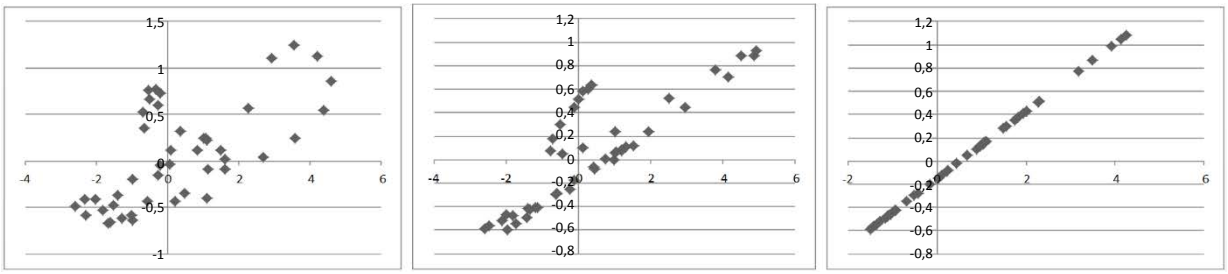

Figure 4.6 Nuages de points des pentes lissées de l'ICC (axe des x) et de l'IMS (axe des y) pour un modèle sans corrélation (graphique de gauche), avec corrélation estimée par le MV (graphique du milieu) et corrélation fixée à un (graphique de droite).

La figure 4.7 donne une comparaison de la série observée de l'IMS avec la tendance lissée obtenue sous le modèle bivarié. La figure 4.8 donne une comparaison des estimations directes de la série de l'ICC avec la tendance lissée plus la valeur aberrante sous le modèle univarié et le modèle bivarié. Comme le montre la figure 4.8, le niveau et l'évolution des estimations lissées de la série de l'ICC sont presque identiques sous les modèles univarié et bivarié.

À la figure 4.9, nous comparons les erreurs-types des estimations directes de la série de l'ICC avec la tendance lissée plus la valeur aberrante sous le modèle univarié et le modèle bivarié. Pour que la comparaison soit juste, les résultats pour le modèle univarié et le modèle bivarié sont fondés sur des séries de même longueur. Pour cela, le modèle univarié est réestimé en se basant sur la série allant de juin 2010 à mars 2015. La figure 4.9 révèle que l'erreur-type sous le modèle bivarié est légèrement plus petite que l'erreur-type sous le modèle univarié si les deux modèles sont appliqués à des séries de longueur égale, comme prévu étant donné la forte corrélation positive significative entre les termes de perturbation des tendances des deux séries. En revanche, si le modèle univarié est appliqué à la série disponible à partir de décembre 2000, les erreurs-types pour les estimations lissées sous le modèle univarié sont légèrement plus faibles que sous le modèle bivarié comme le montre la figure 4.10.

Pour résumer, il apparaît que le modèle bivarié décèle une forte corrélation entre les séries de l'ICC et de l'IMS. L'utilisation de la série de l'IMS comme série auxiliaire améliore légèrement la précision des estimations fondées sur un modèle pour l'ICC. Puisque la série de l'ICC est neuf années plus longue que celle de l'IMS, dans le modèle univarié, la plus grande précision obtenue avec la série auxiliaire est compensée par l'information supplémentaire disponible avant 2010 dans la série de l'ICC.

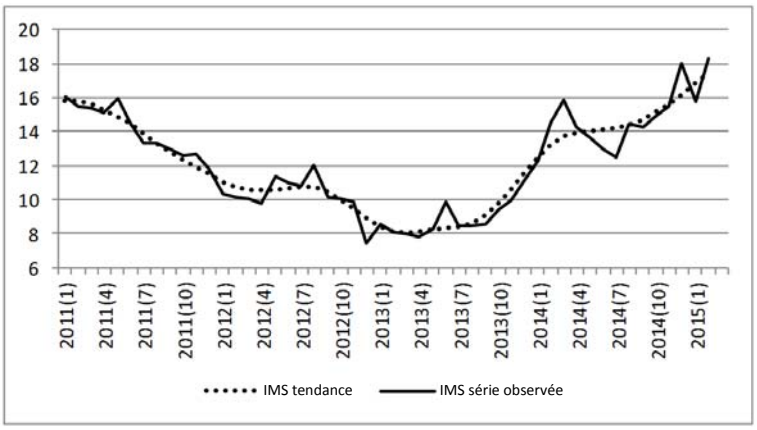

Figure 4.7 Série observée et tendance lissée de l'IMS.

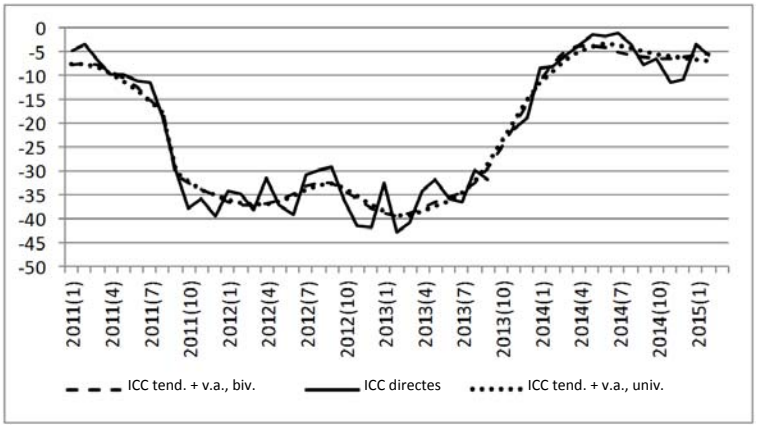

Figure 4.8 Comparaison des estimations directes et de la tendance lissée plus la valeur aberrante pour l'ICC sous les modèles bivarié et univarié de l'ICC.

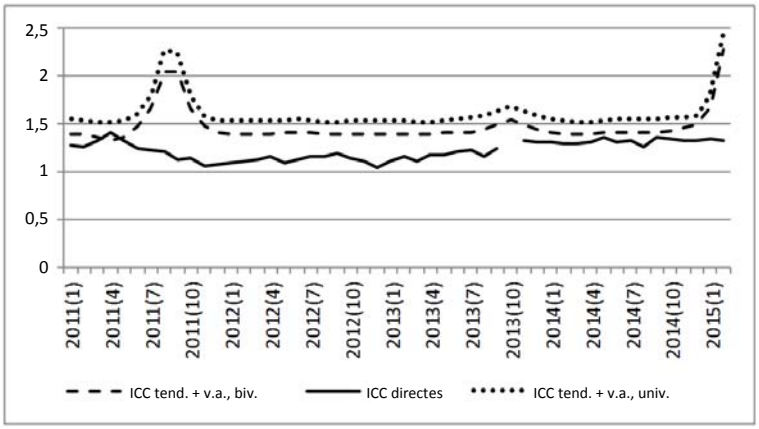

Figure 4.9 Comparaison des erreurs-types des estimations directes de l'ICC et de la tendance lissée plus valeur aberrante sous les modèles bivarié et univarié pour l'ICC si les deux modèles sont appliqués à des séries de même longueur (juin 2010 à mars 2015).

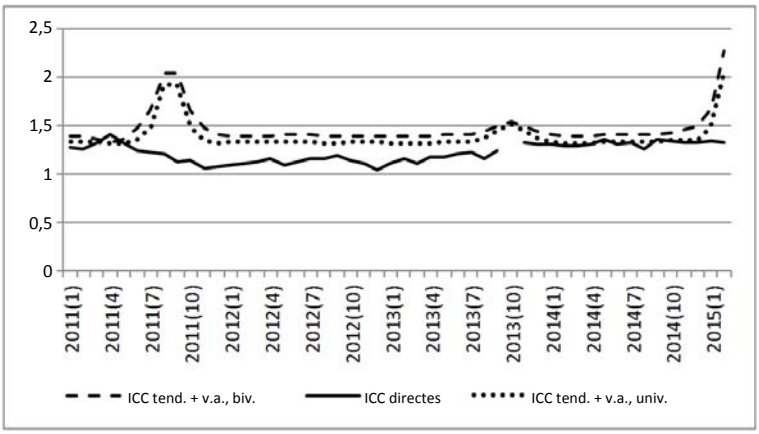

Figure 4.10 Comparaison des erreurs-types des estimations directes de l'ICC et de la tendance lissée plus valeur aberrante sous les modèles bivarié et univarié pour l'ICC si le modèle univarié est appliqué à la série complète de l'ICC (décembre 2000).

Le graphique supérieur de la figure 4.11 compare les estimations directes du mouvement mois-à-mois aux estimations lissées obtenues au moyen des modèles univarié et bivarié de séries chronologiques (tous deux fondés sur la série observée à partir de juin 2010). Le graphique inférieur donne une comparaison des erreurs-types de ces estimations. Durant la période observée à partir de 2011(1), l'EMRET pour les estimations lissées sous le modèle univarié vaut 39 % et sous le modèle bivarié, 43 %. L'EMRET pour les estimations filtrées sous le modèle univarié vaut 7 % et sous le modèle bivarié, 14 %. Comme dans le cas du modèle univarié, l'approche de modélisation de séries chronologiques produit des estimations plus stables et plus précises du mouvement mois-à-mois. L'utilisation de la série de l'IMS améliore légèrement la précision des mouvements mois-à-mois comparativement au modèle univarié.

Une fois que l'estimation directe de l'ICC pour le mois t devient disponible, la valeur ajoutée de la série de l'IMS est limitée à améliorer l'estimation de la série chronologique pour l'ICC pour le mois t. Toutefois, un inconvénient des sondages est que les données sont généralement moins à jour que celles provenant des médias sociaux. La valeur ajoutée de l'IMS devient plus évidente lorsque l'on profite de la plus grande fréquence de cette série pour produire des prédictions précoces ou prédictions immédiates pour l'ICC au moyen du modèle espace-état bivarié. Si une première prédiction immédiate pour l'ICC est nécessaire durant le mois t ou directement à la fin de ce mois, le modèle univarié produit seulement une prévision une étape à l'avance. Par contre, dès que les résultats pour la série de l'IMS sont disponibles durant le mois t ou à la fin du mois t, le modèle bivarié exploite la forte corrélation entre les séries pour faire une prédiction plus précise pour l'ICC, déjà avant que l'estimation directe pour le mois t devienne disponible.

Afin d'illustrer la valeur ajoutée qu'offre l'IMS dans une procédure de prédiction immédiate pour l'ICC, au graphique supérieur de la figure 4.12, nous comparons les prévisions une étape à l'avance pour la tendance plus valeur aberrante de la série de l'ICC obtenue avec le modèle univarié à l'estimation obtenue avec le modèle bivarié si l'IMS pour le mois t est disponible, mais que l'estimation directe de l'ICC manque encore. Les estimations lissées pour la tendance plus valeur aberrante de l'ICC obtenues au moyen du modèle univarié sont incluses comme référence. Dans le graphique inférieur de la figure, nous comparons les erreurs-types de ces trois estimations.

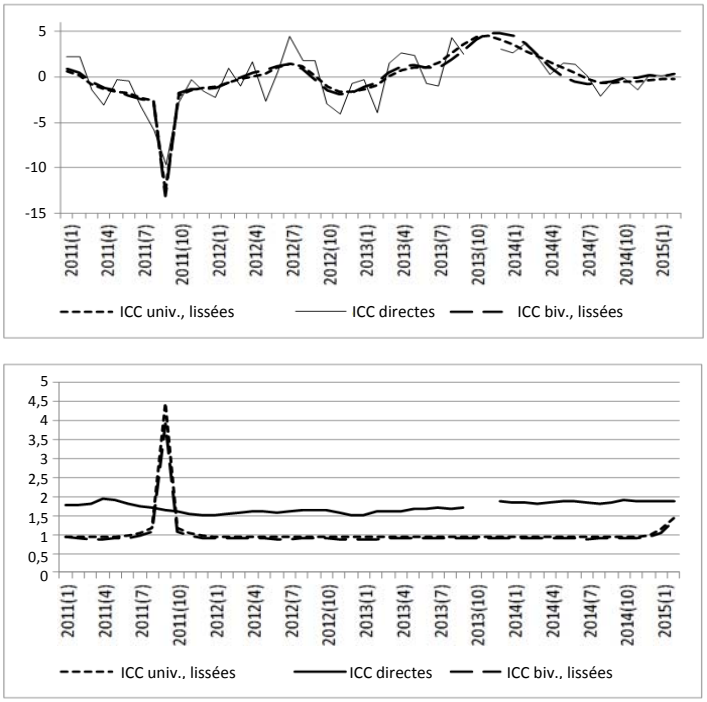

Figure 4.11 Comparaison des estimations sous le modèle bivarié, sous le modèle univarié et directes des mouvements mois-à-mois. Graphique supérieur : estimations lissées, graphique inférieur : erreurs-types.

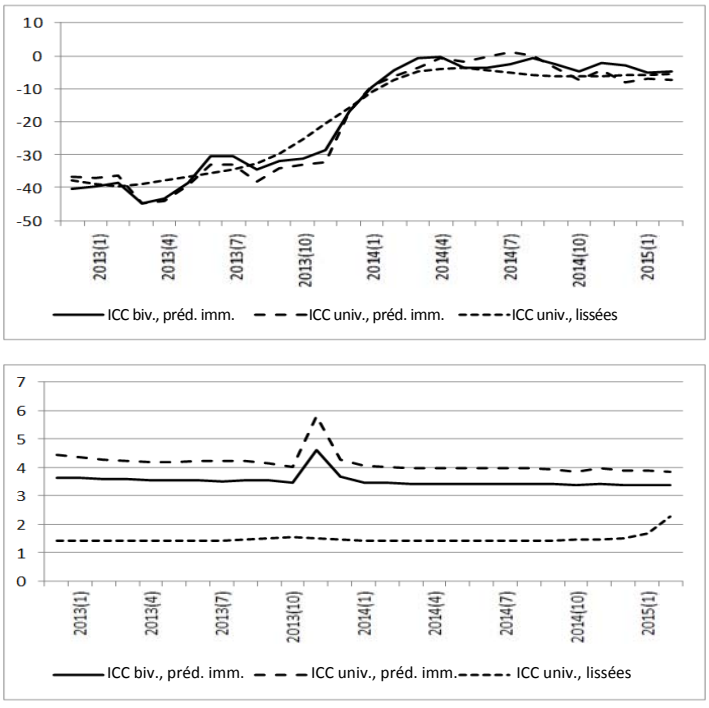

Figure 4.12 Comparaison des estimations pour la tendance plus valeur aberrante de la série de l'ICC; modèle univarié avec prévisions une étape à l'avance (ICC univ., prédiction immédiate), modèle bivarié si l'IMS pour le mois *t* est disponible, mais que l'estimation directe de l'ICC manque (ICC biv., prédiction immédiate) et estimations lissées avec le modèle univarié (ICC univ., lissées). Graphique supérieur : comparaison des estimations ponctuelles. Graphique inférieur : comparaison des erreurs-types.

Lorsque les estimations lissées issues du modèle univarié sont prises comme référence, nous utilisons comme mesure de la grandeur de la révision l'écart moyen absolu relatif (EMAR) entre les prédictions immédiates et les estimations lissées, lequel est défini par $\text{EMAR} = 100/(T-t') * \sum_{t=t'}^{T} \left|\theta_{t|T} - \theta_{t|t-1}\right| \Big/ \left|\theta_{t|T}\right|$, où $\theta_t = L_t + \beta^{11}\delta_t^{11}$ désigne la tendance plus valeurs aberrantes de la série de l'ICC. En se basant sur les mois observés à partir de $t =$ 2013(1), l'EMAR pour les prédictions immédiates obtenues au moyen du modèle univarié égale 35 % et au moyen du modèle bivarié, 31 %. Cela montre que les révisions sont un peu plus petites, et donc plus stables si l'on utilise les prédictions immédiates pour l'ICC obtenues au moyen du modèle bivarié. La différence de précision entre les prédictions immédiates obtenues avec le modèle univarié et le modèle bivarié est mesurée par l'EMRET qui est défini ici comme étant $\text{EMRET} = 100/(T-t') * \sum_{t=t'}^{T} \left[\text{et}\left(\theta_{t|t-1}^{\text{uni}}\right) - \text{et}\left(\theta_{t|t-1}^{\text{biv}}\right)\right] \Big/ \text{et}\left(\theta_{t|t-1}^{\text{biv}}\right)$. En se basant sur les mois observés à partir de $t =$ 2013(1), la différence de précision entre les deux prédictions immédiates selon cet EMRET est égale à 17 %. La figure 4.12, ainsi que l'EMAR et l'EMRET illustrent que l'IMS améliore la stabilité et la précision des prédictions immédiates pour l'ICC.

5 Discussion

Pendant des décennies, les instituts nationaux de statistique se sont appuyés sur l'échantillonnage probabiliste pour produire les statistiques officielles. Cette approche est fondée sur une théorie éprouvée en vue de faire des inférences statistiques valides pour de grandes populations cibles finies en partant d'échantillons aléatoires relativement petits. Au cours des dernières décennies, de plus en plus de sources de données de rechange, comme les données administratives et les mégadonnées, sont devenues disponibles et la question qui se pose est celle de savoir comment utiliser ces sources de données pour produire les statistiques officielles. Un important problème concerne la généralisation des résultats obtenus au moyen de ces sources à une population cible finie. Puisque le processus de génération des données est habituellement inconnu, la façon de faire des inférences valides au moyen de ces sources de données n'est pas évidente.

Le présent article traite de la façon d'utiliser les sources de données administratives et de mégadonnées pour produire des statistiques officielles. L'approche la plus extrême consiste à remplacer les données d'enquête par des données provenant de sources de rechange reliées, en courant le risque d'introduire, par exemple, un biais de sélection. Puisque la plupart des enquêtes sont réalisées de manière répétée, nous proposons une approche de modélisation de séries chronologiques pour déterminer dans quelle mesure les sources de données de rechange reliées reflètent une même évolution que la série obtenue au moyen d'une enquête répétée. Un modèle espace-état multivarié permet de modéliser la corrélation entre les composantes sous-jacentes non observées des deux séries. Le cas où les composantes du modèle de séries chronologiques sont cointégrées constitue un fort indice que les deux sources de données ont pour moteur le même facteur sous-jacent. Le cas échéant, cela permettrait d'arguer qu'une source de rechange peut remplacer les enquêtes existantes, puisqu'elles reflètent la même évolution d'un processus, généralement à un niveau différent.

La théorie qui sous-tend l'échantillonnage probabiliste pour l'inférence sur des populations finies est plus solide que l'application du concept de cointégration. Les séries obtenues à partir des médias sociaux ou de Google Trends sont choisies de manière à maximiser la corrélation avec la série issue de l'enquête par sondage et ne mesurent pas nécessairement le même concept que l'enquête. Rien ne garantit que cette corrélation est basée sur une causalité réelle ni que la corrélation persistera dans l'avenir. En revanche, la théorie de l'échantillonnage repose sur une théorie mathématique rigoureuse montrant qu'une stratégie

d'échantillonnage correcte, c'est-à-dire la combinaison appropriée d'un échantillon probabiliste et d'un estimateur approximativement sans biais sous le plan, aboutit à des inférences statistiques valides pour les populations cibles choisies.

Même dans le cas de séries cointégrées, une évaluation approfondie du modèle, par exemple une forme de validation croisée, sera nécessaire pour avoir la certitude que la source de données de rechange est un remplacement valide. Consulter aussi, dans ce contexte, Eichler (2013) pour une discussion de l'utilisation de la causalité de Granger pour l'inférence causale dans le cas de séries chronologiques multiples. Au lieu de remplacer une enquête périodique par des sources de données reliées, ces dernières peuvent être utilisées comme série auxiliaire dans une approche de modélisation de séries chronologiques multivariée en vue d'améliorer la précision des estimations directes ou des estimations du mouvement période-à-période des estimations directes obtenues au moyen d'une enquête périodique. Un autre avantage important des sources de mégadonnées consiste à profiter de la plus grande fréquence de ces sources de données pour faire des prédictions précoces ou des prédictions immédiates plus précises quand, en temps réel, les estimations issues de l'enquête ne sont pas encore disponibles, mais que la covariable l'est déjà. Le modèle de séries chronologiques appliqué dans le présent article, proposé pour la première fois par Harvey et Chung (2000), représente une approche générique pour une procédure d'estimation fondée sur un modèle pour des enquêtes périodiques. Naturellement, les enquêtes par sondage posent aussi des problèmes. À titre d'exemple, les taux de réponse continuellement à la baisse et les modes de collecte des données ne permettant pas d'atteindre la population cible produisent aussi un biais de sélection. Dans cette situation, la cointégration avec une série reliée dérivée des médias sociaux pourrait être un signe qu'il existe des similarités entre le biais de sélection dans les sources de mégadonnées non probabilistes et le biais de couverture et de sélection de la non-réponse dans un sondage, comme l'ont souligné Baker et coll. (2013).

Dans l'application à l'ICC, l'approche de modélisation de séries chronologiques ne réduit pas la variance de l'estimateur direct quand elle est utilisée pour produire des estimations de niveau. Cela tient au fait que l'erreur-type du modèle de séries chronologiques reflète l'erreur d'échantillonnage et le bruit blanc du paramètre de population. L'erreur-type de l'estimateur direct reflète uniquement l'erreur d'échantillonnage. Dans le cas de l'ICC, la composante de variance du bruit blanc du paramètre de population est aussi grande que la variance de l'erreur d'échantillonnage. L'approche du modèle espace-état est néanmoins utile pour produire les chiffres officiels de l'ICC, puisqu'elle filtre une tendance plus stable de l'opinion des personnes interrogées au sujet du climat économique à partir de la série observée d'estimations directes. Cependant, la situation change si le modèle de séries chronologiques est utilisé pour estimer les mouvements mois-à-mois. Les estimations de la tendance stable résultent d'une forte corrélation positive entre les estimations de périodes subséquentes. Par conséquent, les erreurs-types du mouvement mois-à-mois obtenues au moyen du modèle de séries chronologiques sont nettement plus petites que celles des estimations directes. Les erreurs-types des mouvements mois-à-mois lissés sont environ 47 % plus petites que celles des estimations directes. Les erreurs-types des estimations filtrées sont environ 17 % plus petites que celles des estimations directes.

L'utilisation de l'IMS comme série auxiliaire dans un modèle espace-état bivarié réduit légèrement l'erreur-type des estimations modélisées de l'ICC. Cependant, puisque la série de l'IMS disponible est relativement courte, la réduction obtenue au moyen de cette série auxiliaire ne surpasse pas la perte de l'information de la série de l'ICC qui a été observée durant la période avant que l'IMS soit disponible. Toutefois, puisque les deux séries présentent une même évolution et que les données des médias sociaux

sont disponibles rapidement, l'IMS s'est avéré utile comme série auxiliaire dans le modèle bivarié pour produire des prédictions immédiates plus fiables pour l'ICC en temps réel, au moment où l'IMS devient disponible tandis que l'ICC ne l'est pas encore. Dans cette application, l'IMS réduit d'environ 17 % les erreurs-types de l'ICC dans une procédure de prédiction immédiate.

D'aucuns se demanderont si, dans sa mise en œuvre actuelle, l'IMS mesure le même concept que celui qu'essaie de mesurer l'ICC et comment on pourrait tirer parti de toutes les possibilités qu'offrent les données des médias sociaux et d'autres sources de mégadonnées pour obtenir de meilleures mesures de la confiance des consommateurs que celles produites par les ICC et IMS actuels. Au lieu de construire un indice basé sur les médias sociaux en prenant la différence entre les messages jugés positifs et négatifs, on pourrait construire un IMS en examinant les concepts qui sous-tendent les questions utilisées pour l'ICC. Par exemple, si l'on mesure la confiance des consommateurs par le nombre d'achats de biens coûteux au cours des 12 derniers mois ou par la tendance des ménages à acheter des biens coûteux, les indices fondés sur les médias sociaux devraient mesurer la recherche de ce genre de biens sur Internet (automobiles, maisons, produits blancs, etc.), ainsi que les achats réels de ces biens durant les mois précédents. Le grand avantage de cette approche est qu'elle permet de mesurer directement le comportement actuel des ménages, alors qu'une enquête le mesure indirectement, ce qui induit une plus grande erreur de mesure. Cela pourrait aboutir à des séries cointégrées qui mesurent des concepts similaires et améliorent davantage ou même remplacent l'ICC.

Remerciements

Les auteurs sont reconnaissants au rédacteur associé et aux examinateurs pour leur lecture attentive d'une version antérieure de cet article et pour leurs commentaires constructifs ayant contribué à améliorer significativement le contenu de l'article. Les opinions exprimées dans le présent article sont celles des auteurs et ne reflètent pas forcément les politiques de *Statistics Netherlands*.

Annexe A

Diagnostics du modèle

Tableau A.1
Modèle univarié (3.8) pour l'ICC 172-24 observations

Diagnostic	Valeur	Valeur *p*	IC à 95 %	
			Inf.	Sup.
Log-vraisemblance	-464			
Moyenne des innovations standardisées	0,0152			
Variance des innovations standardisées	1,0851			
Asymétrie des innovations standardisées	0,0276			
Aplatissement des innovations standardisées	2,8901			
Test de Bowman-Shenton[1] sur la normalité des innovations standardisées	0,0926	0,955		
Test de Ljung-Box[2] sur la corrélation sériale des innovations standardisées	24,108	0,287		
Test de Durban-Watson[3] sur la corrélation sériale des innovations standardisées $(T = 148)$	2,082		1,68	2,32
Test F[4] sur l'hétéroscédasticité des innovations standardisées $(df_{\text{num}} = df_{\text{denom}} = 60)$	0,913		0,60	1,67

1) Statistique de Bowman-Shenton : loi du χ^2_2.
2) Statistique du test de Ljung-Box pour la corrélation sériale dans les 24 premiers décalages temporels : loi du χ^2_{21}.
3) Statistique du test de Durban-Watson approximée par $N(2, 4/T)$.
4) Statistique F : loi de $F^{df_{\text{num}}}_{df_{\text{denom}}}$.

Tableau A.2
Modèle bivarié (3.9) pour l'ICC 57-24 observations

Diagnostic	Valeur	Valeur *p*	IC à 95 %	
			Inf.	Sup.
Log-vraisemblance	-230			
Moyenne des innovations standardisées	-0,0872			
Variance des innovations standardisées	0,9777			
Asymétrie des innovations standardisées	0,0982			
Aplatissement des innovations standardisées	2,545			
Test de Bowman-Shenton[1] sur la normalité des innovations standardisées	0,3382	0,844		
Test de Ljung-Box[2] sur la corrélation sériale des innovations standardisées	18,060	0,645		
Test de Durban-Watson[3] sur la corrélation sériale des innovations standardisées ($T = 33$)	2,133		1,32	2,68
Test F[4] sur l'hétéroscédasticité des innovations standardisées ($df_{\text{num}} = df_{\text{denom}} = 15$)	0,783		0,35	2,86

Tableau A.3
Modèle bivarié (3.9) pour l'IMS 57-12 observations

Diagnostic	Valeur	Valeur *p*	IC à 95 %	
			Inf.	Sup.
Log-vraisemblance	-230			
Moyenne des innovations standardisées	0,0954			
Variance des innovations standardisées	1,0437			
Asymétrie des innovations standardisées	-0,1311			
Aplatissement des innovations standardisées	2,5331			
Test de Bowman-Shenton[1] sur la normalité des innovations standardisées	0,5377	0,764		
Test de Ljung-Box[2] sur la corrélation sériale des innovations standardisées	24,208	0,283		
Test de Durban-Watson[3] sur la corrélation sériale des innovations standardisées ($T = 45$)	2,028		1,42	2,58
Test F[4] sur l'hétéroscédasticité des innovations standardisées ($df_{\text{num}} = df_{\text{denom}} = 20$)	0,329		0,41	2,46

1) Statistique de Bowman-Shenton : loi du χ^2_2.
2) Statistique du test de Ljung-Box pour la corrélation sériale dans les 24 premiers décalages temporels : loi du χ^2_{21}.
3) Statistique du test de Durban-Watson approximée par $N(2, 4/T)$.
4) Statistique F : loi de $F^{df_{\text{num}}}_{df_{\text{denom}}}$.

Bibliographie

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. et Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-143, publié pour la première fois en ligne le 26 septembre 2013, doi:10.1093/jssam/smt008.

Bell, W.R. (2005). Some considerations of seasonal adjustment variances. Census Bureau. Article accessible à l'adresse https://www.census.gov/ts/papers/jsm2005wrb.pdf.

Bell, W.R., et Hillmer, S.C. (1990). Estimation dans les enquêtes à passages répétés au moyen de séries chronologiques. *Techniques d'enquête*, 16, 2, 205-227. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1990002/article/14535-fra.pdf.

Binder, D.A., et Dick, J.P. (1989). Enquêtes répétées – Modélisation et estimation. *Techniques d'enquête*, 15, 1, 31-48. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1989001/article/14579-fra.pdf.

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.

Blight, B.J.N., et Scott, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 35, 61-66.

Blumenstock, J., Cadamuro, G. et On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350, 1073-1076.

Bollineni-Balabay, O., van den Brakel, J.A. et Palm, F. (2015). Multivariate state-space approach to variance reduction in series with level and variance breaks due to sampling redesigns. Accepté pour publication dans *Journal of the Royal Statistical Society, Series A*.

Bollineni-Balabay, O., van den Brakel, J.A. et Palm, F. (2017). 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. *Techniques d'enquête*, 43, 1, 47-75. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2017001/article/14819-fra.pdf.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin de l'Institut International de Statistique*, 22, Supplément au livre 1, 6-62.

Buelens, B., Burger, J. et van den Brakel, J.A. (2015). Predictive inference for non-probability samples: A simulation study. Document de discussion 2015-13, Statistics Netherlands, Heerlen.

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

Daas, P., et Puts, M. (2014a). Big data as a source of statistical information. *The Survey Statistician*, 69, 22-31.

Daas, P., et Puts, M. (2014b). Social media sentiment and consumer confidence. European Central Bank Statistics paper series No. 5, Frankfurt Allemagne.

Doornik, J.A. (2009). *An Object-oriented Matrix Programming Language Ox 6*. Londres: Timberlake Consultants Press.

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

Eichler, M. (2013). Causal inference with multiple time series: Principles and problems. *Philosophical transactions of the Royal Statistical Society A*, 371, édition 1997.

Feder, M. (2001). Time series analysis of repeated surveys: The state-space approach. *Statistica Neerlandica*, 55, 182-199.

Hansen, M.H., et Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics* 14, 333-362.

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

Harvey, A.C., et Chung, C.H. (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, Series A*, 163, 303-339.

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

Koopman, S.J., Harvey, A., Shephard, N. et Doornik, J.A. (2009). *STAMP 8.2*, Londres: Timberlake Consultants Press.

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

Lind, J.T. (2005). Repeated surveys and the Kalman filter. *Econometrics Journal*, 8, 418-427.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Perdreschi, D., Rinzivillo, S., Pappalardo, L. et Gabrielli, L. (2015). Small area model-based estimators using big data sources. *Journal of Official Statistics*, 31, 263-281.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

Pang, B., et Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2, 1-135.

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

Pfeffermann, D., et Burck, L. (1990). Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales. *Techniques d'enquête*, 16, 2, 229-249. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/1990002/article/14534-fra.pdf.

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 Sverchkov, M. (2014). Estimation of mean squared error of X-11-ARIMA and other estimators of time series components. *Journal of Official Statistics*, 30, 811-838.

Pfeffermann, D., et Tiller, R. (2006). Small area estimation with state space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101, 1387-1397.

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 & Economic Statistics*, 16, 339-348.

Rao, J.N.K., et Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics*, 22, 511-528.

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

Scott, A.J., et Smith, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of* the *American Statistical Association*, 69, 674-678.

Scott, A.J., Smith, T.M.F. et Jones, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review/Revue Internationale de Statistique*, 45, 13-28.

Tam, S.-M. (1987). Analysis of repeated surveys using a dynamic linear model. *International Statistical Review/Revue Internationale de Statistique*, 55, 1, 63-73.

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

van den Brakel, J.A., 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.A., 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.

Techniques d'enquête, décembre 2017
Vol. 43, nº 2, pp. 231-254
Statistique Canada, nº 12-001-X au catalogue

Décomposition des inégalités salariales selon le sexe par calage : application à l'Enquête suisse sur la structure des salaires

Mihaela-Catalina Anastasiade et Yves Tillé[1]

Résumé

L'article propose une nouvelle approche de décomposition de l'écart salarial entre les hommes et les femmes fondée sur une procédure de calage. Cette approche généralise deux méthodes de décomposition courantes, qui sont réexprimées en se servant des poids de sondage. La première est la méthode de Blinder-Oaxaca et la seconde est une méthode de repondération proposée par DiNardo, Fortin et Lemieux. La nouvelle approche offre un système de pondération qui nous permet d'estimer des paramètres d'intérêt tels que les quantiles. Une application aux données de l'Enquête suisse sur la structure des salaires illustre l'intérêt de cette approche.

Mots-clés : Blinder-Oaxaca; discrimination salariale selon le sexe; quantiles; repondération; salaires.

1 Introduction

La discrimination salariale peut être fondée sur différents critères, dont le sexe, la race ou la religion. La discrimination salariale selon le sexe a lieu quand un homme et une femme sont rémunérés différemment pour un travail qui requiert les mêmes qualifications ou qui implique une productivité identique (voir, par exemple, Neumark, 1988; Gardeazabal et Ugidos, 2005). Puisqu'il faut quantifier la discrimination pour en évaluer l'importance, le sujet a suscité l'intérêt des statisticiens. La technique originale proposée par Blinder (1973) et Oaxaca (1973) consiste à estimer la part de l'écart entre les salaires moyens des hommes et des femmes qui est due à la discrimination. Cependant, en général, la répartition des hommes et des femmes entre les emplois n'est pas uniforme (voir, par exemple Bielby et Baron, 1986). Si les membres d'un de ces deux groupes, habituellement les femmes, sont concentrés dans les emplois faiblement rémunérés, l'écart entre les salaires moyens pourrait ne pas être très pertinent. Donc, au lieu d'analyser le niveau de discrimination dans les salaires moyens, il serait peut-être intéressant de voir si la discrimination se produit uniformément dans tous les types d'emplois. Une bibliographie détaillée des différents articles statistiques consacrés à l'estimation de la discrimination peut être consultée dans Fortin, Lemieux et Firpo (2011).

Bien que les méthodes de décomposition offertes dans la littérature soient nombreuses, deux d'entre elles seulement seront discutées dans le présent article. Ces deux méthodes ne sont pas présentées dans leur forme originale, mais plutôt en tenant compte des poids de sondage. Il s'agit de la méthode de Blinder-Oaxaca (ci-après nommée BO) et la méthode semi-paramétrique élaborée par DiNardo, Fortin et Lemieux (1996) (ci-après nommée DFL). La méthode BO originale analyse la différence entre les salaires moyens des hommes et les salaires moyens des femmes. Cependant, elle ne permet pas d'analyser l'écart salarial pour d'autres paramètres, comme les quantiles. La méthode DFL originale résout ce problème. Son point de départ est un modèle logistique dans lequel, pour chaque observation, la probabilité d'être un homme ou une femme est modélisée comme une fonction des caractéristiques observées. Le ratio de ces probabilités est utilisé pour

1. Mihaela-Catalina Anastasiade et Yves Tillé, Institut de statistique, Université de Neuchâtel, 51 Avenue de Bellevaux, 2000 Neuchâtel, Suisse. Courriel : mihaela.anastasiade@unine.ch; yves.tille@unine.ch.

construire un facteur de repondération. L'objectif de la méthode est de rapprocher la distribution des caractéristiques des femmes de la distribution des caractéristiques des hommes. En disposant de distributions similaires des caractéristiques, il est possible d'obtenir une estimation du niveau de discrimination pour d'autres paramètres que la moyenne. Cependant, la variance du facteur de repondération peut être grande dans les cas où une ou plusieurs caractéristiques sont de bons prédicteurs du sexe. En outre, la distribution repondérée des caractéristiques des femmes peut ne pas concorder avec la distribution des caractéristiques des hommes. Nous abordons les problèmes associés aux deux méthodes par une approche de calage. L'idée qui sous-tend le calage est la même que celle qui sous-tend la méthode DFL. Elle consiste à rapprocher la distribution des caractéristiques des femmes de celle des hommes, afin d'estimer le niveau de discrimination tout le long de la distribution des salaires.

La présentation de l'article est la suivante. À la section 2, nous définissons la notation et à la section 3, nous réexprimons la décomposition BO en nous servant de données d'enquête. Les poids de sondage sont pris en compte afin de corriger l'écart entre l'échantillon et la population d'intérêt. Par conséquent, la décomposition sera nommée « BO pondérée ». Nous présentons aussi le concept clé de la distribution contrefactuelle des salaires des femmes. Elle est définie comme étant la distribution des salaires des femmes si celles-ci avaient les mêmes caractéristiques que les hommes. Ensuite, nous discutons de l'utilisation de la distribution contrefactuelle des salaires dans la décomposition de l'écart salarial. À la section 4, nous développons la méthode DFL, de nouveau en utilisant les poids de sondage. Puisque la méthode originale n'inclut pas ces poids, nous nommons cette nouvelle méthode « DFL pondérée ». Ensuite, à la section 5, nous proposons une nouvelle approche pour calculer la distribution contrefactuelle des salaires, en utilisant la méthode de calage (Deville et Särndal, 1992). Nous discutons de deux cas particuliers de calage, à savoir le calage linéaire et le calage par raking ratio (ou ratissage croisé). Le premier cas produit le même résultat que la méthode BO pondérée pour les salaires moyens. Le deuxième cas présente une approche similaire à la méthode DFL pondérée, mais sans émettre l'hypothèse d'un modèle logistique. Autrement dit, la technique proposée peut être considérée comme une généralisation des deux méthodes susmentionnées. La section 6 comprend un aperçu de l'ensemble de données utilisé, ainsi que les statistiques descriptives sur les salaires observés. Nous donnons aussi une brève description du modèle utilisé et des résultats obtenus en appliquant les méthodes discutées. Enfin, à la section 7, nous résumons les conclusions et à l'annexe B, nous présentons le calcul de la variance du salaire contrefactuel.

2 Problème et notation

La question d'intérêt est l'estimation des écarts salariaux entre les hommes et les femmes, plus précisément, quelle part de ces écarts est attribuable à la discrimination. Supposons qu'il existe une population finie U de taille N pouvant être divisée en deux sous-populations, les femmes (F) et les hommes (M), et que nous désignons par $U_h, h \in \{F, M\}$, de taille N_h. En outre, nous tirons de U un échantillon aléatoire S qui contient à la fois des hommes et des femmes. L'échantillon S est sélectionné selon un plan d'échantillonnage $p(s) = \Pr(S = s)$ pour tout $s \subset U$, où

$$p(s) \geq 0 \quad \text{et} \quad \sum_{s \subset U} p(s) = 1.$$

L'échantillon S peut être divisé en deux sous-échantillons, $S_h, h \in \{F, M\}$, de femmes et d'hommes, tels que $S = \cup S_h$. La variable d'intérêt, désignée par y, est ici le logarithme du salaire. Les totaux de la variable d'intérêt dans les deux sous-populations sont donnés par

$$Y_h = \sum_{k \in U_h} y_k, h \in \{F, M\},$$

où y_k est le logarithme du salaire du k^{e} individu. Puisque l'on n'observe pas toutes les unités des sous-populations, les totaux peuvent être estimés par

$$\hat{Y}_h = \sum_{k \in S_h} d_k y_k, h \in \{F, M\},$$

où d_k est un poids de sondage affecté à la k^{e} unité de l'échantillon. Les poids de sondage sont obtenus après plusieurs traitements statistiques (par exemple, ajustement pour la non-réponse).

Les moyennes de population des logarithmes des salaires sont données par

$$\bar{Y}_h = \frac{1}{N_h} \sum_{k \in U_h} y_k, h \in \{F, M\},$$

et peuvent être estimées par

$$\hat{\bar{Y}}_h = \frac{\sum_{k \in S_h} d_k y_k}{\sum_{k \in S_h} d_k}, h \in \{F, M\}.$$

En outre, supposons que, pour chaque k^{e} individu dans l'un ou l'autre des deux sous-échantillons, il existe un vecteur de p variables auxiliaires désigné par

$$\mathbf{x}_k = \left(x_{k1}, \ldots, x_{kj}, \ldots, x_{kp}\right)^\top \in \mathbb{R}^p.$$

Ce vecteur est supposé connu pour chaque unité sélectionnée dans l'échantillon. Les variables auxiliaires contiennent certaines caractéristiques de l'individu, par exemple l'âge, le niveau d'études ou le niveau d'ancienneté. Il peut exister des variables qualitatives ou quantitatives, de sorte que x_{kj} peut être une variable catégorielle ou une quantité. En outre, supposons que la première variable auxiliaire est une constante, c'est-à-dire $x_{k1} = 1,$ pour tout $k \in U$.

Les totaux de ces variables auxiliaires au niveau de la sous-population sont donnés par

$$\mathbf{X}_h = \sum_{k \in U_h} x_k, h \in \{F, M\}.$$

En utilisant les poids d_k définis plus haut, ces deux totaux peuvent être estimés par

$$\hat{\mathbf{X}}_h = \sum_{k \in S_h} d_k \mathbf{x}_k, h \in \{F, M\}.$$

Les vecteurs des valeurs moyennes peuvent être estimés de manière analogue. Les valeurs moyennes au niveau des sous-populations sont données par

$$\bar{\mathbf{X}}_h = \frac{1}{N_h} \sum_{k \in U_h} \mathbf{x}_k, h \in \{F, M\},$$

et estimées par

$$\hat{\bar{\mathbf{X}}}_h = \frac{\sum_{k\in S_h} d_k \mathbf{x}_k}{\sum_{k\in S_h} d_k}, h \in \{F, M\}. \tag{2.1}$$

3 La décomposition BO pondérée

3.1 La décomposition

Partant des conditions établies à la section 2, nous résumons les constatations de Blinder (1973) et d'Oaxaca (1973) dans le contexte de la théorie du sondage, à savoir en utilisant les poids de sondage. Supposons que, dans chaque échantillon, une relation linéaire entre les p caractéristiques disponibles et le logarithme du salaire est appropriée. Une régression est effectuée séparément dans chaque sous-population $U_h, h = \{M, F\}$. Au niveau de la sous-population, les valeurs des coefficients de régression sont données par

$$\boldsymbol{\beta}_h = \left(\sum_{k\in U_h} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k\in U_h} \mathbf{x}_k y_k.$$

Elles peuvent être estimées d'après l'échantillon par

$$\hat{\boldsymbol{\beta}}_h = \left(\sum_{k\in S_h} d_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k\in S_h} d_k \mathbf{x}_k y_k, \tag{3.1}$$

où d_k désigne les poids de sondage. Les coefficients de régression $\hat{\boldsymbol{\beta}}_h$, que nous appelons structure salariale de groupe ou rendement des caractéristiques, représentent la contribution de chaque caractéristique au salaire.

Résultat 1 *Une condition suffisante pour obtenir les égalités suivantes*

$$\bar{Y}_h = \bar{\mathbf{X}}_h^\top \boldsymbol{\beta}_h \quad et \quad \hat{\bar{Y}}_h = \hat{\bar{\mathbf{X}}}_h^\top \hat{\boldsymbol{\beta}}_h$$

est qu'il existe un vecteur $\boldsymbol{\varsigma} \in \mathbb{R}^p$, *tel que* $\boldsymbol{\varsigma}^\top \mathbf{x}_k = 1$, *pour tout* $k \in U_h$.

Puisque nous supposons que $x_{k1} = 1$ pour tout $k \in U$, avec $\boldsymbol{\varsigma}^\top = (1\ 0 \ldots 0)$, l'égalité est toujours vérifiée. La preuve du résultat 1 figure à l'annexe A. En regroupant le résultat susmentionné et les équations (2.1) et (3.1), l'écart moyen entre les salaires des deux groupes peut s'écrire

$$\Delta = \hat{\bar{Y}}_M - \hat{\bar{Y}}_F = \left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F \right)^\top \hat{\boldsymbol{\beta}}_F + \hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F \right). \tag{3.2}$$

L'écart entre les salaires moyens des groupes contient deux éléments, à savoir une partie expliquée, également appelée *effet de composition* $\left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F \right)^\top \hat{\boldsymbol{\beta}}_F$ et une partie inexpliquée, ou *effet de structure* $\hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F \right)$. Le premier effet englobe les différences de caractéristiques entre les deux groupes. Le second représente la différence de rendement des caractéristiques entre les deux groupes, soit la partie qui n'est pas attribuable à des facteurs objectifs (Oaxaca, 1973; Blinder, 1973). On l'obtient en utilisant les caractéristiques comme mesure de substitution de la productivité. L'estimation de l'*effet de structure* est l'élément central du présent article. L'équation (3.2) contient les mêmes éléments que celle proposée par Oaxaca (1973) et Blinder (1973). La méthodologie appliquée pour estimer les valeurs moyennes et les

coefficients diffère de la technique de régression classique. La méthode BO utilise les coefficients de régression estimés obtenus par les moindres carrés ordinaires (MCO) et les vecteurs des valeurs moyennes des variables explicatives observées. L'approche proposée tient compte des poids de sondage. Cependant, la méthode BO pondérée est la même que la méthode BO originale si les poids de sondage valent tous 1.

3.2 Une note sur l'effet de structure

Les deux éléments de l'équation 3.2 ont des noms différents dans les divers textes publiés. Le premier effet, pour lequel nous avons retenu le nom d'*effet de composition* est également nommé *effet de dotation*. Le second, que nous appelons *effet de structure* est également trouvé dans la littérature sous les noms de *résidu inexpliqué, effet du prix, effet du sexe, effet calculé* ou *traitement inégal* (Weichselbaumer et Winter-Ebmer, 2006). En utilisant la méthode BO, l'effet de structure est une estimation du niveau de discrimination. Toutefois, la discrimination est un phénomène complexe qui pourrait ne pas être toujours entièrement observé. Les variables inobservables, le biais de sélection ou certains mécanismes du marché du travail peuvent contribuer à l'accroissement de la part inexpliquée de l'écart salarial. En outre, Weichselbaumer et Winter-Ebmer (2005) notent deux problèmes possibles concernant le modèle choisi. Premièrement, si les caractéristiques choisies dans le modèle linéaire sont elles-mêmes sujettes à la discrimination, alors l'effet de structure résultant sera surestimé. Deuxièmement, si les caractéristiques ne représentent pas une mesure appropriée de la productivité, de nouveau, l'effet de structure pourrait être sous- ou surestimé. Weichselbaumer et Winter-Ebmer (2006) mettent en garde au sujet de la légitimité des caractéristiques en tant qu'indicateurs de la productivité, puisque les salaires pourraient aussi être déterminés par le pouvoir de négociation, les différences de rémunération ou les salaires basés sur le rendement. Cependant, pour simplifier, dans la suite, nous supposerons que ce genre de problème n'existe pas et que l'effet de structure estimé est le résultat de la discrimination sur le marché du travail. En outre, nous n'examinons pas le biais de sélection de l'échantillon ni d'autres mécanismes qui sous-tendent la distribution des hommes et des femmes dans certains emplois.

3.3 La distribution contrefactuelle des salaires

En général, la distribution contrefactuelle des salaires est une distribution artificielle obtenue en utilisant les caractéristiques d'un groupe pour estimer les salaires d'un autre groupe (voir, par exemple, Bourguignon, Ferreira et Leite, 2002). Des exemples de distributions contrefactuelles sont donnés dans DiNardo et coll. (1996) ou DiNardo (2002). Le terme $\hat{\bar{\mathbf{X}}}_M \hat{\boldsymbol{\beta}}_F$ qui figure dans l'équation (3.2) est appelé salaire moyen contrefactuel des femmes. Il est interprété comme étant le salaire moyen estimé des femmes si celles-ci avaient les mêmes caractéristiques moyennes que les hommes et que leur rendement des caractéristiques demeurait inchangé. La distribution contrefactuelle des salaires s'obtient en utilisant les caractéristiques des hommes $(\mathbf{X}_M)$ et la structure des salaires des femmes $(\boldsymbol{\beta}_F)$. Pour ce qui est de l'interprétation, il s'agit de la distribution des salaires des femmes, si celles-ci avaient les mêmes caractéristiques que les hommes.

En utilisant le résultat 1 de la section précédente, le salaire moyen contrefactuel des femmes est égal à

$$\bar{Y}_{F|M} = \bar{\mathbf{X}}_M^\top \boldsymbol{\beta}_F,$$

et est estimé d'après l'échantillon par

$$\hat{\bar{Y}}_{F|M} = \hat{\bar{\mathbf{X}}}_M^\top \hat{\boldsymbol{\beta}}_F,$$

où $\hat{\bar{\mathbf{X}}}_M$ est estimé dans l'équation (2.1) et $\hat{\boldsymbol{\beta}}_F$ représente les coefficients estimés au moyen de l'équation (3.1). Selon cette notation, la décomposition BO donnée en (3.2) se réexprime sous la forme

$$\Delta = \hat{\bar{Y}}_M - \hat{\bar{Y}}_F = \left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F\right)^\top \hat{\boldsymbol{\beta}}_F + \hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F\right) = \left(\hat{\bar{Y}}_{F|M} - \hat{\bar{Y}}_F\right) + \left(\hat{\bar{Y}}_M - \hat{\bar{Y}}_{F|M}\right). \quad (3.3)$$

3.4 Utilisation de la distribution contrefactuelle pour estimer les effets de composition et de structure

Construire le salaire moyen contrefactuel permet d'estimer les deux effets qui constituent l'écart salarial au niveau moyen. Partant de l'équation (3.3), l'effet de composition est égal à

$$\left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F\right)^\top \hat{\boldsymbol{\beta}}_F = \left(\hat{\bar{Y}}_{F|M} - \hat{\bar{Y}}_F\right).$$

L'effet de composition peut être interprété comme la différence entre ce que les femmes pourraient gagner, en moyenne, si elles avaient les caractéristiques des hommes et ce qu'elles gagnent effectivement. Donc, il reflète l'inégalité due aux différences de caractéristiques. L'effet de structure dans l'équation (3.3) est égal à

$$\hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F\right) = \left(\hat{\bar{Y}}_M - \hat{\bar{Y}}_{F|M}\right).$$

L'effet de structure est la différence entre le salaire moyen réel des hommes et ce que les femmes gagneraient si elles avaient les caractéristiques moyennes des hommes et la structure salariale propre à ceux-ci. Les équations susmentionnées expriment les effets de composition et de structure aux niveaux moyens, puisqu'il s'agit de la limite de la méthode BO. À la section suivante, nous présentons une méthode qui permet de construire la distribution contrefactuelle complète. Cela, à son tour, donne la capacité d'estimer les effets de composition et de structure tout le long de la distribution des salaires.

4 La méthode DFL pondérée

4.1 La méthode

La méthode proposée par DiNardo et coll. (1996) fait appel à une fonction de repondération au moyen de laquelle la distribution des caractéristiques des femmes est rendue similaire à la distribution des caractéristiques des hommes. La distribution repondérée est la distribution contrefactuelle des caractéristiques des femmes. La méthode DFL est présentée en utilisant les poids de sondage, afin de tenir compte du plan de sondage.

La fonction de repondération est égale à

$$\psi(\mathbf{x}_k) = \frac{\Pr(D_{Mk} = 1 | \mathbf{x}_k) / \Pr(D_{Mk} = 1)}{\Pr(D_{Mk} = 0 | \mathbf{x}_k) / \Pr(D_{Mk} = 0)},$$

où $D_{Mk} = 1$ si l'individu k est un homme et $D_{Mk} = 0$ sinon, et $\mathbf{x}_k$ est le vecteur des caractéristiques observées pour l'individu k. Manifestement, $\Pr(D_{Mk} = 1 | \mathbf{x}_k)$ et $\Pr(D_{Mk} = 0 | \mathbf{x}_k)$ doivent être estimées. Pour ce type d'estimation, DiNardo et coll. (1996) ont proposé d'utiliser un modèle logit ou probit. En se servant de l'information provenant de l'échantillon,

$$\hat{\psi}(\mathbf{x}_k) = \frac{\widehat{\Pr}(D_{Mk} = 1 | \mathbf{x}_k) \big/ \widehat{\Pr}(D_{Mk} = 1)}{\widehat{\Pr}(D_{Mk} = 0 | \mathbf{x}_k) \big/ \widehat{\Pr}(D_{Mk} = 0)}. \tag{4.1}$$

En utilisant le facteur de repondération, la moyenne contrefactuelle des salaires des femmes est estimée par

$$\hat{\bar{Y}}_{F|M}^{\mathrm{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)\, y_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)}, \tag{4.2}$$

et les moyennes contrefactuelles des caractéristiques des femmes sont estimées par

$$\hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)\, \mathbf{x}_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)}. \tag{4.3}$$

Le facteur de repondération estimé défini en l'équation (4.1) sera égal à

$$\hat{\psi}(\mathbf{x}_k) = \hat{a}\ \exp\left(\mathbf{x}_k^{\top} \hat{\boldsymbol{\gamma}}\right),$$

où $\hat{\boldsymbol{\gamma}}$ est l'estimation de $\boldsymbol{\gamma}$ d'après l'échantillon en utilisant la vraisemblance empirique et $\hat{a}$ est le ratio des proportions estimées de femmes et d'hommes. Il est donné par :

$$\hat{a} = \frac{\widehat{\Pr}(D_{Mk} = 0)}{\widehat{\Pr}(D_{Mk} = 1)} = \frac{\sum_{k \in S_F} d_k}{\sum_{k \in S_M} d_k}.$$

Puisque la méthode DFL est présentée en tenant compte des poids de sondage, le facteur de repondération ψ_k sera multiplié par $d_k, k \in S_F$. Nous appellerons ce facteur résultant « facteur DFL pondéré ». La moyenne contrefactuelle des salaires estimée des femmes peut être réexprimée sous la forme

$$\hat{\bar{Y}}_{F|M}^{\mathrm{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)\, y_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)} = \frac{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^{\top} \hat{\boldsymbol{\gamma}}\right) y_k}{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^{\top} \hat{\boldsymbol{\gamma}}\right)}. \tag{4.4}$$

Les moyennes contrefactuelles des caractéristiques des femmes sont estimées par

$$\hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)\, \mathbf{x}_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)} = \frac{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^{\top} \hat{\boldsymbol{\gamma}}\right) \mathbf{x}_k}{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^{\top} \hat{\boldsymbol{\gamma}}\right)}.$$

En utilisant le facteur de repondération, les coefficients contrefactuels dans l'échantillon de femmes sont donnés par

$$\boldsymbol{\beta}_F^{\mathrm{DFL}} = \left(\sum_{k \in U_F} \psi(\mathbf{x}_k)\, \mathbf{x}_k \mathbf{x}_k^{\top} \right)^{-1} \sum_{k \in U_F} \psi(\mathbf{x}_k)\, \mathbf{x}_k y_k,$$

et estimés par

$$\hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}} = \left(\sum_{k \in S_F} d_k \hat{\psi}\left(\mathbf{x}_k\right) \mathbf{x}_k \mathbf{x}_k^{\top} \right)^{-1} \sum_{k \in S_F} d_k \hat{\psi}\left(\mathbf{x}_k\right) \mathbf{x}_k y_k . \tag{4.5}$$

Les coefficients susmentionnés doivent être calculés, parce que, sous la même condition que dans le résultat 1, la moyenne contrefactuelle des salaires des femmes définie dans (4.2) est donnée par

$$\hat{\bar{Y}}_{F|M}^{\mathrm{DFL}} = \hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}^{\top}} \hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}} .$$

La formule de la décomposition BO peut maintenant être exprimée sous la forme

$$\begin{aligned} \hat{\bar{Y}}_M - \hat{\bar{Y}}_F &= \left(\hat{\bar{Y}}_{F|M}^{\mathrm{DFL}} - \hat{\bar{Y}}_F \right) + \left(\hat{\bar{Y}}_M - \hat{\bar{Y}}_{F|M}^{\mathrm{DFL}} \right) \\ &= \left(\hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}^{\top}} \hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}} - \hat{\bar{\mathbf{X}}}_F^{\top} \hat{\boldsymbol{\beta}}_F \right) + \left(\hat{\bar{\mathbf{X}}}_M^{\top} \hat{\boldsymbol{\beta}}_M - \hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}^{\top}} \hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}} \right), \end{aligned} \tag{4.6}$$

où $\hat{\boldsymbol{\beta}}_M$ et $\hat{\boldsymbol{\beta}}_F$ sont définis dans (3.1). Le premier terme de l'équation (4.6) est l'effet de composition et le second, l'effet de structure.

4.2 Décomposition plus poussée de l'effet de structure

Comme le soulignent Fortin et coll. (2011), l'objectif du facteur de repondération DFL est de rendre la distribution des caractéristiques des femmes identique à celle des hommes. Cela implique que les moyennes des variables auxiliaires dans les deux groupes doivent être égales. Or, cela n'est pas le cas pour la méthode DFL. En effet,

$$\hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}} \neq \hat{\bar{\mathbf{X}}}_M \tag{4.7}$$

(voir, par exemple, Fortin et coll. 2011; Donzé, 2013). Le facteur de repondération n'arrive donc pas à faire concorder parfaitement les deux distributions.

L'effet de structure dans l'équation (4.6) peut être subdivisé en les éléments suivants

$$\left(\hat{\bar{\mathbf{X}}}_M^{\top} \hat{\boldsymbol{\beta}}_M - \hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}^{\top}} \hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}} \right) = \hat{\bar{\mathbf{X}}}_M^{\top} \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}} \right) + \left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}} \right) \hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}}, \tag{4.8}$$

où $\hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{DFL}}$ et $\hat{\boldsymbol{\beta}}_F^{\mathrm{DFL}}$ sont définis dans les équations (4.3) et (4.5), respectivement (Fortin et coll. 2011). Le premier terme du deuxième membre de l'équation (4.8) est l'*effet pur* et le deuxième, l'*effet résiduel* ou *erreur totale de repondération* (Fortin et coll. 2011). L'effet pur est la part inexpliquée réelle de l'écart salarial. L'effet résiduel contient l'inadéquation du modèle, autrement dit, ce que le facteur de repondération n'a pas réussi à faire concorder entre les distributions des caractéristiques des hommes et des femmes. Cette méthode permet de construire une distribution contrefactuelle des salaires. Cette nouvelle distribution peut alors être comparée aux distributions observées des salaires des femmes et des hommes. L'inconvénient de la méthode est qu'il peut arriver qu'au moins une caractéristique soit un bon prédicteur du sexe (par exemple, le secteur économique). Cette situation implique que $\Pr\left(D_{Mk} = 1 \middle| \mathbf{x}_k\right)$ peut s'approcher de 1 et que le facteur de repondération prendra une grande valeur (Fortin et coll. 2011). Cela mène de toute évidence à une grande variance de ce facteur, comme nous le montrerons à la section 8.

5 L'approche du calage

5.1 La méthode de calage

La méthode de calage a été introduite par Deville et Särndal (1992). L'idée qui sous-tend la technique est d'utiliser l'information sur certaines variables auxiliaires disponibles au niveau de la population pour estimer une fonction d'une variable d'intérêt. Habituellement, les variables auxiliaires et la variable d'intérêt sont corrélées. Les estimations résultantes sont convergentes et efficaces.

Sous l'hypothèse que l'on dispose des poids de sondage d_k et que l'on connaît les totaux des données auxiliaires au niveau de la population exprimés par

$$\mathbf{X} = \sum_{k \in U} \mathbf{x}_k ,$$

de nouveaux poids $w_k , k \in S$ doivent être construits, de manière que la contrainte (ou équation de calage) qui suit soit respectée

$$\sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k . \tag{5.1}$$

Les poids sont déterminés en résolvant en $\boldsymbol{\lambda}$ les équations de calage qui deviennent

$$\sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in S} d_k F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right) \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k ,$$

où $F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right)$ est la fonction de calage. L'estimation par calage résultante de Y est

$$\hat{Y} = \sum_{k \in S} d_k y_k F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right). \tag{5.2}$$

Dans la partie qui suit, nous utiliserons le cas linéaire, où la pseudo fonction de distance est la distance du khi carré et la fonction de calage est donnée par $F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right) = 1 + \mathbf{x}_k^\top \boldsymbol{\lambda}.$ Dans le second cas, nous appliquerons la méthode du raking ratio, qui utilise la pseudo-distance d'entropie et où la fonction de calage est donnée par $F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right) = \exp \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right)$.

5.2 Calage des caractéristiques des femmes sur les caractéristiques des hommes

Supposons que, pour toutes les unités de l'échantillon, il existe un poids de sondage donné d_k. Dans le présent contexte, les variables auxiliaires utilisées dans le processus de calage sont certaines caractéristiques mesurées pour chaque individu. L'objectif est de « détourner » la technique de calage afin de calculer un système de pondération qui ajuste les totaux des variables auxiliaires des femmes sur les totaux des hommes. La variable d'intérêt est le logarithme du salaire.

Dans l'échantillon de femmes, de nouveaux poids w_k proches de d_k sont calculés, de manière que $\sum_{k \in S_F} G \left(w_k , d_k \right)$ soit minimisée. L'équation de calage qui suit est satisfaite

$$\sum_{k \in S_F} w_k \mathbf{x}_k = \hat{\hat{\mathbf{X}}}_M , \tag{5.3}$$

où le vecteur $\hat{\hat{\mathbf{X}}}_M$ contient les totaux des caractéristiques des hommes ajustés sur le total des poids des femmes divisé par le total des poids des hommes.

$$\hat{\tilde{\mathbf{X}}}_M = \frac{\sum_{k\in S_F} d_k}{\sum_{k\in S_M} d_k} \sum_{k\in S_M} d_k \mathbf{x}_k .$$

La division de l'équation de calage (5.3) par $\sum_{k\in S_F} d_k$ donne

$$\frac{\sum_{k\in S_F} w_k \mathbf{x}_k}{\sum_{k\in S_F} d_k} = \frac{\hat{\tilde{\mathbf{X}}}_M}{\sum_{k\in S_F} d_k} = \hat{\tilde{\mathbf{X}}}_M . \tag{5.4}$$

Donc, avec les nouveaux poids w_k, les nouvelles moyennes des caractéristiques des femmes sont égales à celles des hommes. Une autre égalité intéressante est

$$\sum_{k\in S_F} w_k = \sum_{k\in S_F} d_k , \tag{5.5}$$

qui est vérifiée parce que $x_{k1} = 1, k \in S_M$ et le calage est effectué dessus. Si

$$\hat{\tilde{\tilde{\mathbf{X}}}}_M = \frac{\sum_{k\in S_F} w_k \mathbf{x}_k}{\sum_{k\in S_F} w_k},$$

en regroupant les équations (5.4) et (5.5), cela signifie que

$$\hat{\tilde{\tilde{\mathbf{X}}}}_M = \hat{\tilde{\mathbf{X}}}_M . \tag{5.6}$$

L'estimateur de la moyenne contrefactuelle des salaires des femmes est donc

$$\hat{\tilde{\tilde{Y}}}_{F|M} = \frac{\sum_{k\in S_F} w_k y_k}{\sum_{k\in S_F} d_k} = \frac{\sum_{k\in S_F} w_k y_k}{\sum_{k\in S_F} w_k} .$$

5.3 Calage linéaire

Résultat 2 *La moyenne contrefactuelle des salaires des femmes obtenue en utilisant le calage linéaire est égale à la moyenne contrefactuelle des salaires obtenue en utilisant la méthode BO pondérée, c'est-à-dire* $\hat{\tilde{\tilde{Y}}}_{F|M} = \hat{\tilde{\mathbf{X}}}^\top \hat{\boldsymbol{\beta}}_F$.

Preuve

Afin de déterminer le vecteur $\boldsymbol{\lambda}$ dans le cas où l'on utilise la pseudo-distance du khi carré, l'équation qui suit doit être résolue

$$\begin{aligned}\hat{\tilde{\tilde{\mathbf{X}}}}_M &= \sum_{k\in S_F} d_k \mathbf{x}_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) = \sum_{k\in S_F} d_k \mathbf{x}_k \left(1 + \mathbf{x}_k^\top \boldsymbol{\lambda}\right) \\ &= \sum_{k\in S_F} d_k \mathbf{x}_k + \left(\sum_{k\in S_F} d_k \mathbf{x}_k \mathbf{x}_k^\top\right) \boldsymbol{\lambda}.\end{aligned}$$

Donc,

$$\boldsymbol{\lambda} = \left(\sum_{k\in S_F} d_k \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \left(\hat{\tilde{\tilde{\mathbf{X}}}}_M - \sum_{k\in S_F} d_k \mathbf{x}_k\right) = \mathbf{T}^{-1}\left(\hat{\tilde{\tilde{\mathbf{X}}}}_M - \hat{\mathbf{X}}_F\right),$$

où

$$\mathbf{T} = \sum_{k \in S_F} d_k \mathbf{x}_k \mathbf{x}_k^\top .$$

Donc,

$$w_k = d_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) = d_k \left\{1 + \mathbf{x}_k^\top \mathbf{T}^{-1}\left(\hat{\bar{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right)\right\}.$$

En utilisant le résultat de l'équation précédente, le numérateur de l'expression (5.2) devient

$$\begin{aligned} \hat{Y}_{F|M}^{\mathrm{LC}} &= \sum_{k \in S_F} d_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) y_k \\ &= \sum_{k \in S_F} d_k y_k + \left(\hat{\bar{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right)^\top \mathbf{T}^{-1} \sum_{k \in S_F} d_k \mathbf{x}_k y_k , \end{aligned} \tag{5.7}$$

où $\hat{Y}_{F|M}^{\mathrm{LC}}$ désigne le total des logarithmes des salaires dans l'échantillon de femmes, quand le total est construit en utilisant la pseudo-distance du khi carré. Soit

$$\hat{\boldsymbol{\beta}}_F = \mathbf{T}^{-1} \sum_{k \in S_F} d_k \mathbf{x}_k y_k .$$

Le vecteur $\hat{\boldsymbol{\beta}}_F$ a déjà été défini de la même façon dans l'équation (3.1) pour la méthode BO pondérée. Nous réécrivons l'équation (5.7) sous la forme

$$\begin{aligned} \hat{Y}_{F|M}^{\mathrm{LC}} &= \sum_{k \in S_F} d_k y_k + \left(\hat{\bar{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right)^\top \hat{\boldsymbol{\beta}}_F \\ &= \hat{Y}_F + \left(\hat{\bar{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right)^\top \hat{\boldsymbol{\beta}}_F \\ &= \hat{\bar{\mathbf{X}}}_M^\top \hat{\boldsymbol{\beta}}_F , \end{aligned} \tag{5.8}$$

parce que, sous la condition du résultat 1, $\hat{\mathbf{X}}_F^\top \hat{\boldsymbol{\beta}}_F = \hat{Y}_F$. En divisant (5.8) par $\sum_{k \in S_F} w_k$, nous obtenons le résultat 2.

Si l'on utilise la pseudo-distance du khi carré, les poids résultants n'ont pas de bornes. Cela signifie que les poids de calage pourraient être négatifs. Même si cette approche de calage donne les mêmes résultats que la méthode BO pour les salaires moyens, nous recommandons d'utiliser une approche qui donne des poids non négatifs.

5.4 Calage par raking ratio

La deuxième approche de calage s'appuie sur la pseudo-distance d'entropie. Elle est également appelée calage par « raking ratio ». Si l'on utilise la pseudo-distance d'entropie, l'équation (5.3) devient

$$\hat{\bar{\mathbf{X}}}_M = \sum_{k \in S_F} d_k \mathbf{x}_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) = \sum_{k \in S_F} d_k \mathbf{x}_k \exp\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right). \tag{5.9}$$

Ce système d'équations résultant ne peut pas être résolu analytiquement. Cependant, on peut trouver la valeur de $\boldsymbol{\lambda}$ au moyen de l'algorithme de Newton-Raphson.

L'équation (5.2) peut maintenant s'écrire

$$\hat{Y}_{F|M}^{\mathrm{RRC}} = \sum_{k \in S_F} d_k \exp\left(\mathbf{x}_k^{\top}\boldsymbol{\lambda}\right) y_k,$$

où $\hat{Y}_{F|M}^{\mathrm{RRC}}$ désigne le total du logarithme du salaire dans l'échantillon de femmes, quand le total est construit en utilisant le calage par raking ratio. La moyenne contrefactuelle des salaires des femmes s'écrit

$$\hat{\bar{Y}}_{F|M}^{\mathrm{RRC}} = \frac{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}_k^{\top}\boldsymbol{\lambda}\right) y_k}{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}_k^{\top}\boldsymbol{\lambda}\right)}.$$

L'équation qui précède est très similaire à l'équation (4.4). La seule différence tient à l'estimation des paramètres $\boldsymbol{\lambda}$ et $\boldsymbol{\gamma}$. Le vecteur $\boldsymbol{\lambda}$ contient les multiplicateurs de Lagrange résolvant l'équation 5.9 sous la contrainte (5.1), tandis que l'on trouve le vecteur $\boldsymbol{\gamma}$ par la méthode du maximum de vraisemblance.

Après le calcul des poids de calage w_k définis en (5.3) et en utilisant l'information qui figure dans l'équation (5.6), il résulte que

$$\hat{\bar{\mathbf{X}}}_M = \hat{\bar{\mathbf{X}}}_{F|M}^{\mathrm{RRC}} = \frac{\sum_{k \in S_F} \mathbf{x}_k w_k}{\sum_{k \in S_F} w_k},$$

qui assure que la part résiduelle de l'effet de structure définie dans l'équation (4.8) sera nulle. Il s'agit d'une solution au problème présenté à la section 4.3. Cette approche de calage remédie aussi au problème des poids négatifs que l'on peut obtenir en utilisant la pseudo-distance du khi carré.

6 Application à l'Enquête suisse sur la structure des salaires

6.1 Description des données

L'ensemble de données utilisé contient des renseignements recueillis en 2008 par l'Office fédéral de la statistique suisse au moyen d'une enquête appelée Enquête suisse sur la structure des salaires. Un questionnaire a été envoyé à des organisations publiques et privées des secteurs secondaire et tertiaire afin de recueillir des renseignements sur des aspects particuliers. Ces aspects comprennent la taille de l'organisation, les types de contrats d'emploi et la rémunération des employés au sein de l'organisation. Le questionnaire a été rempli par un membre autorisé de l'organisation et non par des employés. Cela améliore la fiabilité des données et les rend moins susceptibles aux approximations. Les analyses qui suivent ont été limitées au secteur privé. Les observations valides incluses dans les analyses correspondaient aux individus sans valeurs manquantes, qui avaient travaillé plus d'une heure par semaine et pour lesquels la différence entre l'âge et les années d'expérience de travail était égale ou supérieure à 15 (d'après les lois suisses sur l'emploi, cela représente l'âge minimum légal pour pouvoir travailler). Donc, 29 048 cas ont été exclus de l'ensemble de données original. L'ensemble de données final comptait 647 139 hommes et 435 507 femmes. L'ensemble de données fourni par l'Office fédéral de la statistique contenait aussi les poids de sondage, de sorte qu'aucun traitement ni calcul de ces poids n'a été effectué dans la présente application.

Dans les tableaux qui suivent, les valeurs exprimées en francs suisses sont données entre parenthèses. Cependant, les chiffres sont représentés graphiquement en utilisant les logarithmes des salaires. Les valeurs sont obtenues en tenant compte des poids de sondage.

Le tableau 6.1 contient la médiane et la moyenne salariale pour l'échantillon complet, ainsi que pour les femmes et pour les hommes.

Tableau 6.1
Moyenne et médiane des salaires calculées pour l'ensemble de données complet, les femmes et les hommes, en francs suisses

	Moyenne	Médiane
Ensemble de données complet	6 977	5 905
Femmes	5 843	5 220
Hommes	7 725	6 346

Les valeurs de la moyenne ainsi que de la médiane des salaires des hommes sont supérieures aux valeurs pour l'ensemble de données complet, tandis que les valeurs pour les femmes sont inférieures. Le tableau 6.2 montre la répartition des hommes et des femmes dans les emplois faiblement et fortement rémunérés. Les quantiles pondérés des salaires pour l'ensemble de données complet sont calculés à la première ligne. Les deux lignes suivantes montrent les proportions cumulées de femmes et d'hommes qui gagnent moins que la valeur du quantile.

Tableau 6.2
Quantiles pondérés du logarithme du salaire et proportions de femmes et d'hommes qui gagnent moins que la valeur qui représente un quantile particulier du salaire calculé pour l'ensemble de données complet (les valeurs en francs suisses sont données entre parenthèses)

	Quantile										
	1 %	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	99 %
Logarithme du salaire	7,89 (2 683)	8,27 (3 897)	8,39 (4 412)	8,50 (4 905)	8,59 (5 400)	8,68 (5 905)	8,78 (6 488)	8,89 (7 233)	9,03 (8 380)	9,27 (10 667)	10,09 (24 202)
Proportion cumulée de femmes	0,02	0,17	0,32	0,43	0,53	0,63	0,72	0,81	0,89	0,96	1
Proportion cumulée d'hommes	0,006	0,06	0,12	0,21	0,31	0,42	0,52	0,63	0,74	0,86	0,99

Alors que 43 % de femmes ont un salaire inférieur à 4 905 CHF (comparativement à seulement 21 % d'hommes), seulement 11 % de femmes ont un salaire compris entre 8 380 CHF et 24 202 CHF (comparativement à 25 % d'hommes). En outre, 63 % de femmes gagnent moins que la valeur médiane des salaires pour l'ensemble complet de données, comparativement à seulement 42 % d'hommes. Les mécanismes possibles à l'origine de cette répartition doivent être étudiés. Néanmoins, ce n'est pas l'objet du présent article. Afin de mieux comprendre la distribution des salaires dans chaque échantillon, le tableau 6.3 montre les quantiles pondérés des logarithmes des salaires des femmes et des hommes, ainsi que l'écart entre eux. Une valeur surprenante de l'écart entre les salaires s'observe au quantile d'ordre 1 %. En principe, ces emplois sont d'un type qui ne requiert ni de grandes qualifications ni un niveau d'études élevé. Alors que seulement 0,6 % d'hommes occupent ce genre d'emplois (voir le tableau 6.3), ils gagnent plus

que les 2 % de femmes ayant des emplois comparables. La figure 6.1 montre les données du tableau 6.3 qui suit sous forme graphique.

Tableau 6.3
Salaires des femmes et des hommes et écart entre les salaires des hommes et des femmes, exprimés en logarithmes (les valeurs en francs suisses sont données entre parenthèses)

	Quantile										
	1 %	**10 %**	**20 %**	**30 %**	**40 %**	**50 %**	**60 %**	**70 %**	**80 %**	**90 %**	**99 %**
Femmes	7,80	8,19	8,30	8,38	8,47	8,56	8,66	8,76	8,88	9,06	9,67
	(2 432)	(3 602)	(4 005)	(4 344)	(4 756)	(5 220)	(5 743)	(6 353)	(7 154)	(8 577)	(15 761)
Hommes	8,01	8,36	8,49	8,58	8,67	8,76	8,86	8,98	9,14	9,38	10,26
	(3 000)	(4 259)	(4 850)	(5 344)	(5 820)	(6 346)	(7 012)	(7 908)	(9 291)	(11 905)	(28 571)
Écart	0,21	0,17	0,19	0,21	0,20	0,20	0,20	0,22	0,26	0,33	0,59
	(568)	(657)	(845)	(1 000)	(1 064)	(1 126)	(1 269)	(1 555)	(2 137)	(3 328)	(12 810)

La distance entre les deux ensembles de points augmente vers les quantiles de niveau plus élevé, ce qui signifie que l'écart entre les salaires devient plus important. Il reste à établir quelle part de ces écarts n'est pas attribuable à des différences de caractéristiques entre les hommes et les femmes. En guise de preuve graphique finale des inégalités salariales, la figure 6.2 montre les distributions des logarithmes des salaires des femmes et des hommes.

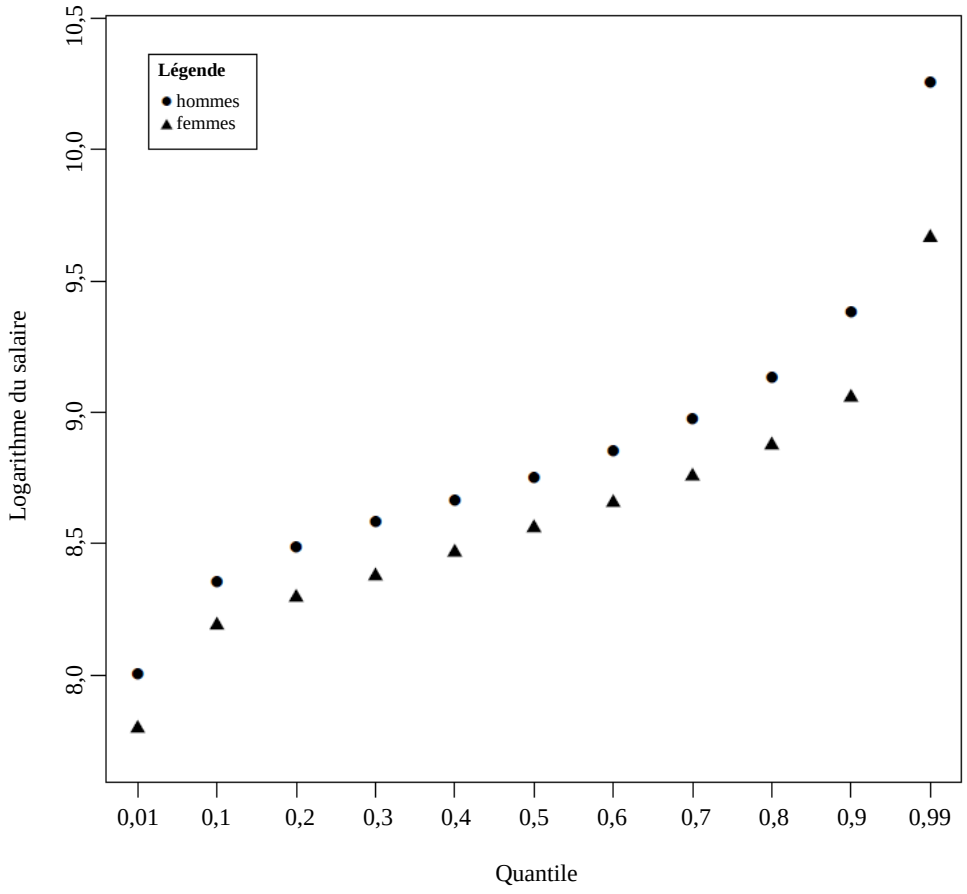

Figure 6.1 Quantiles pondérés du logarithme des salaires des femmes et des hommes.

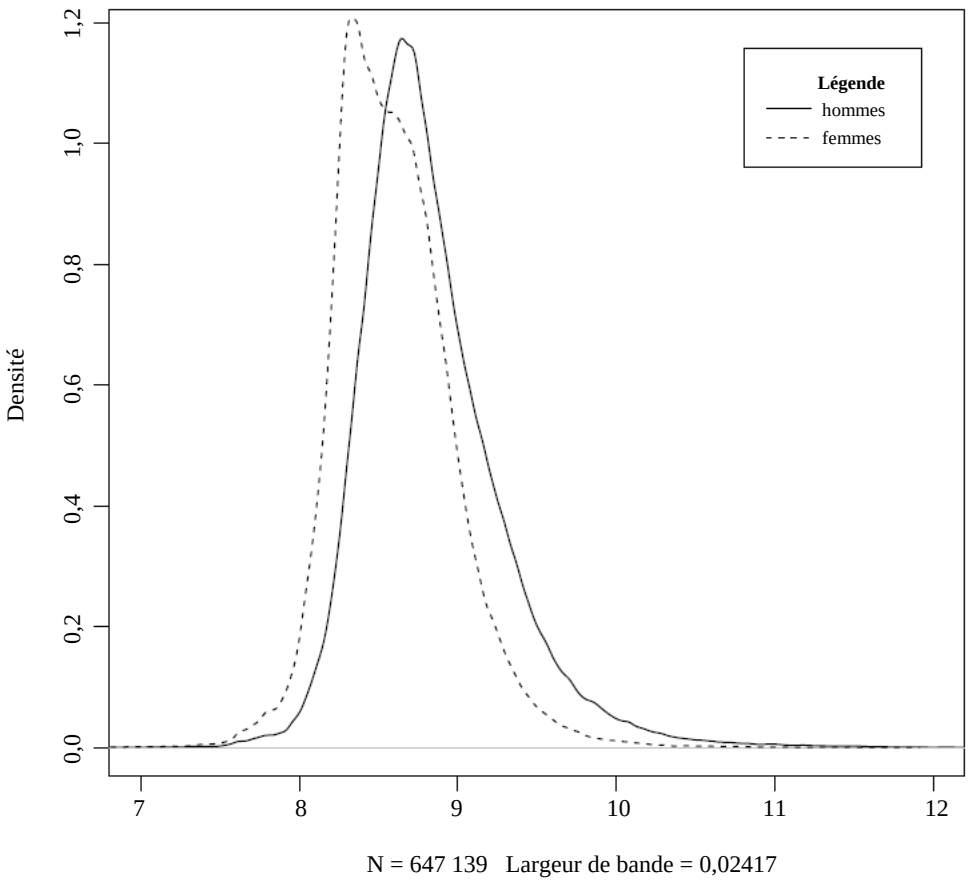

Figure 6.2 Densités estimées des logarithmes des salaires des hommes et des femmes.

6.2 Le modèle

Le modèle de régression contient huit variables explicatives :

- niveau d'études : variable nominale comprenant 9 catégories indiquant le niveau d'études le plus élevé atteint;
- nombre d'années de service dans l'emploi courant (variable de substitution pour l'expérience de travail);
- qualifications requises : variable ordinale comprenant 4 niveaux indiquant le degré de qualification requis pour le poste;
- région de l'institution : variable nominale comptant 7 catégories;
- secteur économique : variable nominale comptant 10 catégories;
- degré d'activité : le taux d'activité de l'employé (si la valeur est 1, l'employé travaille à temps plein);
- âge : l'âge réel;

- carré de l'âge : le carré de l'âge est également inclus, parce qu'il a été constaté que le salaire augmente jusqu'à un certain âge, puis diminue par après (voir, par exemple, Williams, 2010).

Le modèle a été choisi parmi un certain nombre de modèles contenant plusieurs variables en utilisant le critère AIC minimal. La variable dépendante est le logarithme du salaire standardisé. Par salaire standardisé d'une personne, nous entendons le salaire calculé pour cette personne si elle travaillait à temps plein. Cette variable est fournie par l'Office fédéral de la statistique suisse dans l'ensemble de données, de sorte qu'aucun calcul n'a été effectué par les auteurs.

6.3 Poids et distributions contrefactuelles

La présente section comprend uniquement les résultats exprimés en logarithmes. Lorsqu'on utilise la méthode BO, l'écart entre les salaires moyens des hommes et des femmes est de 0,23, dont 0,09 seulement représentent la part expliquée et 0,14 représentent la part inexpliquée. Nous comparons les résultats obtenus par les méthodes présentées plus haut. La méthode de calage au moyen de la pseudo-distance du khi carré est appelée « linéaire », le calage au moyen de la divergence de Kullback-Leibler est appelé « raking ratio » et la méthode proposée par DiNardo et coll. (1996), ajustée pour tenir compte des poids de sondage est appelée « DFL ». Premièrement, le tableau 6.4 montre les valeurs minimale et maximale des poids, ainsi que les écarts-types, obtenus en utilisant le calage linéaire, le calage par raking ratio et la méthode DFL pondérée.

Tableau 6.4
Poids minimal et maximal et écart-type

Méthode	Minimal	Maximal	Écart-type
Linéaire	-39,06	319,8	4,97
Raking ratio	0,0011	904,7	6,79
DFL pondérée	0,0022	804,4	6,16

Le cas linéaire donne les mêmes résultats que la méthode BO pondérée. Cependant, comme le montre le tableau 6.4, ce cas particulier produit des poids négatifs. Le nombre de ces poids négatifs était de 69 553 (14,59 %). L'option du raking ratio donne toujours des poids positifs, mais l'écart-type des poids est plus grand. Le facteur DFL pondéré possède un plus petit écart-type que les poids obtenus par la méthode de calage par raking ratio. Il existe 1 319 cas où la probabilité conditionnelle d'être un homme est plus grande que 0,98. Initialement, le facteur DFL est multiplié par le ratio entre la somme des poids de sondage des femmes et la somme des poids de sondage des hommes. Puisque $\hat{a}$ est plus petit que 1, le facteur de repondération diminuera. Si, d'autre part, $\hat{a}$ est plus grand que 1 (par exemple, pour les secteurs tels que le secteur public), le facteur de repondération pourrait être plus grand. Le tableau 6.5 montre l'effet de structure estimé aux niveaux moyens des salaires. Les deux approches de calage donnent des effets de structure et de composition égaux. L'utilisation du facteur de repondération DFL résulte en un effet de structure légèrement plus faible et un effet de composition plus grand que les deux autres méthodes.

Tableau 6.5
Effets de composition et de structure estimés dans la différence entre les moyennes

Méthode	Effet de composition	Effet de structure	Total
Linéaire	0,09	0,14	0,23
Raking ratio	0,09	0,14	0,23
DFL pondérée	0,10	0,13	0,23

Étant donné que des poids négatifs sont obtenus dans le premier cas de calage, la densité estimée correspondante ne peut pas être représentée graphiquement. Seules les distributions contrefactuelles des salaires des femmes obtenues en utilisant le raking ratio et le facteur de repondération DFL sont construites. Elles sont présentées à la figure 6.3.

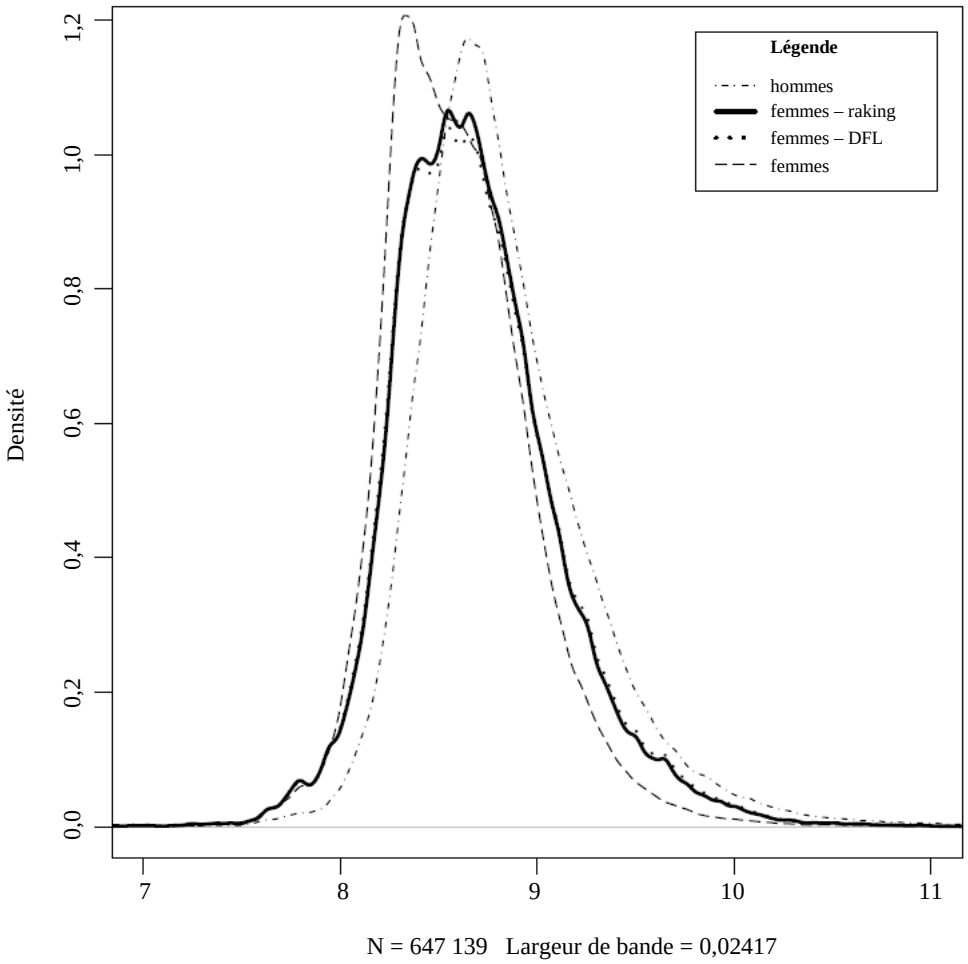

Figure 6.3 Densités estimées du logarithme du salaire des femmes et des hommes et distributions contrefactuelles du logarithme du salaire des femmes construites en utilisant le raking ratio et le facteur DFL repondéré, respectivement.

La figure 6.3 montre que les deux distributions contrefactuelles des salaires sont très proches l'une de l'autre autour des queues. Cependant, vers le milieu, les deux méthodes ne donnent pas les mêmes résultats. Comme nous l'avons mentionné plus haut, l'utilisation des méthodes de repondération DFL et de calage permet d'estimer les effets de composition et de structure non seulement aux niveaux moyens, mais aussi tout le long de la distribution. Le tableau 6.6 donne les effets de structure et de composition estimés des écarts salariaux entre les hommes et les femmes calculés en utilisant les trois méthodes pour certains quantiles.

Tableau 6.6
Effets de composition et de structure estimés des écarts salariaux pour certains quantiles

Quantile	Méthode	Effet de composition (%)	Effet de structure (%)	Total
1 %	Linéaire	0,01 (3 %)	0,20 (97 %)	0,21
	Raking	-0,01 (-3,5 %)	0,22 (103,5 %)	0,21
	DFL pondérée	-0,01 (-3,4 %)	0,22 (103,4 %)	0,21
10 %	Linéaire	0,05 (28,8 %)	0,12 (71,2 %)	0,17
	Raking	0,04 (22,4 %)	0,13 (77,6 %)	0,17
	DFL pondérée	0,03 (19,4 %)	0,14 (80,6 %)	0,17
20 %	Linéaire	0,07 (34,2 %)	0,13 (65,8 %)	0,20
	Raking	0,06 (29,7 %)	0,13 (70,3 %)	0,19
	DFL pondérée	0,05 (28,2 %)	0,14 (71,8 %)	0,19
50 %	Linéaire	0,09 (46,3 %)	0,10 (53,7 %)	0,19
	Raking	0,09 (44,7 %)	0,11 (55,3 %)	0,20
	DFL pondérée	0,09 (45,7 %)	0,11 (54,3 %)	0,20
80 %	Linéaire	0,11 (43,9 %)	0,15 (56,1 %)	0,26
	Raking	0,12 (46,5 %)	0,14 (53,5 %)	0,26
	DFL pondérée	0,13 (50,8 %)	0,13 (49,2 %)	0,26
90 %	Linéaire	0,15 (46,0 %)	0,18 (54,0 %)	0,33
	Raking	0,17 (51,6 %)	0,16 (48,4 %)	0,33
	DFL pondérée	0,19 (58,0 %)	0,14 (42,0 %)	0,33
99 %	Linéaire	0,24 (40,0 %)	0,36 (60,0 %)	0,60
	Raking	0,27 (45,3 %)	0,33 (54,7 %)	0,60
	DFL pondérée	0,29 (49,4 %)	0,30 (50,6 %)	0,59

La proportion de l'effet de structure dans l'écart salarial entier entre les hommes et les femmes diminue à mesure qu'augmente l'ordre du quantile. Cela signifie qu'une plus grande part de l'écart salarial peut être expliquée par des différences de caractéristiques des groupes pour les emplois à salaires élevés que pour les emplois à faibles salaires. Le raking ratio et le facteur de repondération DFL donnent des résultats similaires jusqu'au quantile d'ordre 90 %. Au premier percentile, l'effet de composition estimé est négatif, ce qui signifie qu'à ce point, les différences de salaires sont dues uniquement à la discrimination.

La figure 6.4 montre les quantiles pondérés des logarithmes du salaire des hommes et du salaire des femmes, et contraste les distributions contrefactuelles obtenues au moyen du calage par raking ratio et du facteur de repondération DFL. Comme le calage linéaire donnait des poids négatifs, le graphique correspondant n'est pas présenté.

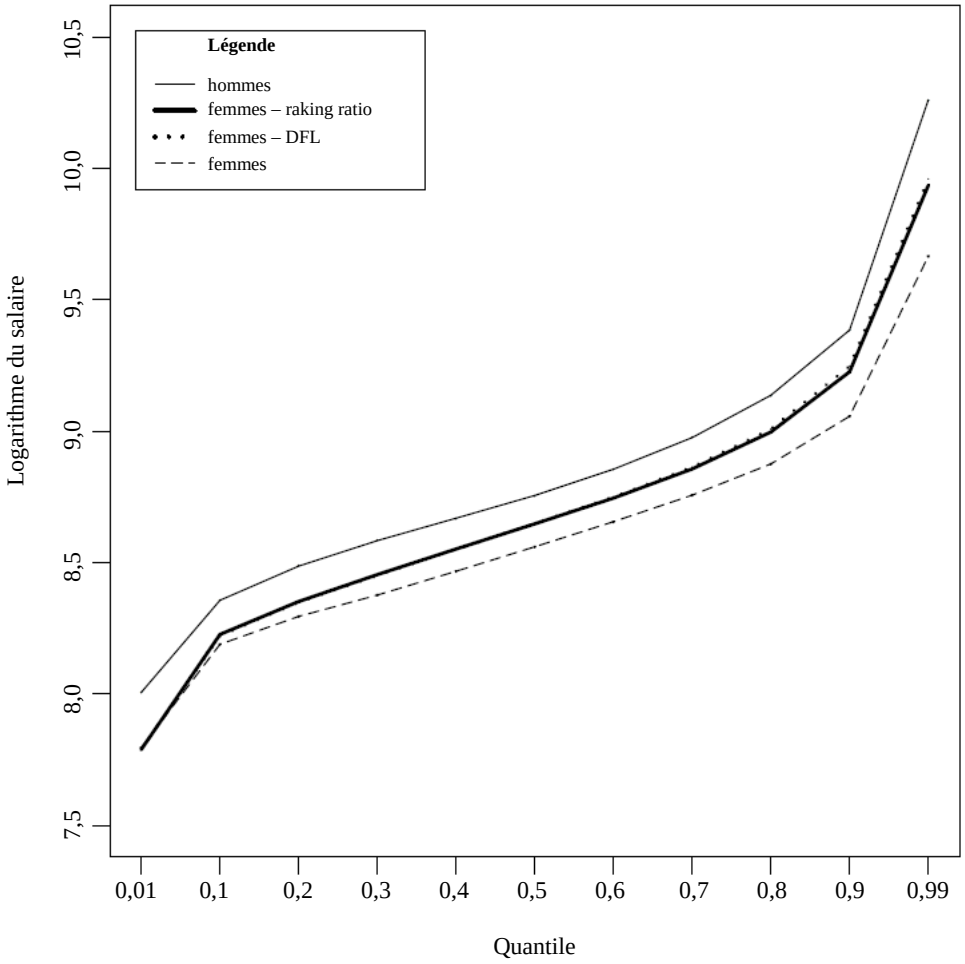

Figure 6.4 Quantiles pondérés des logarithmes des salaires des femmes et des hommes, et quantiles pondérés des distributions contrefactuelles du logarithme du salaire des femmes construites en utilisant le calage par raking ratio et le facteur DFL pondéré.

6.4 Décomposition plus poussée de l'effet de structure

Un modèle logistique de la probabilité d'être un homme donne des valeurs estimées comprises entre 0,002 et 0,99. C'est pour les variables « années d'occupation du poste courant », « âge » et « carré de l'âge » que les écarts entre les valeurs moyennes des hommes et les valeurs repondérées des femmes, calculées en utilisant le facteur de repondération, sont les plus grands. Dans l'équation (4.8), l'effet de structure est composé de l'effet pur et l'effet résiduel. En utilisant le facteur de repondération DFL, l'effet résiduel vaut -0,00474. Par contre, en utilisant l'une ou l'autre des techniques de calage, cet effet est nul dans les deux cas. En outre, l'approche de calage permet de contourner le calcul des coefficients de régression contrefactuels, parce que la technique assure l'égalité entre les moyennes $\hat{\bar{\mathbf{X}}}_M$ et $\hat{\bar{\mathbf{X}}}_{F|M}$. Donc, le calage représente une généralisation de la méthode du facteur de repondération DFL, car il permet une estimation plus précise de l'effet de structure, vu que la valeur résultante n'inclut que l'effet pur.

7 Conclusion

Le phénomène de discrimination présente de multiples facettes et peut être produit par de nombreux mécanismes. Néanmoins, nous n'examinons ici son estimation que d'un point de vue méthodologique. Les deux approches de calage prises en considération représentent une généralisation de deux méthodes de décomposition existantes, à savoir la méthode de Blinder (1973) et Oaxaca (1973) et la méthode semi-paramétrique de DiNardo et coll. (1996), toutes deux exprimées en utilisant les poids de sondage. Les méthodes originales peuvent aussi être obtenues si tous les poids de sondage sont considérés comme égaux à 1. Le cas linéaire donne le même résultat que la méthode BO. Cependant, puisque les poids résultants ne sont pas bornés, des valeurs négatives sont parfois observées. Tout comme la méthode DFL, l'approche de calage permet de décomposer les écarts salariaux à d'autres points que la moyenne, tels que les quantiles. Cependant, le calage par raking ratio représente une amélioration de la méthode DFL, en ce sens que l'estimation de l'effet de structure comprendra toujours un effet résiduel nul. Donc, l'effet de structure sera composé uniquement de l'effet pur. La décomposition des écarts salariaux le long des quantiles permet de conclure que, dans les emplois faiblement rémunérés, les inégalités sont dues uniquement à la discrimination. Dans le présent article, l'accent est mis sur la généralisation de deux méthodes de décomposition bien établies au moyen de l'approche de calage.

Remerciements

Les auteurs remercient l'Office fédéral de la statistique suisse de son soutien financier, et son département des salaires (LOHN) d'avoir fourni les données. Cependant, les opinions exprimées dans le présent article ne reflètent pas nécessairement celles de l'Office fédéral de la statistique suisse.

Annexe A

Preuve du résultat 1

$$\begin{aligned}\hat{\mathbf{X}}_h\hat{\boldsymbol{\beta}}_h &= \hat{\mathbf{X}}_h\left(\sum_{k\in S_h} d_k\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\sum_{l\in S_h} d_l\mathbf{x}_l y_l \\ &= \sum_{j\in S_h} d_j\mathbf{x}_j^\top\left(\sum_{k\in S_h} d_k\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\sum_{l\in S_h} d_l\mathbf{x}_l y_l \\ &= \left(\sum_{j\in S_h}\boldsymbol{\varsigma}^\top d_j\mathbf{x}_j\mathbf{x}_j^\top\right)\left(\sum_{k\in S_h} d_k\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\sum_{l\in S_h} d_l\mathbf{x}_l y_l \\ &= \sum_{l\in S_h}\boldsymbol{\varsigma}^\top d_l\mathbf{x}_l y_l = \sum_{l\in S_h} d_l y_l = \hat{Y}_h.\end{aligned}$$

En divisant cette équation par $\sum_{k\in S_h} d_k$, on obtient le résultat 1.

Annexe B

B.1 Linéarisation des moyennes

Afin de calculer la variance des moyennes et des moyennes contrefactuelles, nous avons utilisé la méthode de linéarisation proposée par Graf (2011). L'auteur propose de calculer la dérivée partielle de l'estimateur par rapport à l'indicateur d'échantillon. Cette dérivée fournit la variable linéarisée qui peut être insérée dans l'estimateur de variance. Les moyennes sont définies par :

$$\hat{\bar{Y}}_F = \frac{\sum_{k\in S_F} d_k y_k}{\sum_{k\in S_F} d_k},$$

et

$$\hat{\bar{Y}}_M = \frac{\sum_{l\in S_M} d_l y_l}{\sum_{l\in S_M} d_l}.$$

Pour les deux salaires moyens, nous obtenons les variables linéarisées :

$$\frac{\partial\hat{\bar{Y}}_F}{\partial I_j} = \begin{cases}\dfrac{d_j\left(y_j - \hat{\bar{Y}}_F\right)}{\sum_{k\in S_F} d_k} & j\in S_F \\ 0 & j\in S_M\end{cases},$$

et

$$\frac{\partial\hat{\bar{Y}}_M}{\partial I_j} = \begin{cases}\dfrac{d_j\left(y_j - \hat{\bar{Y}}_M\right)}{\sum_{l\in S_M} d_l} & j\in S_M \\ 0 & j\in S_F\end{cases}.$$

B.2 Linéarisation de la moyenne contrefactuelle

Afin de calculer la moyenne contrefactuelle, nous calculons les poids v_k définis par le système

$$\mathbf{A} = \sum_{k \in S_F} v_k d_k \mathbf{x}_k = \frac{\sum_{k \in S_F} d_k}{\sum_{l \in S_M} d_l} \sum_{l \in S_M} d_l \mathbf{x}_l = \hat{\bar{\mathbf{X}}}_M \sum_{k \in S_F} d_k,$$

avec

$$v_k = F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right).$$

Pour les variables linéarisées, nous avons considéré deux cas :

- Si $j \in S_F$

$$\frac{\partial \mathbf{A}}{\partial I_j} = v_j d_j \mathbf{x}_j + \left[\sum_{k \in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k \mathbf{x}_k^\top\right] \frac{\partial \boldsymbol{\lambda}}{\partial I_j} = d_j \hat{\bar{\mathbf{X}}}_M.$$

  Donc,

$$\frac{\partial \boldsymbol{\lambda}}{\partial I_j} = -\mathbf{T}^{-1} d_j \left(v_j \mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right),$$

  où

$$\mathbf{T} = \sum_{k \in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k \mathbf{x}_k^\top.$$

- Si $j \in S_M$

$$\frac{\partial \mathbf{A}}{\partial I_j} = \left[\sum_{k \in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k \mathbf{x}_k^\top\right] \frac{\partial \boldsymbol{\lambda}}{\partial I_j} = d_j \left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right) \frac{\sum_{k \in S_F} d_k}{\sum_{l \in S_M} d_l}.$$

  Donc,

$$\frac{\partial \boldsymbol{\lambda}}{\partial I_j} = \mathbf{T}^{-1} d_j \left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right) \frac{\sum_{k \in S_F} d_k}{\sum_{l \in S_M} d_l}.$$

Puisque nous avons supposé qu'il existe un vecteur $\boldsymbol{\gamma}$ tel que $\boldsymbol{\gamma}^\top \mathbf{x}_k = 1$ pour tout $k \in U$, alors nous avons

$$\boldsymbol{\gamma}^\top \mathbf{A} = \sum_{k \in S_F} v_k d_k = \sum_{k \in S_F} d_k.$$

Considérons maintenant

$$\hat{\bar{Y}}_{F|M} = \frac{\sum_{k \in S_F} v_k d_k y_k}{\sum_{k \in S_F} v_k d_k} = \frac{\sum_{k \in S_F} v_k d_k y_k}{\sum_{k \in S_F} d_k}.$$

De nouveau, deux cas doivent être considérés :

- Si $j \in S_F$

$$\frac{\partial \hat{\bar{Y}}_{F|M}}{\partial I_j} = \frac{d_j\left(v_j y_j - \hat{\bar{Y}}_{F|M}\right) + \frac{\partial \boldsymbol{\lambda}^\top}{\partial I_j}\left[\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k\right]}{\sum_{k\in S_F} d_k}$$

$$= \frac{d_j\left(v_j y_j - \hat{\bar{Y}}_{F|M}\right) - d_j\left(v_j \mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \mathbf{T}^{-1}\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k}{\sum_{k\in S_F} d_k}$$

$$= \frac{d_j\left[v_j y_j - \hat{\bar{Y}}_{F|M} - \left(v_j \mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \mathbf{B}_F\right]}{\sum_{k\in S_F} d_k},$$

où

$$\mathbf{B}_F = \mathbf{T}^{-1}\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k.$$

- Si $j \in S_M$

$$\frac{\partial \hat{\bar{\mathbf{Y}}}_{F|M}}{\partial I_j} = \frac{\partial \boldsymbol{\lambda}^\top}{\partial I_j}\frac{\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k}{\sum_{k\in S_F} d_k}$$

$$= d_j\left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \frac{\sum_{k\in S_F} d_k}{\sum_{l\in S_M} d_l}\mathbf{T}^{-1}\frac{\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k}{\sum_{k\in S_F} d_k}$$

$$= d_j\left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \frac{1}{\sum_{l\in S_M} d_l}\mathbf{B}_F.$$

Donc, la variable linéarisée est

$$z_k = \begin{cases} \dfrac{d_j\left[v_j y_j - \hat{\bar{Y}}_{F|M} - \left(v_j \mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \mathbf{B}_F\right]}{\sum_{k\in S_F} d_k} & \text{si} \quad j \in S_F \\ \dfrac{d_j\left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \mathbf{B}_F}{\sum_{l\in S_M} d_l} & \text{si} \quad j \in S_M. \end{cases}$$

La variable linéarisée doit uniquement être insérée dans l'estimateur de variance correspondant au plan de sondage. Notons que la variance de la moyenne contrefactuelle dépend de la variance calculée pour l'échantillon d'hommes en ce qui concerne la part qui est expliquée par la régression, et de la variance calculée pour l'échantillon de femmes en ce qui concerne la part qui demeure inexpliquée.

Bibliographie

Bielby, W.T., et Baron, J.N. (1986). Men and women at work: Sex segregation and statistical discrimination. *American Journal of Sociology*, 759-799.

Blinder, A.S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 8(4), 436-455.

Bourguignon, F., Ferreira, F.H. et Leite, P.G. (2002). Beyond Oaxaca-Blinder: Accounting for differences in household income distributions across countries. *Inequality and Economic Development in Brazil*, 105.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382.

DiNardo, J. (2002). Propensity score reweighting and changes in wage distributions. Document de discussion, University of Michigan.

DiNardo, J., Fortin, N.M. et Lemieux, T. (1996). Labor market Institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, 64(5), 1001-44.

Donzé, L. (2013). Erreurs de spécification dans la décomposition de l'inégalité salariale. Dans l'*International Conference Ars Conjectandi*, 1713-2013.

Fortin, N., Lemieux, T. et Firpo, S. (2011). Decomposition methods in economics. Dans *Handbook of Labor Economics*, (Éds., O. Ashenfelter et D. Card), 4, 1-102. Elsevier.

Gardeazabal, J., et Ugidos, A. (2005). Gender wage discrimination at quantiles. *Journal of Population Economics*, 18(1), 165-179.

Graf, M. (2011). Use of survey weights for the analysis of compositional data. Dans *Compositional Data Analysis: Theory and Applications*, (Éds., V. Pawlowsky-Glahn et A. Buccianti), 114-127. Wiley, Chichester.

Neumark, D. (1988). Employers' discriminatory behavior and the estimation of wage Discrimination. *Journal of Human Resources*, 23(3), 279-295.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3), 693-709.

Weichselbaumer, D., et Winter-Ebmer, R. (2005). A meta-analysis of the international gender wage gap. *Journal of Economic Surveys*, 19(3), 479-511.

Weichselbaumer, D., et Winter-Ebmer, R. (2006). Rhetoric in economic research: The case of gender wage differentials. *Industrial Relations: A Journal of Economy and Society*, 45(3), 416-436.

Williams, C. (2010). Bien-être économique. Femmes au Canada : rapport statistique fondé sur le sexe, Statistique Canada.

Techniques d'enquête, décembre 2017
Vol. 43, nº 2, pp. 255-260
Statistique Canada, nº 12-001-X au catalogue

Une note sur les intervalles de couverture de Wilson pour les proportions estimées sur des échantillons complexes

Phillip S. Kott[1]

Résumé

La présente note expose les fondements théoriques de l'extension de l'intervalle de couverture bilatéral de Wilson à une proportion estimée à partir de données d'enquêtes complexes. Il est démontré que l'intervalle est asymptotiquement équivalent à un intervalle calculé en partant d'une transformation logistique. Une légèrement meilleure version est examinée, mais les utilisateurs pourraient préférer construire un intervalle unilatéral déjà décrit dans la littérature.

Mots-clés : Taille effective de l'échantillon; intervalle de confiance; transformation logistique.

1 Introduction

Brown, Cai et Dasgupta (2001) montrent qu'une méthode proposée par Wilson (1927) peut produire des intervalles de couverture bilatéraux se comportant raisonnablement bien pour une proportion sous échantillonnage aléatoire simple *avec* remise. La section 2 de la présente note expose les fondements théoriques de l'extension de cette méthode de construction de l'intervalle à des proportions estimées à partir d'une enquête complexe. La section 3 montre que cet intervalle de type Wilson peut être asymptotiquement équivalent à un intervalle calculé en partant d'une transformation logistique. La section 4 offre certaines conclusions.

Le terme « intervalle de couverture » est préféré ici au terme « intervalle de confiance » plus fréquent, parce qu'un intervalle de couverture de Wilson à 95 % ne vise pas à couvrir la proportion vraie dans au moins 95 % des cas quelle que soit cette proportion. Il tente plutôt simplement de couvrir la proportion vraie dans 95 % des cas pour les valeurs raisonnables de cette proportion. Pour certaines valeurs, il donne une sur-couverture, et pour d'autres, une sous-couverture, comme le montrent Brown et coll. (2001). Si on limite son applicabilité à des intervalles de couverture bilatéraux, la méthodologie de Wilson permet (en majeure partie) de ne pas tenir compte de l'asymétrie de la distribution d'une proportion estimée.

2 L'extension

Il n'est pas difficile de généraliser les intervalles de couverture de Wilson (également appelés « intervalles de score ») à des données d'enquêtes complexes. Consulter, par exemple, Kott et Carr (1997). Comme pour l'intervalle de Wilson proprement dit, on résout simplement l'équation qui suit pour la proportion vraie P :

$$\frac{(p-P)^2}{\left[\dfrac{P(1-P)}{n^*}\right]} \le z^2_{1-\alpha/2}, \tag{2.1}$$

1. Phillip S. Kott, RTI International, 6110 Executive Blvd., Rockville, MD 20852, É.-U. Courriel : pkott@rti.org.

où p est un estimateur convergent de P sous la théorie de l'échantillonnage probabiliste, et $z_{1-\alpha/2}$ est le score z normal pour $(1-\alpha/2)$, sachant que l'objectif est de produire un intervalle de couverture à $(1-\alpha)\,\%$ (α est souvent fixé à 0,05). L'élément manquant dans l'équation (2.1) est n^*, communément appelé « taille effective d'échantillon », qui, dans la formulation classique de Wilson, est la taille d'échantillon n. Dans notre contexte plus général, $n^* = p(1-p)/\mathrm{var}(p)$, où $\mathrm{var}(p)$ est un estimateur convergent de la variance de p, $\mathrm{Var}(p)$.

Afin de calculer n^*, il faut que les termes $p(1-p)$ et $\mathrm{var}(p)$ soient tous deux positifs. En outre, supposons que $1/n^* = O_P\left(1/n^a\right)$ pour une certaine valeur positive $a \le 1$, $p - P = O_P\left(1/\sqrt{n^*}\right)$, $0 < \mathrm{Var}(p) = O(1/n^*)$, et $\mathrm{var}(p)/\mathrm{Var}(p)$ est d'ordre $1 + O_P\left(1/\sqrt{n^*}\right)$. Notons que les trois dernières contraintes sont toujours vérifiées sous échantillonnage aléatoire simple avec remise à condition que $P(1-P) \ge B > 0$.

En laissant tomber les termes $O_P\left(1/[n^*]^{3/2}\right)$, mais en permettant que $p(1-p)$ soit petit (effectivement $o_p(1)$), on peut calculer cet intervalle de type Wilson pour P en partant de l'équation (2.1) :

$$p + \frac{1-2p}{n^*}\frac{z_{1-\alpha/2}^2}{2} - z_{1-\alpha/2}\left(\frac{p(1-p)}{n^*} + \frac{z_{1-\alpha/2}^2}{4(n^*)^2}\right)^{1/2} \le P$$

$$\le p + \frac{1-2p}{n^*}\frac{z_{1-\alpha/2}^2}{2} + z_{1-\alpha/2}\left(\frac{p(1-p)}{n^*} + \frac{z_{1-\alpha/2}^2}{4(n^*)^2}\right)^{1/2}. \qquad (2.2)$$

Nous pouvons l'appeler « intervalle de couverture de Wilson sous échantillonnage complexe ». WesVar (2007) calcule une variante de cet intervalle en n'écartant pas les termes $O_P\left(1/[n^*]^{3/2}\right)$. Ils le sont ici parce que d'autres termes de cette taille seront abandonnés plus loin dans la présente note.

S'il est raisonnable de laisser tomber les termes $O_P\left(1/[n^*]^{3/2}\right)$ pour obtenir l'équation (2.2), on peut également ignorer sans risque la différence entre $1/n$ et $1/(n-1)$. Sous échantillonnage aléatoire simple *sans* remise, $n^* = n/(1-f)$ (ou $(n-1)/(1-f)$) où f est la fraction d'échantillonnage. Quand f est très petite, on peut omettre la distinction entre l'échantillonnage avec et sans remise.

Observons que, sous échantillonnage aléatoire simple avec remise, le dénominateur du pivot qui figure dans le premier membre de l'équation (2.1) n'a pas de variance du tout. En revanche, le dénominateur du pivot de Wald classique, $\mathrm{var}(p) = p(1-p)/(n-1)$, peut avoir une variance considérable, surtout quand p ou $1-p$ est petit. C'est la raison pour laquelle les intervalles de Wilson donnent de meilleurs résultats sous échantillonnage aléatoire simple, avec ou sans remise.

Cette supériorité s'observe aussi dans le cas de l'échantillonnage complexe (voir, par exemple, Kott, Andersson et Nerman, 2001), où le dénominateur du pivot est

$$\frac{P(1-P)}{n^*} = \mathrm{var}(p)\frac{P(1-P)}{p(1-p)} = \mathrm{var}(p)\left[1 - \frac{(p-P) - \left(p^2 - P^2\right)}{p(1-p)}\right]$$

$$= \mathrm{var}(p)\left[1 - \frac{(p-P) - (p-P)(p+P)}{p(1-p)}\right]$$

$$= \mathrm{var}(p) - \frac{1-2P}{n^*}(p-P) + \mathrm{O}_P\left(1/[n^*]^2\right),$$

dont la variance sera vraisemblablement plus faible que $\text{var}(p)$ dans la plupart des applications. Pour comprendre intuitivement pourquoi il en est ainsi, observons qu'un estimateur de variance putatif de la forme $\text{var}_1(p) = \text{var}(p) - b(p - P)$ est minimisé quand $b = \text{Cov}[\text{var}(p), p]/\text{Var}(p)$. Sous échantillonnage aléatoire simple, avec ou sans remise, b est exactement $(1-2P)/n^*$.

Bien que le b minimisant ne soit pas exactement égal à $(1-2P)/n^*$, sous des plans d'échantillonnage plus complexes, le b optimal est vraisemblablement plus proche de $(1-2P)/n^*$ que de 0. Il n'est donc pas étonnant que la variance de $\text{var}(p) - [(1-2P)/n^*](p-P)$ soit habituellement plus petite que la variance de $\text{var}(p)$. Néanmoins, l'intervalle de couverture de Wilson sous échantillonnage complexe peut être amélioré légèrement en remplaçant n^* dans l'équation (2.2) par

$$\tilde{n} = [(1-2p)\,\text{var}(p)]/\text{cov}[\text{var}(p), p]$$

quand $\text{cov}[\text{var}(p), p]$, un estimateur convergent de $\text{Cov}[\text{var}(p), p]$, existe (voir Kott et coll., 2001).

Comme dans le cas de l'intervalle de Wilson classique, le centre de l'intervalle de Wilson sous échantillonnage complexe dans l'équation (2.2) diffère légèrement de p quand p n'est pas égal à ½ :

$$C = p + \frac{1-2p}{n^*}\,\frac{z_{1-\alpha/2}}{2}.$$

Sa longueur L semble être plus grande que celle de l'intervalle de Wald :

$$L = z_{1-\alpha/2}\left(\frac{p(1-p)}{n^*} + \frac{z^2_{1-\alpha/2}}{4(n^*)^2}\right)^{1/2} > z_{1-\alpha/2}\left(\frac{p(1-p)}{n^*}\right)^{1/2}.$$

Quand $P(1-P) \geq B > 0$, cependant,

$$\begin{aligned}\left(\frac{p(1-p)}{n^*} + \frac{z^2_{1-\alpha/2}}{4(n^*)^2}\right)^{1/2} &= \left(\frac{p(1-p)}{n^*}\right)^{1/2}\left(1 + \frac{\frac{1}{4}z^2_{1-\alpha/2}}{n^* p(1-p)}\right)^{1/2} \\ &= \left(\frac{p(1-p)}{n^*}\right)^{1/2} + \text{o}_p\left(\frac{1}{n^*}\right). \qquad (2.3)\end{aligned}$$

3 La transformation logistique

L'intervalle de couverture de Wilson sous échantillonnage complexe s'avère être très semblable à l'intervalle de couverture bilatéral obtenu en utilisant une transformation logistique (voir Brown et coll., 2001) :

$$f^{-1}\left\{f(p) - z_{1-\alpha/2}\sqrt{\text{var}[f(p)]}\right\} \leq P \leq f^{-1}\left\{f(p) + z_{1-\alpha/2}\sqrt{\text{var}[f(p)]}\right\}, \qquad (3.1)$$

où $f(p) = \log(p) - \log(1-p)$, et $\text{var}[f(p)] = \left[f'(p)\right]^2 \text{var}(p) = [1/p + 1/(1-p)]^2\, p(1-p)/n^* = 1/[n^* p(1-p)]$. La justification originale de cet intervalle semble être qu'il possède la propriété désirable

de ne pas pouvoir contenir des valeurs inférieures à 0 ou supérieures à 1, qui seraient absurdes pour une proportion.

Le premier membre de l'équation (3.1) peut se réécrire sous la forme $g(x-h)$, où

$$g(y) = f^{-1}(y) = [1+\exp(-y)]^{-1},\ x = f(p) = \log\left(\frac{p}{1-p}\right),$$

et

$$h = \frac{z_{1-\alpha/2}}{\sqrt{n^* p(1-p)}}.$$

Les dérivées première et seconde de $g(y)$ sont $g'(y) = g(y)[1-g(y)]$ et $g''(y) = g(y)[1-g(y)][1-2g(y)]$. En vertu du théorème de la valeur moyenne, il existe un h^* entre 0 et h tel que

$$\begin{aligned} g(x-h) &= g(x) - g'(x)\,h + \frac{1}{2}\,g''(x-h^*)\,h^2 \\ &= p - p(1-p)\frac{z_{1-\alpha/2}}{\sqrt{n^* p(1-p)}} \\ &+ \frac{1}{2}\left[1+\left(\frac{1-p}{p}\right)e^{h^*}\right]^{-1}\left\{1-\left[1+\left(\frac{1-p}{p}\right)e^{h^*}\right]^{-1}\right\}\left\{1-2\left[1+\left(\frac{1-p}{p}\right)e^{h^*}\right]^{-1}\right\}\frac{z^2_{1-\alpha/2}}{n^* p(1-p)} \\ &= p - p(1-p)\frac{z_{1-\alpha/2}}{\sqrt{n^* p(1-p)}} \\ &+ \frac{1}{2}\,\frac{p}{1+(1-p)\left(e^{-h^*}-1\right)}\,\frac{(1-p)-(1-p)\left(e^{h^*}-1\right)}{1+(1-p)\left(e^{h^*}-1\right)}\,\frac{(1-2p)-(1-p)\left(e^{h^*}-1\right)}{1+(1-p)\left(e^{h^*}-1\right)}\,\frac{z^2_{1-\alpha/2}}{n^* p(1-p)}, \end{aligned}$$

en utilisant

$$\left[1+\left(\frac{1-p}{p}\right)e^{h^*}\right]^{-1} = \frac{p}{1+(1-p)\left(e^{h^*}-1\right)}.$$

Un calcul analogue peut être effectué pour le deuxième membre de l'équation (3.1).

Par conséquent,

$$\begin{aligned} p + \frac{1-2p}{n^*}\,\frac{z^2_{1-\alpha/2}}{2} - z_{1-\alpha/2}\left(\frac{p(1-p)}{n^*}\right)^{1/2} + o_P\left(\frac{1}{n^*}\right) &\le P \\ &\le p + \frac{1-2p}{n^*}\,\frac{z^2_{1-\alpha/2}}{2} + z_{1-\alpha/2}\left(\frac{p(1-p)}{n^*}\right)^{1/2} + o_P\left(\frac{1}{n^*}\right). \end{aligned}$$

En vertu de l'égalité asymptotique dans l'équation (2.3) et en laissant tomber les termes $o_P(1/n^*)$, le dernier ensemble d'inégalités équivaut à l'intervalle de Wilson dans l'équation (2.2) à condition que n^* soit suffisamment grand et que $P(1-P) > 0$, cette dernière contrainte signifiant que la proportion vraie n'est égale ni à 0 ni à 1.

4 Certaines conclusions

L'équivalence asymptotique d'un intervalle de couverture fondé sur une transformation logistique et de l'intervalle de Wilson reposant sur la théorie constitue la principale contribution du présent article. Même si, dans le cadre asymptotique, $P(1-P)$ est fixé et positif quand n^* devient grand, en pratique, c'est la taille de $p(1-p)n^*$ qui importe lorsque l'on compare les intervalles de type Wilson à ceux fondés sur la transformation logistique. Il faut pour cela que $P(1-P)$ ne soit pas trop petit.

Brown et coll. (2001) montrent empiriquement que, sous échantillonnage aléatoire simple (avec $n = 50$), les intervalles de couverture calculés en partant de la transformation logistique ont tendance à être plus grands que les intervalles de Wilson correspondants pour les petites valeurs de $P(1-P)$. Kott et Liu (2009) font la même constatation pour des intervalles unilatéraux fondés sur des échantillons complexes, ce qui appuie la notion qu'il s'agit d'un meilleur choix.

L'équivalence asymptotique de l'intervalle par transformation logistique et de l'intervalle de Wilson explique la supériorité empirique du premier signalée dans la littérature (par exemple, Brown et coll., 2001) comparativement à un intervalle analogue construit en utilisant une transformation arcsinus. Comme $\arcsin(p)$ possède une variance constante en grand échantillon sous échantillonnage aléatoire simple, peu importe la valeur vraie de P (à condition que $P(1-P) > 0$), d'aucuns espéraient que la transformation arcsinus soit un moyen idéal de construire les intervalles.

Les intervalles de couverture unilatéraux pour P calculés en utilisant un développement d'Edgeworth sur $p - P$ décrits dans Kott et Liu (2009) sont meilleurs qu'un intervalle de Wilson, mais, autant que je sache, ils n'ont encore été incorporés dans aucun progiciel. Cette méthode produit l'intervalle bilatéral qui suit :

$$p + \frac{1-2p}{\tilde{n}}\left(\frac{1}{6} + \frac{z_{1-\alpha/2}^2}{3}\right) - z_{1-\alpha/2}\left(\text{var}(p) + \left[\frac{1-2p}{\tilde{n}}\left(\frac{1}{6} + \frac{z_{1-\alpha/2}^2}{3}\right)\right]^2\right)^{1/2} \leq P$$

$$\leq p + \frac{1-2p}{\tilde{n}}\left(\frac{1}{6} + \frac{z_{1-\alpha/2}^2}{3}\right) + z_{1-\alpha/2}\left(\text{var}(p) + \left[\frac{1-2p}{\tilde{n}}\left(\frac{1}{6} + \frac{z_{1-\alpha/2}^2}{3}\right)\right]^2\right)^{1/2},$$

où $\tilde{n} = [(1-2p)\,\text{var}(p)] / \text{cov}[\text{var}(p),\ p]$, et $\text{cov}[\text{var}(p),\ p]$, un estimateur convergent de $\text{Cov}[\text{var}(p),\ p]$, existe et est égal à un estimateur convergent du troisième moment de p. Notons que $\text{cov}[\text{var}(p),\ p]$ n'existe pas pour les plans ne comportant que deux unités primaires d'échantillonnage par strate. En outre, il ne s'agit pas d'un estimateur convergent pour le troisième moment de p quand il importe de faire une correction pour population finie.

Observons que $\tilde{n}$ remplace de nouveau n^*. De plus, $1/6 + z_{1-\alpha/2}/3$ remplace $z_{1-\alpha/2}/2$, ce qui signifie que le centre sera souvent plus proche de p quand on utilise cet intervalle plutôt que celui de Wilson. Les bonnes propriétés de couverture de cet intervalle disparaissent, comme dans le cas de l'intervalle de Wilson, quand le coefficient d'asymétrie de $p\left(\text{E}\left[(p-P)^3\right] / \left[\text{Var}(p)\right]^{3/2}\right)$ devient trop grand en valeur absolue, quoiqu'il faille encore déterminer le seuil de grandeur excessive.

Enfin, SAS/STAT (SAS Institute Inc., 2010) offre un intervalle de couverture de Wilson pour les proportions estimées dans sa procédure SURVEYFREQ. La méthode d'ajustement de la taille effective

d'échantillon de la procédure, que l'on peut – et doit – désactiver, n'est pas reliée au $\tilde{n}$ discuté ici. Elle est plutôt fondée sur un ajustement t ponctuel qui, malheureusement, n'est pas relié à la variance de la variance du dénominateur du pivot de Wilson.

Remerciements

L'auteur remercie Per Gösta Andersson de lui avoir fait découvrir ce domaine de recherche et un examinateur anonyme d'avoir corrigé les erreurs dans une version antérieure du manuscrit. Les erreurs qui persistent sont imputables à l'auteur.

Bibliographie

Brown, L.D., Cai, T. et Dasgupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16, 101-133.

Kott, P.S., et Carr, D.A. (1997). Developing an estimation strategy for a pesticide data program. *Journal of Official Statistics*, 13, 367-383.

Kott, P.S., et Liu, Y.K. (2009). One-sided coverage intervals for a proportion estimated from a stratified simple random sample. *International Statistical Review*/*Revue Internationale de Statistique*, 77, 251-265.

Kott, P.S., Andersson, P.G. et Nerman, O. (2001). Two-sided coverage intervals for small proportion based on survey data. Présenté à la Federal Committee on Statistical Methodology Research Conference, Washington, DC. http://fcsm.sites.usa.gov/files/2014/05/2001FCSM_Kott.pdf.

SAS Institute Inc. (2010). *SAS/STAT® 9.22 User's Guide*. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_surveyfreq_a0000000252.htm.

WesVar (2007). *WesVar® 4.3 Users' Guide*, B28-B29.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22, 209-212.

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article ou plus durant l'année 2017.

P. Ardilly, *INSEE*
J.-F. Beaumont, *Statistique Canada*
A. Bianchi, *University of Bergamo*
P. Biemer, *RTI International*
H.J. Boonstra, *Statistics Netherlands*
J.M. Brick, *Westat Inc.*
P. Brodie, *Office for National Statistics*
P.J. Cantwell, *U.S. Bureau of the Census*
H. Cardot, *Université de Bourgogne*
K.-C. Chang, *Fu Jen Catholic University*
G. Chauvet, *ENSAI/IRMAR*
B. Chen, *Bureau of Economic Analysis*
J. Chipperfield, *Australian Bureau of Statistics*
T. Deroyon, *INSEE*
J. Dever, *RTI International*
L. Dudoignon, *Médiamétrie*
D. Elliot, *Office of National Statistics*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
S. Er, *University of Cape Town*
S. Falorsi, *ISTAT*
S. Fortier, *Statistique Canada*
C. Franco, *U.S. Census Bureau*
W.A. Fuller, *Iowa State University*
M. Furno, *Università degli Studi di Napoli 'Federico II'*
J. Gambino, *Statistique Canada*
C. Girard, *Statistique Canada*
C. Goga, *Université de Bourgogne*
A. Goia, *Università del Piemonte Orientale*
A. Grafström, *Sweedish University of Agricultural Studies*
E. Gros, *INSEE*
R. Gutman, *Brown University*
D. Haziza, *Université de Montréal*
Y. He, *National Center for Health Statistics*
S.G. Heeringa, *University of Michigan*
M. Hidiroglou, *Statistique Canada*
J. Hua, *Dartmouth College*
B. Hulliger, *University of Applied Sciences Northwestern Switzerland*
J. Im, *Iowa State University*
D. Judkins, *Abt Associates*
E. Kabzinska, *University of Southampton*
J.K. Kim, *Iowa State University*
N. Kim, *Carnegie Mellon University*
P. Kott, *RTI International*
M. Kozak, *University of Inf. Techmology and Management in Rzeszow*
P. Lahiri, *University of Maryland*
A. Lauger, *U.S. Census Bureau*
P. Lavallée, *Statistique Canada*
C. Léger, *Université de Montréal*
J. Legg, *Amagen Inc.*
E. Lesage, *INSEE*
R. Lethonen, *Université d'Helsinki*
D. Liao, *RTI International*
S. Lohr, *Westat Inc.*
A.G. Matei, *Université de Neuchâtel*
K. McConville, *Swarthmore College*
I. Molina, *Universidad Carlos III de Madrid*
P.K. Mukhopadhyay, *SAS Institute Inc.*
C.O. Nambeu, *Statistique Canada*
E. Neusy, *Statistique Canada*
J. Opsomer, *Colorado State University*
P. Ouwehand, *Statistics Netherland*
D. Pavlopoulos, *Free university of Amsterdam*
D. Pfeffermann, *Hebrew University*
R. Powers, *U.S. Bureau of Labor Statistics*
J.N.K. Rao, *Carleton University*
A. Rebecq, *INSEE*
L.-P. Rivest, *Université Laval*
A. Ruiz-Gazen, *Université Toulouse*
F. Scheuren, *National Opinion Research Center*
T. Schmid, *Freie Universität Berlin*
P.L.N.D. Silva, *Escola Nacional de Ciências Estatísticas*
P. Smith, *University of Southampton*
D. Steel, *University of Wollongong*
M. Sverchkov, *U.S. Bureau of Labor Statistics*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
R.B. Tiller, *U.S. Bureau of Labor Statistics*
D. Toth, *U.S. Bureau of Labor Statistics*
J. van den Brakel, *Statistics Netherland and Maastricht University*
P. van Kerm, *Luxembourg Institute of Socio-Economic Research*
V. Vehovar, *University of Ljubljana*
B.T. West, *University of Michigan-Ann Arbor*
C. Wu, *University of Waterloo*
S. Yang, *North Carolina State University*
Y. You, *Statistique Canada*
A. Zaslavsky, *Harvard University*
G. Zhang, *University of New Mexico*
L.-C. Zhang, *University of Southampton*
T. Zimmermann, *German Federal Statistical Office*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2017 : Céline Ethier de la Division de la coopération internationale et des méthodes statistiques institutionnelles; Joana Bérubé de la Division des méthodes d'enquêtes auprès des entreprises; l'équipe de la Division de la diffusion, en particulier Chantal Chalifoux, Christina Jaworski, Kathy Charbonneau, Jacqueline Luffman, Jenna Waite, Darquise Pellerin et Joseph Prince de même que nos partenaires de la Division des communications.

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. 2, June 2017

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. 3, September 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. 3, September/septembre 2017

Issue Information

Review Article

Original Articles

CONTENTS TABLE DES MATIÈRES

Volume 45, No. 4, December/décembre 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.