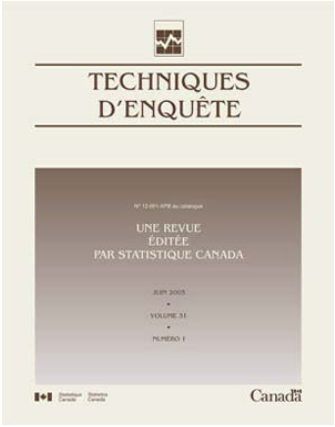
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles ?



par Jean-François Beaumont

Date de diffusion : le 30 juin 2020

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 suivants :

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

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.

Techniques d'enquête, juin 2020
Vol. 46, nº 1, pp. 1-30
Statistique Canada, nº 12-001-X au catalogue

Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles ?

Jean-François Beaumont[1]

Résumé

Depuis plusieurs décennies, les agences nationales de statistique dans le monde utilisent des enquêtes probabilistes comme outil privilégié pour répondre à des besoins d'informations au sujet d'une population d'intérêt. Au cours des dernières années, on a observé un vent de changement et on considère de plus en plus d'autres sources de données. Cette tendance peut être expliquée par cinq facteurs principaux : le déclin des taux de réponse dans les enquêtes probabilistes, les coûts de collecte élevés, l'accroissement du fardeau sur les répondants, le désir d'avoir accès à des statistiques en « temps réel » et la prolifération des sources de données non probabilistes. Certaines personnes en sont même venues à croire que les enquêtes probabilistes pourraient graduellement disparaître. Dans cet article, nous passons en revue quelques approches qui permettent de réduire, voire éliminer, l'utilisation d'enquêtes probabilistes tout en conservant un cadre d'inférence statistique valide. Toutes les approches que nous considérons utilisent des données d'une source non probabiliste accompagnées, dans la plupart des cas, de données d'une enquête probabiliste. Certaines d'entre elles reposent sur la validité d'hypothèses de modèle ce qui contraste avec les approches fondées sur le plan de sondage probabiliste. Ces dernières sont généralement moins efficaces mais, en contrepartie, elles ne sont pas affectées par le risque de biais découlant d'une mauvaise spécification d'un modèle.

Mots-clés : Appariement statistique; Calage; Données non probabilistes; Intégration de données; Modèle de Fay-Herriot; Score de propension.

1 Introduction

En 1934, Jerzy Neyman posait les fondements de la théorie des enquêtes probabilistes et de son approche d'inférence fondée sur le plan de sondage avec la parution d'un article publié dans la revue *Journal of the Royal Statistical Society*. L'article de Neyman (1934) suscita l'intérêt de plusieurs statisticiens de l'époque et la théorie fut développée plus en profondeur dans les années subséquentes. Encore aujourd'hui, on trouve de nombreux articles sur ce sujet dans les revues de statistique. On réfère le lecteur à Rao (2005) pour une excellente revue de différents développements de la théorie des enquêtes probabilistes au cours du vingtième siècle (voir aussi Bethlehem, 2009; Rao et Fuller, 2017; et Kalton, 2019). De nos jours, les agences nationales de statistique, comme Statistique Canada et l'Institut National de la Statistique et des Études Économiques (INSEE) en France, ont la plupart du temps recours à des enquêtes probabilistes pour obtenir l'information désirée sur une population d'intérêt.

La popularité des enquêtes probabilistes pour la production de statistiques officielles découle en grande partie du caractère non paramétrique de l'approche d'inférence élaborée par Neyman (1934). En d'autres mots, les enquêtes probabilistes permettent de faire des inférences valides sur une population sans avoir recours à des hypothèses de modèle. C'est une propriété attrayante, voire même indispensable selon Deville (1991), pour les agences nationales de statistique qui produisent des statistiques officielles. Ces agences ont d'ailleurs été historiquement réticentes à la prise inutile de risques inhérente aux approches dépendant de la validité d'hypothèses de modèle, surtout lorsque celles-ci sont difficilement vérifiables.

1. Jean-François Beaumont, Statistique Canada, Édifice R.H. Coats, 100 promenade Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6. Courriel : jean-francois.beaumont@canada.ca.

Les estimations issues d'enquêtes probabilistes peuvent cependant s'avérer inefficaces, au point même d'être inutilisables, particulièrement lorsque la taille d'échantillon est petite (voir, par exemple, Rao et Molina, 2015). De plus, elles reposent sur l'hypothèse que les erreurs non dues à l'échantillonnage, telles que les erreurs de mesure, de couverture ou de non-réponse, sont négligeables. Afin de minimiser ces erreurs, les agences nationales de statistique mettent souvent beaucoup de ressources en oeuvre. Par exemple, les questionnaires sont testés pour s'assurer qu'ils sont bien compris par les répondants, les données de l'enquête sont validées au moyen de différentes règles de vérification, les répondants sont recontactés au besoin pour vérifier les données recueillies, un suivi des non-répondants est effectué pour minimiser l'impact de la non-réponse sur les estimations, etc. Malgré tous ces efforts, des erreurs non dues à l'échantillonnage subsistent en pratique. Il existe bien sûr des adaptations de la théorie pour tenir compte de ces erreurs. Ces adaptations sont nécessairement accompagnées par l'introduction d'hypothèses de modèle et ainsi par le risque de biais résultant de l'inadéquation des hypothèses. Les enquêtes probabilistes ne sont donc pas une panacée mais on reconnaît généralement qu'elles représentent une source fiable d'informations sur une population sauf dans les cas où les erreurs non dues à l'échantillonnage deviennent prépondérantes. Brick (2011) pousse l'argument plus loin et défend l'idée qu'une enquête probabiliste avec un faible taux de réponse fournit des estimations généralement moins biaisées, si elle est bien conçue, qu'une enquête non probabiliste menée auprès de volontaires. Dutwin and Buskirk (2017) montrent des résultats empiriques qui corroborent cet argument.

Depuis quelques années, un vent de changement souffle sur les agences nationales de statistique et on considère de plus en plus d'autres sources de données. Cette tendance peut être expliquée par cinq facteurs principaux : i) le déclin des taux de réponse dans les enquêtes probabilistes au cours des dernières années; ii) les coûts de collecte élevés; iii) l'accroissement du fardeau sur les répondants; iv) le désir d'avoir accès à des statistiques en « temps réel » (Rao, 2020), c'est-à-dire de pouvoir produire des statistiques pratiquement au même moment ou très peu de temps après que le besoin d'informations ait été formulé; et v) la prolifération de sources de données non probabilistes (Rancourt, 2019) telles que les sources administratives, les médias sociaux, les enquêtes Web, etc. Afin de contrôler les coûts de collecte des enquêtes probabilistes et réduire les effets indésirables de la non-réponse sur la qualité des estimations, plusieurs auteurs ont proposé et évalué des méthodes de collecte adaptatives (ex. : Laflamme et Karaganis, 2010; Lundquist et Särndal, 2013; Schouten, Calinescu et Luiten, 2013; Beaumont, Haziza et Bocci, 2014; et Särndal, Lumiste et Traat, 2016). Tourangeau, Brick, Lohr et Li (2017) passent en revue différentes méthodes et soulignent leur succès mitigé à réduire le biais de non-réponse et les coûts. Särndal et coll. (2016) en arrivent également à la même conclusion en ce qui a trait au biais. Pour certaines enquêtes menées par des agences nationales de statistique, on observe encore des taux de réponse très faibles et il devient hasardeux de se fier uniquement aux méthodes de collecte et d'estimation pour corriger les biais potentiels de non-réponse. Plusieurs auteurs (ex. : Rivers, 2007; Elliott et Valliant, 2017) soulignent d'ailleurs la ressemblance entre une enquête probabiliste avec un très faible taux de réponse et une enquête non probabiliste. Cette dernière possède cependant l'avantage d'avoir une taille d'échantillon généralement beaucoup plus grande tout en étant moins coûteuse. Considérant les éléments discutés ci-

dessus, certaines personnes en sont venues à croire que les enquêtes probabilistes pourraient graduellement disparaître (voir Couper, 2000; Couper, 2013; et Miller, 2017).

Les données de sources non probabilistes ne viennent toutefois pas sans défis, tel que noté entre autres par Couper (2000), Baker, Brick, Bates, Battaglia, Couper, Dever, Gile et Tourangeau (2013) et Elliott et Valliant (2017). Par exemple, il est bien connu que les enquêtes non probabilistes recueillant des données auprès de volontaires peuvent souvent mener à des estimations entachées d'un biais de sélection (ou biais de participation) important. Bethlehem (2016) donne une expression du biais et argue que le risque de biais est généralement plus élevé pour une enquête non probabiliste que pour une enquête probabiliste affectée par la non-réponse. Meng (2018) illustre que le biais devient dominant à mesure que la taille de l'échantillon non probabiliste augmente ce qui réduit considérablement la taille d'échantillon effective. Par conséquent, l'acquisition d'échantillons non probabilistes de grande taille ne peut pas assurer à elle seule la production d'estimations de qualité acceptable. Le sondage préélectoral mené par la revue *Literary Digest* visant à prédire le résultat de l'élection présidentielle américaine de 1936 en est un exemple marquant (Squire, 1988; Elliott et Valliant, 2017). Malgré une taille d'échantillon gigantesque de plus de deux millions de personnes, le sondage ne put prédire la victoire éclatante de Franklin Roosevelt. Il prédit plutôt incorrectement une victoire convaincante pour son adversaire, Alfred Landon. L'ensemble des répondants au sondage, fortement non représentatif de la population d'électeurs, était constitué principalement de propriétaires d'automobiles et de téléphones de même que des abonnés à la revue. Couper (2000) et Elliott et Valliant (2017) citent d'autres exemples plus récents de sondages non probabilistes qui ont mené à des conclusions erronées.

Le biais de sélection n'est pas le seul défi qui doit être relevé quand on utilise des données d'une source non probabiliste. Un autre défi de taille est la présence d'erreurs de mesure (ex. : Couper, 2000). Elles peuvent avoir un impact significatif sur les estimations, particulièrement lorsque les données sont recueillies sans avoir recours à un interviewer expérimenté. C'est le cas de la plupart des sources non probabilistes, notamment les enquêtes Web menées auprès de volontaires.

Le contexte actuel amène à se poser la question suivante : Comment peut-on utiliser des données d'une source non probabiliste afin de minimiser, voire éliminer, les coûts de collecte et le fardeau sur les répondants d'une enquête probabiliste tout en conservant un cadre d'inférence statistique valide et une qualité acceptable ? C'est la question principale à laquelle cet article tente de répondre.

La plupart des méthodes que nous exposerons intègrent des données d'une enquête probabiliste et d'une source non probabiliste. Zhang (2012) discute du concept de validité statistique lorsque des données intégrées sont utilisées pour faire les inférences. Nous soutenons que la détermination d'un cadre statistique qui permette de faire des inférences valides est essentielle pour la production de statistiques officielles, un point qui semble également partagé par Rancourt (2019). Sans un tel cadre, les propriétés habituelles des estimateurs comme le biais et la variance ne sont pas définies. Il devient alors impossible de choisir les estimateurs selon un critère objectif tel que, par exemple, choisir l'estimateur linéaire sans biais qui a la plus petite variance possible. Sans un cadre d'inférence statistique valide, on peut calculer

des estimations mais on perd tous les outils usuels pour déterminer la qualité de ces estimations et tirer des conclusions justes sur les caractéristiques d'intérêt de la population.

Dans le reste de cet article, nous distinguerons les approches d'inférence fondées sur le plan de sondage, décrites à la section 3, des approches d'inférence fondées sur un modèle, décrites à la section 4. Pour chacune des approches, nous considérerons deux scénarios : Dans le premier, les données de la source non probabiliste correspondent exactement aux concepts d'intérêt et ne sont pas entachées d'erreurs de mesure. Ces données peuvent donc être utilisées pour remplacer des données d'une enquête probabiliste. Dans le deuxième scénario, les données de la source non probabiliste ne reflètent pas les concepts d'intérêt ou sont sujettes aux erreurs de mesure. Bien que de telles données ne puissent pas être utilisées pour remplacer directement des données d'une enquête probabiliste, elles peuvent néanmoins être utilisées comme informations auxiliaires pour l'enrichir. À la section 5, nous fournirons quelques réflexions supplémentaires. Commençons tout d'abord avec une mise en contexte à la section 2.

2 Contexte

Une des premières étapes pour répondre à des besoins d'informations est de définir la population cible pour laquelle on désire obtenir ces informations. Nous allons noter cette population cible par U. Ensuite, il faut définir les paramètres d'intérêt, c'est-à-dire ce qu'on veut savoir sur la population cible. En pratique, on est souvent intéressé à estimer un grand nombre de paramètres. Pour simplifier la discussion, nous allons supposer qu'un seul paramètre est d'intérêt : le total de la variable y, $\theta = \sum_{k \in U} y_k$, où y_k est la valeur de la variable y pour l'unité k de la population U. On note par $\mathbf{Y}$, le vecteur contenant les valeurs y_k pour $k \in U$. Finalement, il faut déterminer un ensemble de procédures qui permettront d'estimer le paramètre θ en tenant compte de différents facteurs tels que le budget disponible, le fardeau sur les répondants, la précision souhaitée, etc. Au cours de ce processus, il faudra identifier la ou les sources de données qui seront utilisées, probabilistes ou non, et un cadre d'inférence statistique qui permettra d'évaluer les propriétés des estimations produites telles que le biais et la variance.

La séquence ci-dessus, qui consiste d'abord à déterminer la population cible et les paramètres d'intérêt et ensuite les sources de données et les procédures d'estimation, est en accord avec la proposition de Citro (2014). Elle suggère que les agences nationales de statistique déterminent d'abord les besoins d'informations avec les utilisateurs potentiels. Ensuite, elles peuvent travailler à identifier la ou les sources de données qui permettront de répondre à ces besoins tout en préservant une qualité acceptable des estimations, en maintenant les coûts à l'intérieur du budget établi et en contrôlant le fardeau sur les répondants. Il semble préférable d'éviter la procédure inverse, même si tentante, qui consiste d'abord à identifier des sources de données disponibles et ensuite à déterminer artificiellement les besoins en fonction de ce qui peut être produit par ces sources. Une telle procédure ne peut généralement pas permettre de répondre adéquatement aux besoins réels des utilisateurs.

On va supposer qu'on a accès à des données d'une source non probabiliste (ex. : données administratives, données d'une enquête Web, etc.). On observe les valeurs de quelques variables, dont une variable y^*, pour toutes les unités d'un sous-ensemble de U, noté par s_{NP}. La variable y^* n'est pas

nécessairement égale à y à cause de différences conceptuelles et/ou d'erreurs de mesure. On espère au moins qu'il existe une forte association entre les deux variables. On note par δ_k, l'indicateur d'inclusion dans s_{NP}; c'est-à-dire, $\delta_k = 1$ si l'unité k est dans s_{NP} et $\delta_k = 0$, autrement. Le vecteur des indicateurs d'inclusion δ_k pour $k \in U$ est noté par $\boldsymbol{\delta}$.

On peut aussi avoir accès à des données d'une enquête probabiliste. Pour une telle enquête, un échantillon s_P de la population U est sélectionné aléatoirement avec probabilité $p(s_P \mid \mathbf{Z})$. La matrice $\mathbf{Z}$ contient des informations disponibles sur la base de sondage qui sont utilisées pour définir le plan de sondage, comme, par exemple, les identificateurs de strates pour chaque unité de la population. Les indicateurs d'inclusion dans l'échantillon, $I_k, k \in U$, sont définis comme suit : on pose $I_k = 1$ si l'unité k est choisie dans l'échantillon s_P; autrement, on pose $I_k = 0$. On note par $\mathbf{I}$, le vecteur contenant les indicateurs d'inclusion dans l'échantillon pour $k \in U$. La probabilité que l'unité k de la population U soit choisie dans l'échantillon est notée par $\pi_k = E(I_k \mid \mathbf{Z})$. La plupart du temps, elle est connue ou elle peut être approchée. On suppose que $\pi_k > 0, k \in U$. Pour chaque unité $k \in s_P$, on recueille les valeurs de certaines variables, incluant ou non la variable y.

On va noter par $\boldsymbol{\Omega}$, l'ensemble de toutes les informations auxiliaires utilisées pour faire les inférences. Entre autres, $\boldsymbol{\Omega}$ inclut les informations du plan de sondage, $\mathbf{Z}$, si un échantillon probabiliste est utilisé, et possiblement d'autres variables auxiliaires telles que des variables de calage, des variables d'appariement ou des variables explicatives d'un modèle (voir sections 3 et 4). L'indicateur d'inclusion δ_k peut aussi être utilisé comme variable auxiliaire soit pour stratifier la population ou pour le calage (voir section 3). Le vecteur $\boldsymbol{\delta}$ peut donc être inclus dans $\mathbf{Z}$ et $\boldsymbol{\Omega}$.

Les deux hypothèses suivantes seront utilisées tout au long de l'article :

Hypothèse 1 : $\mathbf{I}$ est indépendant de $\boldsymbol{\Omega}$ et $\mathbf{Y}$ après avoir conditionné sur $\mathbf{Z}$.
Hypothèse 2 : $\boldsymbol{\delta}$ et $\mathbf{I}$ sont indépendants après avoir conditionné sur $\boldsymbol{\Omega}$ et $\mathbf{Y}$.

L'hypothèse 1 implique que les valeurs des variables incluses dans $\boldsymbol{\Omega}$ et $\mathbf{Y}$ ne sont pas affectées par l'appartenance ou non à l'échantillon s_P. Elle est implicite dans la littérature sur les sondages probabilistes et résulte de la définition même du plan de sondage qui ne dépend que de $\mathbf{Z}$. L'hypothèse 2 est automatiquement satisfaite si la source non probabiliste (et ainsi $\boldsymbol{\delta}$) est disponible avant la sélection de l'échantillon probabiliste. Notons que si δ_k est utilisé comme variable auxiliaire pour stratifier la population alors $\boldsymbol{\delta}$ est inclus dans $\boldsymbol{\Omega}$ et l'hypothèse 2 reste satisfaite. Elle ne sera pas satisfaite si le fait d'être choisi dans s_P a une incidence sur le fait de fournir des données à la source non probabiliste. Par exemple, le fait d'être choisi dans s_P (et contacté) peut être un rappel indirect pour l'individu sélectionné de remplir des formulaires exigés par le gouvernement (source non probabiliste). On peut s'attendre à ce que les hypothèses 1 et 2 soient satisfaites dans la plupart des cas.

L'union de $\boldsymbol{\Omega}$, $\boldsymbol{\delta}$, $\mathbf{I}$ et $\mathbf{Y}$ contient toutes les informations utilisées pour faire les inférences. Les différentes approches d'inférence exposées aux sections 3 et 4 diffèrent dans ce qu'elles traitent comme étant fixe et ce qu'elles traitent comme étant aléatoire. Par exemple, dans l'approche d'inférence fondée sur le plan de sondage, tout est considéré comme étant fixe à l'exception du vecteur $\mathbf{I}$; c'est-à-dire que l'inférence est conditionnelle à $\boldsymbol{\Omega}$, $\mathbf{Y}$ et $\boldsymbol{\delta}$. Pour simplifier la notation, on va noter par $\boldsymbol{\Omega}_P$ l'union de

$\boldsymbol{\Omega}$, $\mathbf{Y}$ et $\boldsymbol{\delta}$. Ainsi, les espérances par rapport au plan de sondage seront notées par $E(\cdot \mid \boldsymbol{\Omega}_P)$ plutôt que par $E(\cdot \mid \boldsymbol{\Omega}, \mathbf{Y}, \boldsymbol{\delta})$. Dans l'approche d'inférence fondée sur le plan de sondage, on choisit habituellement un estimateur $\hat{\theta}$ de θ tel que le biais par rapport au plan, $E\left(\hat{\theta} - \theta \mid \boldsymbol{\Omega}_P\right)$, est nul ou négligeable. Sous les hypothèses 1 et 2, on note que $E(I_k \mid \boldsymbol{\Omega}_P) = E(I_k \mid \mathbf{Z}) = \pi_k$. Pour l'estimation du total $\theta = \sum_{k \in U} y_k$, on utilise la plupart du temps un estimateur de la forme $\hat{\theta} = \sum_{k \in s_P} w_k y_k$, où w_k est un poids de sondage pour l'unité k. Le poids de base classique est $w_k = \pi_k^{-1}$. Ce poids assure que l'estimateur $\hat{\theta}$ est exactement sans biais pour θ. On peut ensuite modifier le poids de base au moyen de techniques de calage (ex. : Deville et Särndal, 1992; Haziza et Beaumont, 2017). L'avantage de cette approche est son caractère non paramétrique : aucune hypothèse de modèle n'est nécessaire pour faire des inférences valides sur la population puisque les deux premiers moments du plan de sondage sont contrôlés par le statisticien et habituellement connus. L'approche n'est toutefois pas exempte d'hypothèses, par exemple, pour assurer la convergence et la normalité asymptotique des estimateurs, mais elle ne requiert aucun modèle paramétrique.

En pratique, on observe souvent de la non-réponse dans les enquêtes probabilistes de même que d'autres erreurs non dues à l'échantillonnage. La non-réponse de certaines unités échantillonnées est souvent vue comme une phase additionnelle d'échantillonnage qui n'est pas contrôlée par le statisticien. Autrement dit, le mécanisme de non-réponse n'est pas connu, contrairement au plan de sondage. Sous l'hypothèse d'un modèle adéquat pour le mécanisme de non-réponse, on peut construire des estimateurs qui n'ont peu ou pas de biais, par exemple, en repondérant les unités répondantes par l'inverse de leur probabilité de réponse estimée. Cela requiert toutefois une modélisation minutieuse des indicateurs de réponse. Dans le reste de cet article, on ignore les erreurs non dues à l'échantillonnage et on suppose que les estimations provenant de l'enquête probabiliste sont peu biaisées ou, au moins, que leur biais est petit par rapport au biais des estimations provenant seulement de la source non probabiliste. Cette hypothèse n'est peut-être pas toujours satisfaite en pratique mais elle est raisonnable dans bien des contextes (voir Brick, 2011), en particulier dans les grandes enquêtes menées par les agences nationales de statistique.

L'acquisition de données de sources non probabilistes s'avère généralement peu coûteuse en comparaison du coût de la collecte des données d'une enquête probabiliste. Alors, on aimerait idéalement les utiliser pour remplacer les données d'une enquête probabiliste. Ce remplacement de données n'est valide que si $y_k^* \approx y_k$, $k \in U$. Cette hypothèse ne sera pas satisfaite avec toutes les sources de données non probabilistes mais on peut penser qu'elle est réaliste avec certaines sources de données administratives. Aux sections 3 et 4, nous distinguerons les méthodes fondées sur l'hypothèse que $y_k^* = y_k$ des méthodes ne requérant pas cette hypothèse. Plusieurs des méthodes décrites aux sections 3 et 4 sont également passées en revue dans l'article à venir de Rao (2020) qui a été présenté, entre autres, à Statistique Canada à l'été 2018.

3 Approches fondées sur le plan de sondage

Les approches fondées sur le plan de sondage permettent de construire des estimateurs de θ convergents par rapport au plan de sondage même quand la source non probabiliste produit des estimations comportant un

biais de sélection important. Dans ce contexte, l'utilisation d'un échantillon non probabiliste a pour objectif une réduction de la variance des estimateurs de θ. Les gains d'efficacité réalisés peuvent être utilisés pour justifier une réduction de la taille de l'échantillon probabiliste, réduisant ainsi les coûts de collecte et le fardeau sur les répondants. Les méthodes que nous considérons aux sections 3.1 et 3.2 nécessitent de recueillir les valeurs de la variable d'intérêt y dans l'échantillon probabiliste tout comme les méthodes d'estimation sur petits domaines décrites à la section 4.4. On peut toutefois s'attendre à ce que les gains d'efficacité soient généralement plus modestes que ceux obtenus au moyen des méthodes d'estimation sur petits domaines. À la section 3.1, on considère le scénario $y_k^* = y_k$ tandis qu'à la section 3.2, on considère le scénario $y_k^* \neq y_k$.

3.1 Pondération par l'inverse de la probabilité d'inclusion dans l'échantillon combiné

Le cas idéal survient quand $s_{\text{NP}} = U$. On a alors un recensement et on peut calculer directement la valeur du paramètre d'intérêt $\theta = \sum_{k \in U} y_k$ sans se soucier du biais ou de la variance puisque, dans cette section, on travaille sous l'hypothèse que $y_k^* = y_k$. En général, on se trouvera plutôt dans un contexte de sous-couverture au sens où s_{NP} est de plus petite taille que la population U. Dans une approche fondée sur le plan de sondage, on peut remédier au biais potentiel de sous-couverture en sélectionnant un échantillon probabiliste s_P à partir de U et en recueillant les valeurs de la variable y pour les unités échantillonnées. Idéalement, l'échantillon probabiliste est tiré de $U - s_{\text{NP}}$ mais il est possible que les unités dans s_{NP} ne puissent pas être couplées à celles de la base de sondage U pour ainsi déterminer l'ensemble $U - s_{\text{NP}}$. En général, plus la taille de l'échantillon non probabiliste sera grande, plus il sera possible de réduire la taille de l'échantillon probabiliste sans compromettre la précision souhaitée des estimations.

On aimerait pouvoir estimer θ en utilisant toutes les données recueillies dans l'échantillon combiné $s = s_P \cup s_{\text{NP}}$. On peut définir l'indicateur d'inclusion dans s comme étant $\tilde{I}_k = \delta_k + (1 - \delta_k) I_k$. Pour obtenir une estimation sans biais de θ, on pondère chaque unité $k \in s$ par $\tilde{w}_k = \tilde{\pi}_k^{-1}$ où $\tilde{\pi}_k = E(\tilde{I}_k \mid \mathbf{\Omega}_P)$. Sous les hypothèses 1 et 2, $E(I_k \mid \mathbf{\Omega}_P) = \pi_k$ et nous obtenons

$$\tilde{\pi}_k = E(\tilde{I}_k \mid \mathbf{\Omega}_P) = \delta_k + (1 - \delta_k)\pi_k.$$

L'estimateur qui en résulte s'écrit :

$$\hat{\theta} = \sum_{k \in s} \tilde{w}_k y_k = \sum_{k \in s_{\text{NP}}} y_k + \sum_{k \in s_P} \frac{1}{\pi_k} (1 - \delta_k) y_k. \tag{3.1}$$

Il est à noter que l'estimateur (3.1) nécessite que l'indicateur δ_k soit disponible pour toutes les unités de l'échantillon s_P. Pour les unités $k \in s_P \cap s_{\text{NP}}$, on dispose de deux valeurs : y_k et y_k^*. En principe, on devrait avoir $y_k^* = y_k$ mais il est possible que cette relation ne soit pas satisfaite exactement. Ces unités peuvent servir à valider l'hypothèse $y_k^* \approx y_k$. Si les différences sont trop importantes, il peut être préférable de ne pas considérer cette approche et de s'en remettre aux méthodes de la section 3.2 qui utilisent les données de la source non probabiliste comme données auxiliaires. Si on a

plutôt confiance dans la qualité des données de la source non probabiliste alors il peut être judicieux de ne pas collecter la variable y dans l'échantillon probabiliste pour les unités également présentes dans l'échantillon non probabiliste afin de réduire les coûts de collecte et le fardeau sur les répondants.

On peut voir le problème comme si on avait deux bases de sondages : U et s_{NP}. Un échantillon s_P est tiré aléatoirement de U et on recense toutes les unités de s_{NP}. On peut ainsi calculer la probabilité de sélection dans l'échantillon s pour chaque unité $k \in U$, $\Pr(k \in s \,|\, \mathbf{\Omega}_P)$, et on retrouve l'estimateur (3.1) en pondérant chaque unité $k \in s$ par l'inverse de cette probabilité. Cette approche a été proposée par Bankier (1986) pour traiter le problème des bases de sondage multiples. Dans le contexte de l'intégration d'un échantillon probabiliste et non probabiliste, l'estimateur (3.1) a été proposé par Kim et Tam (2020).

La dernière somme de (3.1) est un estimateur sans biais de $\sum_{k \in U} (1 - \delta_k) y_k = \sum_{k \in U - s_{\text{NP}}} y_k$. Si un vecteur de variables auxiliaires, $\mathbf{x}_k$, est disponible pour $k \in s_P$ de même que le total $\mathbf{T}_\mathbf{x} = \sum_{k \in U} \mathbf{x}_k$ alors on peut remplacer le poids $1 / \pi_k$ dans (3.1) par un poids calé w_k (ex. : Deville et Särndal, 1992; Haziza et Beaumont, 2017). Les poids calés minimisent une fonction de distance entre w_k et $1 / \pi_k$, $k \in s_P$, sous la contrainte de satisfaire l'équation de calage $\sum_{k \in s_P} w_k \mathbf{x}_k = \mathbf{T}_\mathbf{x}$. Idéalement, le calage est fait seulement sur la portion non couverte par l'échantillon non probabiliste, $U - s_{\text{NP}}$, c'est-à-dire que le vecteur de calage utilisé est $(1 - \delta_k)\, \mathbf{x}_k$ et l'équation de calage devient : $\sum_{k \in s_P} w_k (1 - \delta_k)\, \mathbf{x}_k = \sum_{k \in U - s_{\text{NP}}} \mathbf{x}_k$. Ce n'est pas possible quand $\sum_{k \in U - s_{\text{NP}}} \mathbf{x}_k$ n'est pas connu.

Remarque : Si l'hypothèse 2 n'est pas appropriée alors $E(I_k \,|\, \mathbf{\Omega}_P) \neq E(I_k \,|\, \mathbf{Z}) = \pi_k$. Pour contourner ce problème, on peut enlever de s_{NP} toutes les unités dont les données ont été recueillies après la sélection de l'échantillon s_P. L'hypothèse 2 est alors respectée mais on peut omettre ainsi beaucoup de données disponibles. Pour profiter de l'ensemble complet s_{NP}, il faudra se résoudre à faire quelques hypothèses et à s'éloigner partiellement de l'approche fondée sur le plan de sondage. En supposant que $E(I_k \,|\, \mathbf{\Omega}_P) = \Pr(I_k = 1 \,|\, \delta_k, \mathbf{Y}, \mathbf{\Omega})$, on peut montrer en utilisant le théorème de Bayes que

$$\Pr(I_k = 1 \,|\, \delta_k = 0, \mathbf{Y}, \mathbf{\Omega}) = \frac{1 - \Pr(\delta_k = 1 \,|\, I_k = 1, \mathbf{Y}, \mathbf{\Omega})}{1 - \Pr(\delta_k = 1 \,|\, \mathbf{Y}, \mathbf{\Omega})}\, \pi_k,$$

pour les unités $k \in U - s_{\text{NP}}$. L'estimation de $E(I_k \,|\, \mathbf{\Omega}_P)$ nécessite donc de postuler un modèle pour δ_k. Sous certaines hypothèses, l'estimation de $\Pr(\delta_k = 1 \,|\, I_k = 1, \mathbf{Y}, \mathbf{\Omega})$ peut se faire en utilisant les données de l'échantillon probabiliste et, par exemple, un modèle de régression logistique. L'estimation de $\Pr(\delta_k = 1 \,|\, \mathbf{Y}, \mathbf{\Omega})$ peut se faire en utilisant les méthodes décrites à la section 4.3 qui ne reposent pas sur la validité de l'hypothèse 2 telle que la méthode de Chen, Li et Wu (2019). Ces méthodes requièrent que les variables auxiliaires utilisées pour modéliser cette probabilité soient disponibles pour toutes les unités de l'échantillon combiné $s = s_P \cup s_{\text{NP}}$. Contrairement à la section 4.3, on peut ici profiter du fait que les valeurs y_k sont disponibles pour toutes les unités des deux échantillons et on peut utiliser la variable d'intérêt comme variable auxiliaire. On estime ensuite θ en remplaçant π_k dans (3.1) par une estimation de $\Pr(I_k = 1 \,|\, \delta_k = 0, \mathbf{Y}, \mathbf{\Omega})$. Des approches similaires ont été proposées par Beaumont, Bocci et Hidiroglou (2014) pour tenir compte des répondants tardifs à l'Enquête nationale sur les ménages de

Statistique Canada, c'est-à-dire les ménages qui ont répondu au questionnaire initial après que l'échantillon probabiliste de suivi des non-répondants ait été tiré.

3.2 Calage de l'échantillon probabiliste sur la source non probabiliste

Les données de sources non probabilistes telles que, par exemple, les données fournies par des répondants d'un panel Web peuvent être entachées d'erreurs de mesure suffisamment importantes pour mettre en doute l'hypothèse que $y_k^* \approx y_k$. De telles données ne peuvent donc pas être utilisées pour remplacer directement les valeurs de la variable y. On peut cependant les utiliser comme données auxiliaires pour enrichir l'enquête probabiliste au moyen de la technique du calage. La source non probabiliste contient les valeurs y_k^* pour $k \in s_{\text{NP}}$ et possiblement les valeurs d'autres variables. À partir de toutes ces variables, on peut former un vecteur de variables auxiliaires $\mathbf{x}_k^*$, disponibles pour $k \in s_{\text{NP}}$, qui pourrait contenir une ordonnée à l'origine. On note son total par $\mathbf{T}_{\mathbf{x}^*} = \sum_{k \in s_{\text{NP}}} \mathbf{x}_k^* = \sum_{k \in U} \delta_k \mathbf{x}_k^*$. On peut aussi disposer d'un autre vecteur de variables auxiliaires, $\mathbf{x}_k, k \in s_P$, de même que son total pour toute la population U, $\mathbf{T}_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k$. Les poids calés $w_k, k \in s_P$, sont obtenus en minimisant une fonction de distance entre w_k et $1/\pi_k, k \in s_P$, sous la contrainte de satisfaire l'équation de calage

$$\sum_{k \in s_P} w_k \begin{pmatrix} \mathbf{x}_k \\ \delta_k \mathbf{x}_k^* \end{pmatrix} = \begin{pmatrix} \mathbf{T}_{\mathbf{x}} \\ \mathbf{T}_{\mathbf{x}^*} \end{pmatrix}.$$

Il est à noter que ce calage ne peut être effectué que si $\mathbf{x}_k^*$ est disponible dans l'échantillon probabiliste pour toutes les unités $k \in s_P \cap s_{\text{NP}}$. L'estimateur de θ s'écrit encore comme $\hat{\theta} = \sum_{k \in s_P} w_k y_k$, où w_k est le poids calé satisfaisant l'équation de calage ci-dessus. Aucune hypothèse de modèle n'est requise pour la validité de l'approche et les estimations résultantes demeurent convergentes par rapport au plan peu importe la force de la relation entre y_k et les variables auxiliaires $\mathbf{x}_k$ et $\mathbf{x}_k^*$. Une relation forte permettra de réduire la variance d'échantillonnage de $\hat{\theta}$, $\text{var}\left(\hat{\theta} \mid \mathbf{\Omega}_P\right)$. Kim et Tam (2020) discutent de l'emploi d'un tel calage.

L'Enquête sur la population active (EPA) du Canada fournit un exemple d'application potentielle de cette méthode de calage. Le taux de chômage, défini comme étant le nombre de personnes sans emploi divisé par le nombre de personnes dans la population active, est un paramètre d'intérêt important que l'EPA permet d'estimer. Afin d'améliorer la précision des estimations de l'EPA, une variable de calage indiquant si une personne reçoit de l'assurance-emploi ou non pourrait être efficace puisqu'il existe assurément un lien entre recevoir de l'assurance-emploi et être sans emploi. Le total de cette variable de calage, le nombre de bénéficiaires d'assurance-emploi, est nécessaire pour effectuer ce calage et disponible à partir d'une source administrative. L'application de cette méthode nécessiterait toutefois l'addition d'une question à l'EPA pour identifier les répondants de l'EPA qui reçoivent de l'assurance-emploi. Cette information pourrait également être obtenue au moyen d'un appariement entre l'EPA et la source administrative. Il reste à déterminer si une telle variable de calage pourrait mener à des gains significatifs à l'EPA.

4 Approches fondées sur un modèle

Les approches fondées sur un modèle permettent d'éliminer le biais de sélection de la source non probabiliste et d'obtenir des inférences statistiques valides pourvu que leurs hypothèses sous-jacentes tiennent la route. L'objectif des méthodes des sections 4.1, 4.2 et 4.3 est de réduire le fardeau sur les répondants et les coûts en éliminant la collecte de quelques variables d'intérêt dans un échantillon probabiliste. Plus le nombre de variables d'intérêt pour lesquelles on ne recueille pas les valeurs sera grand, plus la réduction des coûts de collecte et du fardeau sur les répondants sera importante. Ces méthodes supposent cependant que les variables d'intérêt sont mesurées sans erreur dans l'échantillon non probabiliste $(y_k^* = y_k)$.

À partir de l'échantillon non probabiliste s_{NP}, on peut obtenir l'estimateur naïf $\hat{\theta}^{\text{NP}} = N\sum_{k\in s_{\text{NP}}} y_k / n^{\text{NP}}$ du total θ, où n^{NP} est le nombre d'unités dans s_{NP} et N est la taille de la population U. Il est bien connu que l'estimateur naïf peut être entaché d'un biais de sélection important (voir, par exemple, Bethlehem, 2016). L'objectif des méthodes des sections 4.1, 4.2 et 4.3 consiste à utiliser un vecteur de variables auxiliaires, $\mathbf{x}_k$, pour réduire le biais dont souffre l'estimateur naïf. On va noter par $\mathbf{X}$, la matrice qui contient les valeurs du vecteur $\mathbf{x}_k, k \in U$. On suppose que $\mathbf{x}_k$ est mesuré sans erreurs dans les deux échantillons s_{NP} et s_P.

À la section 4.4, on discute brièvement d'estimation sur petits domaines et du modèle de Fay et Herriot (1979). Les méthodes d'estimations sur petits domaines sont habituellement utilisées pour améliorer la précision d'estimations pour des sous-groupes (domaines) de la population dont la taille de l'échantillon probabiliste est petite. Elles nécessitent de recueillir la variable y dans l'échantillon probabiliste mais pas dans l'échantillon non probabiliste. Elles ne requièrent donc pas la condition $y_k^* = y_k$. Idéalement, l'échantillon non probabiliste contient des variables corrélées à y.

4.1 Calage de l'échantillon non probabiliste

L'approche la plus naturelle pour corriger le biais de sélection d'une source non probabiliste consiste à modéliser la relation entre la variable d'intérêt y_k et les variables auxiliaires $\mathbf{x}_k$ et ensuite à prédire le total θ en prédisant la variable y_k pour chacune des unités hors de l'échantillon non probabiliste. Cette approche par prédiction est décrite dans Royall (1970) et généralisée dans Royall (1976); voir aussi Elliott et Valliant (2017). On réfère le lecteur à Valliant, Dorfman et Royall (2000) pour plus de détails. Avec cette approche, les inférences sont conditionnelles à $\boldsymbol{\delta}$ et $\mathbf{X}$. Par conséquent, $\mathbf{Y}$ est considéré aléatoirede même que $\boldsymbol{\Omega}$ (sauf si $\boldsymbol{\Omega} = \mathbf{X}$). Si un échantillon probabiliste est utilisé, $\mathbf{I}$ est également considéré aléatoire. On fait habituellement l'hypothèse que le mécanisme de sélection des unités de l'échantillon s_{NP} n'est pas informatif :

Hypothèse 3 : $\mathbf{Y}$ et $\boldsymbol{\delta}$ sont indépendants après avoir conditionné sur $\mathbf{X}$.

L'hypothèse 3 est clé pour éliminer le biais de sélection. Plus on a accès à des variables auxiliaires fortement reliées à y_k et δ_k, plus l'hypothèse 3 devient plausible. Autrement dit, plus $\mathbf{X}$ est riche, plus

l'indépendance conditionnelle entre $\mathbf{Y}$ et $\boldsymbol{\delta}$ devient une hypothèse réaliste. On discute de cette hypothèse, appelée hypothèse d'*échangeabilité*, dans Mercer, Kreuter, Keeter et Stuart (2017). Schonlau et Couper (2017) discutent également du choix des variables auxiliaires et soulignent leur importance pour la réduction du biais de sélection.

Souvent, on considère un modèle linéaire pour lequel on suppose que les observations y_k sont mutuellement indépendantes avec $E(y_k \mid \mathbf{X}) = \mathbf{x}_k'\boldsymbol{\beta}$ et $\text{var}(y_k \mid \mathbf{X}) \propto v_k$, où $\boldsymbol{\beta}$ est un vecteur de paramètres inconnus du modèle et v_k est une fonction connue des variables du vecteur $\mathbf{x}_k$. Le meilleur prédicteur linéaire sans biais de θ (voir, par exemple, Valliant, Dorfman et Royall, 2000) est donné par

$$\hat{\theta}^{\text{BLUP}} = \sum_{k \in s_{\text{NP}}} y_k + \sum_{k \in U - s_{\text{NP}}} \mathbf{x}_k'\hat{\boldsymbol{\beta}} = \mathbf{T}_{\mathbf{x}}'\hat{\boldsymbol{\beta}} + \sum_{k \in s_{\text{NP}}} \left(y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}}\right), \tag{4.1}$$

où

$$\hat{\boldsymbol{\beta}} = \left(\sum_{k \in s_{\text{NP}}} v_k^{-1}\mathbf{x}_k\mathbf{x}_k'\right)^{-1} \sum_{k \in s_{\text{NP}}} v_k^{-1}\mathbf{x}_k y_k.$$

Le prédicteur $\hat{\theta}^{\text{BLUP}}$ peut aussi être ré-écrit sous la forme pondérée $\hat{\theta}^{\text{BLUP}} = \sum_{k \in s_{\text{NP}}} w_k^C y_k$, où

$$w_k^C = 1 + v_k^{-1}\mathbf{x}_k' \left(\sum_{k \in s_{\text{NP}}} v_k^{-1}\mathbf{x}_k\mathbf{x}_k'\right)^{-1} \left(\mathbf{T}_{\mathbf{x}} - \sum_{k \in s_{\text{NP}}} \mathbf{x}_k\right). \tag{4.2}$$

On peut facilement montrer que w_k^C est un poids calé qui satisfait l'équation de calage $\sum_{k \in s_{\text{NP}}} w_k^C \mathbf{x}_k = \mathbf{T}_{\mathbf{x}}$. Par conséquent, l'approche par prédiction est équivalente à faire un calage quand un modèle linéaire est utilisé pour décrire la relation entre y_k et $\mathbf{x}_k$. L'équation de calage permet de satisfaire ce que Mercer et coll. (2017) appellent l'hypothèse de *composition*. Cette approche nécessite de connaître le vecteur des totaux de contrôle $\mathbf{T}_{\mathbf{x}}$. S'il n'est pas connu, une alternative consiste à le remplacer dans (4.1) ou (4.2) par une estimation, $\hat{\mathbf{T}}_{\mathbf{x}} = \sum_{k \in s_P} w_k \mathbf{x}_k$, provenant d'une enquête probabiliste (Elliott et Valliant, 2017). Si les hypothèses 1 à 3 sont satisfaites, on peut montrer que le prédicteur $\hat{\theta}^{\text{BLUP}}$ est sans biais, c'est-à-dire que $E\left(\hat{\theta}^{\text{BLUP}} - \theta \,\middle|\, \boldsymbol{\delta}, \mathbf{X}\right) = 0$, que $\mathbf{T}_{\mathbf{x}}$ ou $\hat{\mathbf{T}}_{\mathbf{x}}$ soit utilisé pourvu que ce dernier soit sans biais par rapport au plan, c'est-à-dire que $E\left(\hat{\mathbf{T}}_{\mathbf{x}} \,\middle|\, \boldsymbol{\Omega}_P\right) = \mathbf{T}_{\mathbf{x}}$. La propriété d'être sans biais du prédicteur $\hat{\theta}^{\text{BLUP}}$ nécessite bien sûr que le modèle linéaire soit valide.

Remarque : En pratique, les variables auxiliaires pour lesquelles le total de population est connu seront généralement peu nombreuses et pas suffisamment prédictives de la variable y pour éliminer le biais de sélection. On les complètera avec d'autres variables auxiliaires dont le total peut être estimé au moyen d'une enquête probabiliste existante. Le vecteur des totaux de population sera donc un mélange de totaux connus et estimés. Si l'enquête probabiliste est elle-même calée sur des totaux connus de population alors on pourra utiliser uniquement les totaux estimés $\hat{\mathbf{T}}_{\mathbf{x}}$ provenant de l'enquête probabiliste.

Un modèle linéaire n'est pas toujours approprié. C'est le cas lorsque la variable y est catégorielle. Un autre exemple typique survient lorsqu'on est intéressé à estimer le total d'une variable quantitative dans un domaine d'intérêt. La variable y est alors définie comme le produit de cette variable quantitative et d'une variable binaire indiquant l'appartenance au domaine d'intérêt. Pour modéliser une telle variable, il est

naturel de considérer un mélange d'une distribution dégénérée à 0 et d'une distribution continue. Lorsque la relation entre y_k et $\mathbf{x}_k$ n'est pas linéaire, le calage assisté par un modèle de Wu et Sitter (2001) peut être utilisé pour préserver la forme pondérée du prédicteur de θ tout en tenant compte de la non-linéarité de la relation. Supposons qu'on remplace le modèle linéaire ci-dessus par un modèle non-linéaire (ou non paramétrique) tel que $E(y_k \mid \mathbf{X}) = h(\mathbf{x}_k)$, où $h(\cdot)$ est une certaine fonction. Le calage de Wu et Sitter (2001) consiste d'abord à prédire y_k par $\hat{y}_k = \hat{h}(\mathbf{x}_k)$, $k \in U$, où $\hat{h}(\mathbf{x}_k)$ est une estimation de $h(\mathbf{x}_k)$ obtenue au moyen du modèle. Ensuite, on calcule le total $T_{\hat{y}} = \sum_{k \in U} \hat{y}_k$ et on trouve des poids, w_k^{MC}, $k \in s_{\mathrm{NP}}$, qui satisfont l'équation de calage :

$$\sum_{k \in s_{\mathrm{NP}}} w_k^{\mathrm{MC}} \begin{pmatrix} 1 \\ \hat{y}_k \end{pmatrix} = \begin{pmatrix} N \\ T_{\hat{y}} \end{pmatrix}.$$

Autrement dit, on peut utiliser l'équation (4.2) en remplaçant $\mathbf{x}_k'$ par $(1, \hat{y}_k)$. Cette méthode requiert de connaître la taille de population N de même que les valeurs du vecteur $\mathbf{x}_k$ pour toutes les unités de la population U. Si N et $T_{\hat{y}}$ ne sont pas connus, on peut les remplacer par des estimations provenant d'une enquête probabiliste. Par exemple, on peut remplacer N par $\hat{N} = \sum_{k \in s_P} w_k$ et $T_{\hat{y}}$ par $\hat{T}_{\hat{y}} = \sum_{k \in s_P} w_k \hat{y}_k$. L'approche peut également se généraliser au cas où on a plusieurs variables d'intérêt.

Nous avons mentionné que le biais de sélection sera réduit considérablement si le vecteur $\mathbf{x}_k$ est riche et contient des variables qui sont à la fois reliées à δ_k et y_k, ce qui rend l'hypothèse 3 plus réaliste. Il peut donc être utile en pratique de considérer un grand nombre de variables auxiliaires potentielles et de choisir les plus importantes au moyen d'une technique de sélection de variables. Chen, Valliant et Elliott (2018) suggèrent la technique du LASSO pour sélectionner les variables auxiliaires et montrent ses bonnes propriétés.

Il est à noter que le prédicteur $\hat{\theta}^{\mathrm{BLUP}}$ se réduit à l'estimateur naïf, $\hat{\theta}^{\mathrm{NP}}$, dans le cas le plus simple possible où on ne considère qu'une seule variable auxiliaire constante : $x_k = 1, k \in U$. L'estimateur naïf est habituellement fortement biaisé. Son biais peut être réduit considérablement si on peut subdiviser la population U en H poststrates, $U_h, h = 1, \ldots, H$, disjointes et exhaustives, de taille N_h. On postule alors le modèle de poststratification, $E(y_k \mid \mathbf{X}) = \beta_h, k \in U_h$, qui est un cas particulier important du modèle linéaire ci-dessus. En supposant la variance $\mathrm{var}(y_k \mid \mathbf{X})$ constante pour $k \in U_h$, le prédicteur $\hat{\theta}^{\mathrm{BLUP}}$ s'écrit : $\hat{\theta}^{\mathrm{BLUP}} = \sum_{h=1}^{H} N_h \hat{\beta}_h$, où $\hat{\beta}_h = \sum_{k \in s_{\mathrm{NP},h}} y_k / n_h^{\mathrm{NP}}$, $s_{\mathrm{NP},h}$ est l'ensemble des unités de U_h qui font partie de l'échantillon s_{NP} et n_h^{NP} est la taille de $s_{\mathrm{NP},h}$. Si les tailles de population N_h sont inconnues, elles peuvent être remplacées par des estimations, $\hat{N}_h = \sum_{k \in s_{P,h}} w_k$, provenant d'une enquête probabiliste, où $s_{P,h}$ est l'ensemble des unités de U_h qui font partie de l'échantillon s_P. Les arbres de régression pourraient s'avérer une approche intéressante pour la formation de poststrates, particulièrement quand les variables auxiliaires sont catégorielles.

Si plusieurs variables auxiliaires catégorielles sont disponibles, il peut être utile de former un grand nombre de poststrates pour réduire le biais de sélection. Si un trop grand nombre de variables auxiliaires sont croisées, les tailles d'échantillon n_h^{NP} pourraient devenir très petites rendant ainsi les estimateurs $\hat{\beta}_h$

très instables. Gelman et Little (1997) suggèrent l'utilisation d'un modèle de régression à plusieurs niveaux pour obtenir des estimateurs $\tilde{\beta}_h$ plus stables que $\hat{\beta}_h$. Ils considèrent ensuite le prédicteur poststratifié : $\hat{\theta}^{\text{MRP}} = \sum_{h=1}^{H} N_h \tilde{\beta}_h$. La méthode est connue de nos jours sous le nom de Mr.P ou MRP (*Multilevel Regression and Poststratification*); voir, par exemple, Mercer et coll. (2017). Une approche similaire consisterait à utiliser des méthodes d'estimation sur petits domaines (Rao et Molina, 2015) pour stabiliser les estimateurs $\hat{\beta}_h$. Bien que de telles méthodes soient susceptibles de produire des estimations beaucoup plus précises de la moyenne de la variable y sur la population U_h, il reste à déterminer si de telles méthodes peuvent produire des gains d'efficacité significatifs pour l'estimation du total global θ par rapport au simple prédicteur poststratifié $\hat{\theta}^{\text{BLUP}} = \sum_{h=1}^{H} N_h \hat{\beta}_h$. Il semble que les arbres de régression fournissent une autre façon de contrôler l'instabilité des estimateurs $\hat{\beta}_h$ puisqu'un critère est habituellement utilisé pour éviter une subdivision trop fine de la population. Ces différentes méthodes méritent d'être investiguées plus en profondeur dans la recherche future. L'estimation précise des tailles de population N_h, si elles ne sont pas connues, est également un problème à ne pas négliger quand la population est divisée en un grand nombre de poststrates.

4.2 Appariement statistique

L'appariement statistique, ou la fusion de données, est une approche qui a été développée pour combiner les données de deux sources différentes qui contiennent des variables propres à chaque source mais aussi des variables communes. On réfère le lecteur à D'Orazio, Di Zio et Scanu (2006) ou Rässler (2012) pour une revue des méthodes d'appariement statistique. Dans le contexte de cet article, l'appariement statistique consiste à modéliser la relation entre y_k et les variables auxiliaires $\mathbf{x}_k$, communes aux deux sources, en utilisant les données de l'échantillon non probabiliste. Tout comme dans l'approche par calage, on doit faire l'hypothèse que le mécanisme de sélection de l'échantillon non probabiliste n'est pas informatif et choisir les variables auxiliaires judicieusement pour rendre l'hypothèse 3 la plus plausible possible. Une fois qu'un modèle a été déterminé, on l'utilise pour prédire les valeurs de y dans un échantillon probabiliste. On peut voir l'appariement statistique comme un problème d'imputation avec un taux d'imputation de 100 %. Le prédicteur de θ, obtenu à partir de l'échantillon probabiliste, prend la forme : $\hat{\theta}^{\text{SM}} = \sum_{k \in s_P} w_k y_k^{\text{imp}}$, où y_k^{imp} est la valeur imputée pour l'unité $k \in s_P$. Tout comme pour le calage, les inférences sont conditionnelles à $\boldsymbol{\delta}$ et $\mathbf{X}$. L'hypothèse 3, dans un contexte d'appariement statistique, peut être vue comme l'analogue de l'hypothèse *Population Missing At Random* (PMAR) introduite par Berg, Kim et Skinner (2016) dans un contexte de non-réponse.

Si le modèle de régression linéaire $E(y_k \mid \mathbf{X}) = \mathbf{x}_k'\boldsymbol{\beta}$ est utilisé alors la valeur imputée pour l'unité $k \in s_P$ est $y_k^{\text{imp}} = \mathbf{x}_k'\hat{\boldsymbol{\beta}}$ et le prédicteur résultant est donné par $\hat{\theta}^{\text{SM}} = \hat{\mathbf{T}}_{\mathbf{x}}'\hat{\boldsymbol{\beta}}$. Si les hypothèses 1 à 3 sont satisfaites et $E\left(\hat{\mathbf{T}}_{\mathbf{x}} \mid \boldsymbol{\Omega}_P\right) = \mathbf{T}_{\mathbf{x}}$, l'appariement statistique produit un prédicteur, $\hat{\theta}^{\text{SM}}$, sans biais, c'est-à-dire que $E\left(\hat{\theta}^{\text{SM}} - \theta \mid \boldsymbol{\delta}, \mathbf{X}\right) = 0$. De plus, si $v_k = \mathbf{x}_k'\boldsymbol{\lambda}$, pour un certain vecteur connu $\boldsymbol{\lambda}$, on peut montrer que $\sum_{k \in s_{\text{NP}}} \left(y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}}\right) = 0$ et le prédicteur $\hat{\theta}^{\text{SM}}$ est équivalent au prédicteur $\hat{\theta}^{\text{BLUP}}$ si on

remplace $\mathbf{T}_\mathbf{x}$ dans (4.1) par $\hat{\mathbf{T}}_\mathbf{x}$. On peut aussi montrer que pour un modèle de poststratification où on impute $y_k, k \in s_{P,h}$, par $y_k^{\text{imp}} = \hat{\beta}_h$, le prédicteur $\hat{\theta}^{\text{SM}}$ se réduit à $\hat{\theta}^{\text{SM}} = \sum_{h=1}^{H} \hat{N}_h \hat{\beta}_h$. L'appariement statistique et le calage produisent donc des prédicteurs semblables, même identiques dans certains cas, lorsqu'un modèle linéaire est postulé et que les totaux $\mathbf{T}_\mathbf{x}$ sont estimés.

Le choix entre l'appariement statistique ou le calage peut dépendre du point de vue de l'utilisateur. Par exemple, si c'est le contenu de la source non probabiliste, en termes de variables d'intérêt, qui est pertinent pour l'utilisateur, alors il semble naturel de pondérer l'échantillon non probabiliste pour ainsi espérer réduire le biais de sélection pour toutes les variables d'intérêt. La technique du calage est un choix évident pour obtenir une telle pondération tout comme les méthodes de la section 4.3. À l'opposé, si c'est plutôt le contenu de l'enquête probabiliste qui est pertinent, alors l'appariement statistique est le choix approprié. Cette méthode permet d'enrichir l'enquête probabiliste en imputant les variables d'intérêt manquantes.

L'appariement statistique se généralise facilement au cas de modèles non linéaires ou non paramétriques tels que $E(y_k \mid \mathbf{X}) = h(\mathbf{x}_k)$. Les valeurs imputées y_k^{imp} sont simplement obtenues en prédisant les valeurs manquantes $y_k, k \in s_P$, au moyen du modèle choisi. Le prédicteur $\hat{\theta}^{\text{SM}} = \sum_{k \in s_P} w_k y_k^{\text{imp}}$ reste sans biais si les hypothèses 1 à 3 sont satisfaites et si $E\left(y_k^{\text{imp}} - y_k \mid \boldsymbol{\delta}, \mathbf{X}\right) = 0$. L'imputation par donneur ou par le plus proche voisin est une méthode d'imputation non paramétrique fréquemment utilisée pour traiter la non-réponse (voir, par exemple, Beaumont et Bocci, 2009) qui ne requiert pas une relation linéaire entre y_k et $\mathbf{x}_k$. Dans le contexte de l'appariement d'un échantillon non probabiliste à un échantillon probabiliste, l'imputation par donneur a été popularisée par Rivers (2007). Pour une unité donnée $k \in s_P$, la méthode consiste à trouver le donneur le plus proche, selon les variables auxiliaires $\mathbf{x}$, parmi les unités de l'échantillon non probabiliste et à remplacer la valeur manquante y_k par la valeur de la variable y de ce donneur. Pour l'imputation par donneur, la condition $E\left(y_k^{\text{imp}} - y_k \mid \boldsymbol{\delta}, \mathbf{X}\right) = 0$ est satisfaite si, pour chaque receveur $k \in s_P$, le donneur a exactement les même valeurs du vecteur $\mathbf{x}$ que le receveur. Lorsqu'une ou plusieurs variables auxiliaires sont continues, cette condition n'est satisfaite qu'asymptotiquement en général. Un échantillon non probabiliste de très grande taille fournit un grand bassin de donneurs, ce qui devrait aider à satisfaire approximativement cette condition.

Remarque : Dans certaines applications, on peut avoir accès à un très grand panel non probabiliste de volontaires, s_{NP}, qui contient quelques variables auxiliaires pour l'appariement, $\mathbf{x}$, mais aucune variable d'intérêt. Idéalement, les variables d'intérêt seraient recueillies pour toutes les unités du panel s_{NP} mais ce n'est pas possible en raison du coût et du fardeau sur les membres du panel. En pratique, on choisira donc un sous-échantillon s_{NP}^* de s_{NP} en utilisant des méthodes d'échantillonnage aléatoires ou non. L'échantillonnage par quotas (ex. : Deville, 1991) est souvent considéré dans ce contexte. En plus de recueillir les variables d'intérêt pour toutes les unités de s_{NP}^*, on pourrait aussi vouloir recueillir d'autres variables auxiliaires pour l'appariement afin d'enrichir le vecteur $\mathbf{x}$. L'appariement peut ensuite être effectué à l'échantillon probabiliste, souvent de beaucoup plus petite taille, pourvu que ce dernier

contienne les mêmes variables auxiliaires que celles du sous-échantillon non probabiliste s_{NP}^{*}. En choisissant judicieusement les variables auxiliaires pour l'appariement, le potentiel de réduction du biais est accru (Schonlau et Cooper, 2017). L'implémentation proposée par Rivers (2007) est légèrement différente. Rivers (2007) suggère d'effectuer l'appariement entre l'échantillon probabiliste et le panel s_{NP} en utilisant les variables auxiliaires disponibles dans les deux sources. La collecte des variables d'intérêt se fait uniquement sur l'ensemble des donneurs dans s_{NP} qui ont été appariés à une unité de l'échantillon probabiliste ce qui permet de réduire considérablement les coûts de collecte et le fardeau. L'hypothèse implicite est que les membres du panel, étant initialement des volontaires, sont plus susceptibles de répondre que des individus choisis au hasard dans la population. Évidemment, la non-réponse est inévitable et ce problème doit être traité, possiblement par imputation. L'avantage de cette méthode est que l'appariement se fait à partir du panel s_{NP} plutôt que d'un sous-échantillon de ce panel. Le bassin de donneurs est donc plus grand. En contrepartie, l'appariement ne peut pas être effectué en utilisant le vecteur enrichi de variables auxiliaires car il n'est pas disponible pour les unités de s_{NP}, ce qui limite le potentiel de réduction du biais.

Lavallée et Brisbane (2016) notent le lien entre l'appariement statistique et l'échantillonnage indirect (Lavallée, 2007; Deville et Lavallée, 2006). Ils proposent un estimateur qui est obtenu en imputant chaque valeur manquante $y_k, k \in s_P$, par une moyenne pondérée des valeurs y de donneurs proches. En réalité, leur estimateur peut également être obtenu de façon équivalente en imputant les valeurs manquantes au moyen de la méthode d'imputation fractionnelle par donneur (par exemple, Kim et Fuller, 2004). L'utilisation de plus d'un donneur pour imputer les valeurs manquantes permet une réduction de variance quoique typiquement modeste.

Plusieurs des méthodes d'imputation utilisées en pratique peuvent être considérées comme étant linéaires (Beaumont et Bissonnette, 2011). C'est le cas de l'imputation par la régression linéaire, de l'imputation par donneur et de l'imputation fractionnelle par donneur. Une méthode d'imputation est dite linéaire si la valeur imputée $y_k^{\text{imp}}, k \in s_P$, peut être écrite sous le forme $y_k^{\text{imp}} = \sum_{l \in s_{\text{NP}}} \omega_{kl} y_l$, où ω_{kl} est une fonction de $\boldsymbol{\delta}$ ou $\mathbf{X}$ mais pas de $\mathbf{Y}$. Par exemple, pour l'imputation par donneur ou plus proche voisin, $\omega_{kl} = 1$ si l'unité $l \in s_{\text{NP}}$ est donneuse pour l'unité receveuse $k \in s_P$; sinon, $\omega_{kl} = 0$. Pour une méthode d'imputation linéaire, on peut réécrire l'estimateur $\hat{\theta}^{\text{SM}} = \sum_{k \in s_P} w_k y_k^{\text{imp}}$ comme une somme pondérée sur l'échantillon non probabiliste : $\hat{\theta}^{\text{SM}} = \sum_{l \in s_{\text{NP}}} W_l y_l$, où $W_l = \sum_{k \in s_P} w_k \omega_{kl}$. Pour les méthodes d'imputation linéaire, l'appariement statistique est donc une alternative au calage et aux méthodes de la section 4.3 si l'objectif est de pondérer adéquatement l'échantillon non probabiliste.

Jusqu'à maintenant, nous n'avons considéré que l'estimation du total $\theta = \sum_{k \in U} y_k$. L'échantillon probabiliste contient cependant d'autres variables et on pourrait être intéressé à la relation entre deux ou plusieurs variables, certaines provenant de l'enquête probabiliste et d'autres étant imputées à partir de l'échantillon non probabiliste. À des fins d'exemple, supposons qu'on veuille estimer le total $\theta = \sum_{k \in U} \tilde{y}_k y_k$, où $\tilde{y}_k$ est une variable qui est recueillie par l'enquête probabiliste mais non disponible dans l'échantillon non probabiliste. Elle pourrait, par exemple, définir l'appartenance à un domaine

d'intérêt. On peut utiliser l'appariement statistique pour estimer ce paramètre par $\hat{\theta}^{\text{SM}} = \sum_{k \in s_P} w_k \tilde{y}_k y_k^{\text{imp}}$. Notons par $\tilde{\mathbf{Y}}$, le vecteur qui contient les valeurs de la variable $\tilde{y}_k, k \in U$. On peut montrer que $\hat{\theta}^{\text{SM}}$ est sans biais, $E\left(\hat{\theta}^{\text{SM}} - \theta \mid \boldsymbol{\delta}, \mathbf{X}, \tilde{\mathbf{Y}}\right) = 0$, si les hypothèses 1 à 3 sont satisfaites en plus de l'hypothèse suivante :

Hypothèse 4 : $\mathbf{Y}$ et $\tilde{\mathbf{Y}}$ sont indépendants après avoir conditionné sur $\boldsymbol{\delta}$ et $\mathbf{X}$.

L'hypothèse 4 est connue comme étant l'hypothèse d'indépendance conditionnelle dans la littérature sur l'appariement statistique.

4.3 Pondération par l'inverse du score de propension

Au lieu de modéliser la relation entre y_k et $\mathbf{x}_k$, on peut modéliser la relation entre δ_k et $\mathbf{x}_k$. L'avantage principal de cette approche est de simplifier l'effort de modélisation quand il y a plusieurs variables d'intérêt puisqu'il n'y a toujours qu'une seule variable δ_k. Avec cette approche, les inférences sont conditionnelles à $\mathbf{Y}$ et $\mathbf{X}$ et on suppose habituellement que l'hypothèse 3 est valide de telle sorte que $\Pr(\delta_k = 1 \mid \mathbf{Y}, \mathbf{X}) = \Pr(\delta_k = 1 \mid \mathbf{X})$. On estime ensuite la probabilité de participation $p_k = \Pr(\delta_k = 1 \mid \mathbf{X})$ par $\hat{p}_k$ et on calcule l'estimation $\hat{\theta}^{\text{PS}} = \sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} y_k$, où $w_k^{\text{PS}} = 1 / \hat{p}_k$. On doit faire l'hypothèse que $p_k > 0, k \in U$. Elle est appelée l'hypothèse de *positivité* par Mercer et coll. (2017). Elle peut également être requise dans les approches par calage et par appariement statistique. Par exemple, on pourrait observer des poststrates vides $\left(n_h^{\text{NP}} = 0\right)$ si elle n'est pas satisfaite. Pour corriger ce problème, on regroupe généralement ces poststrates vides avec d'autres poststrates non vides. Un tel regroupement peut compromettre la validité de l'hypothèse 3 si les poststrates regroupées sont différentes.

L'estimation de p_k peut être effectuée en postulant un modèle paramétrique $p_k = g(\mathbf{x}_k; \boldsymbol{\alpha})$, où g est une certaine fonction, normalement bornée par 0 et 1, et $\boldsymbol{\alpha}$ est un vecteur de paramètres inconnus du modèle. La fonction logistique $g(\mathbf{x}_k; \boldsymbol{\alpha}) = \exp\left(\mathbf{x}_k'\boldsymbol{\alpha}\right) \Big/ \left[1 + \exp\left(\mathbf{x}_k'\boldsymbol{\alpha}\right)\right]$ domine dans les applications (voir Kott, 2019, pour une application récente). On note l'estimateur de $\boldsymbol{\alpha}$ par $\hat{\boldsymbol{\alpha}}$ et la probabilité estimée par $\hat{p}_k = g(\mathbf{x}_k; \hat{\boldsymbol{\alpha}})$. Idéalement, on estimerait $\boldsymbol{\alpha}$ en utilisant $\mathbf{x}_k$ pour toutes les unités de la population U comme on le ferait dans un contexte de non-réponse. Par exemple, en supposant l'utilisation de la fonction logistique, on pourrait estimer $\boldsymbol{\alpha}$ en résolvant l'équation du maximum de vraisemblance

$$\sum_{k \in U} [\delta_k - p_k(\boldsymbol{\alpha})] \mathbf{x}_k = \sum_{k \in s_{\text{NP}}} \mathbf{x}_k - \sum_{k \in U} p_k(\boldsymbol{\alpha}) \mathbf{x}_k = \mathbf{0}. \tag{4.3}$$

Ce n'est pas possible quand $\mathbf{x}_k$ n'est pas connu pour toutes les unités $k \in U - s_{\text{NP}}$, ce qui est presque toujours le cas en pratique. Iannacchione, Milne et Folsom (1991) ont proposé une autre équation d'estimation sans biais pour $\boldsymbol{\alpha}$ (voir aussi Deville et Dupont, 1993) :

$$\sum_{k \in s_{\text{NP}}} \frac{\mathbf{x}_k}{p_k(\boldsymbol{\alpha})} - \sum_{k \in U} \mathbf{x}_k = \mathbf{0}. \tag{4.4}$$

L'avantage principal de l'équation (4.4) est qu'elle ne requiert pas de connaître $\mathbf{x}_k$ pour chaque unité $k \in U - s_{\text{NP}}$. Il faut toutefois avoir accès au vecteur de totaux $\sum_{k \in U} \mathbf{x}_k$ à partir d'une source externe. Une propriété intéressante de l'équation (4.4) est que les poids résultants $w_k^{\text{PS}} = 1 / p_k(\hat{\boldsymbol{\alpha}})$ satisfont l'équation de calage $\sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k$ tout comme les poids w_k^C de l'équation (4.2). On peut d'ailleurs montrer que la solution de (4.4) donne $w_k^{\text{PS}} = w_k^C$ si on utilise le modèle $p_k(\boldsymbol{\alpha}) = \left(1 + v_k^{-1} \mathbf{x}_k' \boldsymbol{\alpha}\right)^{-1}$. C'est toutefois un modèle moins naturel que le modèle logistique ci-dessus pour la modélisation d'une probabilité.

Pour contourner le problème des valeurs manquantes $\mathbf{x}_k, k \in U - s_{\text{NP}}$, Chen et coll. (2019) suggèrent d'estimer $\sum_{k \in U} p_k(\boldsymbol{\alpha}) \mathbf{x}_k$ dans (4.3) au moyen d'une enquête probabiliste. L'équation à résoudre devient :

$$\sum_{k \in s_{\text{NP}}} \mathbf{x}_k - \sum_{k \in s_P} w_k p_k(\boldsymbol{\alpha}) \mathbf{x}_k = \mathbf{0}. \tag{4.5}$$

L'équation (4.5) est sans biais conditionnellement à $\mathbf{Y}$ et $\mathbf{X}$ pourvu que l'enquête probabiliste permette d'estimer sans biais, conditionnellement à $\mathbf{Y}$ et $\boldsymbol{\Omega}$, n'importe quel total de population qui n'est pas une fonction de $\boldsymbol{\delta}$ tel que $\sum_{k \in U} p_k(\boldsymbol{\alpha}) \mathbf{x}_k$. Les hypothèses 1 et 3 sont requises mais pas l'hypothèse 2. En utilisant l'idée de Iannacchione et coll. (1991), une alternative à (4.5) consiste à résoudre :

$$\sum_{k \in s_{\text{NP}}} \frac{\mathbf{x}_k}{p_k(\boldsymbol{\alpha})} - \sum_{k \in s_P} w_k \mathbf{x}_k = \mathbf{0}. \tag{4.6}$$

L'équation (4.6) produit des poids $w_k^{\text{PS}} = 1 / p_k(\hat{\boldsymbol{\alpha}})$ qui satisfont l'équation de calage $\sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} \mathbf{x}_k = \sum_{k \in s_P} w_k \mathbf{x}_k$ (voir également Lesage, 2017; et Rao, 2020). Les estimateurs de $\boldsymbol{\alpha}$ obtenus au moyen de (4.5) ou (4.6) sont vraisemblablement moins efficaces que ceux obtenus au moyen de (4.3) ou (4.4). Si on connaît $\mathbf{x}_k, k \in U - s_{\text{NP}}$, ou le vecteur $\sum_{k \in U} \mathbf{x}_k$ alors on privilégiera l'utilisation de (4.3) ou (4.4). Autrement, on se tournera vers (4.5) ou (4.6). Ces dernières requièrent toutefois que $\mathbf{x}_k$ soit recueilli dans une enquête probabiliste. Il est à noter que les indicateurs δ_k n'ont pas besoin d'être observés dans l'échantillon probabiliste.

Les équations (4.5) et (4.6) peuvent être plus difficiles à résoudre que les équations (4.3) et (4.4) et pourraient ne pas avoir de solutions. Prenons par exemple le cas où on a une seule variable auxiliaire : $x_k = 1$. En utilisant (4.5) ou (4.6), on observe que la probabilité estimée doit être : $\hat{p}_k = n^{\text{NP}} / \sum_{k \in s_P} w_k$. Si la taille de l'échantillon probabiliste est assez grande, on s'attend à ce que $0 < \hat{p}_k < 1$. Pour de petites tailles d'échantillon, il pourrait arriver que $\hat{p}_k > 1$ dû à la variabilité de $\sum_{k \in s_P} w_k$. Dans ce cas, les équations (4.5) et (4.6) n'auraient pas de solution si la fonction logistique est utilisée puisqu'elle exige que $0 < \hat{p}_k < 1$. Pour éviter ce problème, il pourrait être utile de considérer d'autres fonctions non bornées par 1 telle que $g(\mathbf{x}_k; \boldsymbol{\alpha}) = \exp\left(\mathbf{x}_k' \boldsymbol{\alpha}\right)$.

Kim et Wang (2019) proposent d'utiliser l'échantillon probabiliste pour estimer la probabilité de participation. En supposant la fonction logistique, l'équation à résoudre est :

$$\sum_{k \in s_P} w_k \left[\delta_k \; - \; p_k \left(\boldsymbol{\alpha}\right)\right] \mathbf{x}_k \; = \; \sum_{k \in s_P} w_k \delta_k \mathbf{x}_k \; - \; \sum_{k \in s_P} w_k p_k \left(\boldsymbol{\alpha}\right) \mathbf{x}_k \; = \; \mathbf{0}.$$

La méthode requiert de connaître les indicateurs δ_k dans l'échantillon probabiliste et la validité des hypothèses 1, 2 et 3 pour que l'équation d'estimation soit sans biais. De plus, l'échantillon probabiliste est habituellement de petite taille en comparaison à l'échantillon non probabiliste et il peut être numériquement difficile d'estimer $\boldsymbol{\alpha}$, particulièrement quand le vecteur $\mathbf{x}_k$ contient beaucoup de variables et que le chevauchement entre les deux échantillons est petit.

Lee (2006), voir aussi Rivers (2007), Valliant et Dever (2011) et Elliott et Valliant (2017), propose de combiner les deux échantillons et d'estimer ensuite p_k au moyen d'une régression logistique. Il semble qu'on y fait l'hypothèse implicite que les deux échantillons ne se chevauchent pas, c'est-à-dire que $\delta_k = 0$ pour toutes les unités dans s_P. En utilisant encore la fonction logistique, l'équation d'estimation qui en résulte est :

$$\sum_{k \in s_{\mathrm{NP}}} \eta_k^{\mathrm{NP}} \left[1 \; - \; p_k \left(\boldsymbol{\alpha}\right)\right] \mathbf{x}_k \; - \; \sum_{k \in s_P} w_k p_k \left(\boldsymbol{\alpha}\right) \mathbf{x}_k \; = \; \mathbf{0}, \tag{4.7}$$

où η_k^{NP} est un certain poids pour les unités de l'échantillon non probabiliste. La méthode est un peu similaire à celle de Chen et coll. (2019) mais l'équation d'estimation (4.7) n'est pas sans biais, conditionnellement à $\mathbf{Y}$ et $\mathbf{X}$, contrairement aux équations (4.5) et (4.6). Cependant, si on pose $\eta_k^{\mathrm{NP}} = 1$ et si $\max \{p_k; \; k \in U\}$ est petit alors l'équation (4.7) devient approximativement équivalente à l'équation (4.5). Lee (2006) n'utilise toutefois pas directement les probabilités estimées découlant de (4.7). L'auteur les utilise seulement pour ordonner l'union des deux échantillons et ensuite créer des classes homogènes. L'utilisation de classes homogènes apporte une certaine robustesse par rapport à une mauvaise spécification du modèle et peut aider à éviter des probabilités estimées très petites et ainsi des poids très grands. Dans le contexte de la non-réponse, la formation de classes homogènes d'imputation ou de repondération a été étudiée, entre autres, par Little (1986), Eltinge et Yansaneh (1997) et Haziza et Beaumont (2007). Haziza et Lesage (2016) illustrent la robustesse de la méthode quand la fonction $g\left(\mathbf{x}_k; \; \boldsymbol{\alpha}\right)$ est incorrectement spécifiée. La méthode est utilisée régulièrement dans les enquêtes de Statistique Canada pour le traitement de la non-réponse.

Plutôt que d'utiliser (4.7), la formation de classes homogènes pourrait être effectuée en partant des équations non biaisées (4.5) ou (4.6). Notons ces probabilités estimées initiales par $\hat{p}_k^0 = g\left(\mathbf{x}_k; \; \hat{\boldsymbol{\alpha}}\right)$. L'échantillon $s = s_P \cup s_{\mathrm{NP}}$ est ensuite ordonné selon $\hat{p}_k^0$ et réparti en C classes homogènes de tailles égales ou non. Notons par $s_{P,c}$, l'ensemble des unités de s_P qui font partie de la classe c. L'ensemble des unités de s_{NP} qui font partie de la classe c est noté par $s_{\mathrm{NP},c}$. Le poids w_k^{PS} pour une unité $k \in s_{\mathrm{NP},c}$ est égal à l'inverse du taux de participation estimé dans la classe c et est donné par $w_k^{\mathrm{PS}} = \hat{N}_c \; / \; n_c^{\mathrm{NP}}$, où $\hat{N}_c = \sum_{k \in s_{P,c}} w_k$ et n_c^{NP} est le nombre d'unités dans $s_{\mathrm{NP},c}$. Ce poids assure la propriété de calage : $\sum_{k \in s_{\mathrm{NP},c}} w_k^{\mathrm{PS}} = \hat{N}_c$. Le nombre de classes doit être suffisamment grand pour capturer un fort pourcentage de la variabilité des probabilités initiales $\hat{p}_k^0$ et ainsi permettre de réduire le biais. En contrepartie, il ne doit pas être trop grand pour éviter que certaines classes deviennent vides car les poids $w_k^{\mathrm{PS}} = \hat{N}_c \; / \; n_c^{\mathrm{NP}}$ ne peuvent pas être calculés si $n_c^{\mathrm{NP}} = 0$. Les arbres de régression peuvent s'avérer

une alternative efficace pour la formation de classes. Dans un contexte de non-réponse, ils ont été étudié par Phipps et Toth (2012). L'estimateur $\hat{\theta}^{\text{PS}} = \sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} y_k$ obtenu en faisant des classes homogènes a exactement la même forme que l'estimateur poststratifié décrit dans l'approche par calage à la section 4.1, la seule différence étant que les classes sont construites en modélisant δ_k plutôt que y_k.

L'hypothèse 3 peut ne pas être réaliste dans certains contextes de telle sorte que $\Pr(\delta_k = 1 \mid \mathbf{Y}, \mathbf{X}) \neq \Pr(\delta_k = 1 \mid \mathbf{X})$. Dans ce cas, on pourrait vouloir modéliser la probabilité de participation $p_k = \Pr(\delta_k = 1 \mid \mathbf{Y}, \mathbf{X})$ au moyen d'un vecteur de variables explicatives $\mathbf{x}_k^*$, définies en utilisant la variable d'intérêt y_k (ou les variables d'intérêt s'il y en a plusieurs) et possiblement d'autres variables auxiliaires $\mathbf{x}_k$. On peut considérer un modèle paramétrique, $p_k = g(\mathbf{x}_k^*; \boldsymbol{\alpha})$, pour modéliser la probabilité de participation. Les équations (4.5) et (4.6) ne peuvent pas être utilisées pour estimer $\boldsymbol{\alpha}$ parce que y_k (et par conséquent $\mathbf{x}_k^*$) n'est pas disponible dans l'échantillon probabiliste. On peut toutefois utiliser une équation similaire à (4.6) :

$$\sum_{k \in s_{\text{NP}}} \frac{\mathbf{x}_k^I}{g(\mathbf{x}_k^*; \boldsymbol{\alpha})} - \sum_{k \in s_P} w_k \mathbf{x}_k^I = \mathbf{0}. \tag{4.8}$$

Le vecteur $\mathbf{x}_k^I$, de même dimension que $\boldsymbol{\alpha}$, contient des variables de calage, aussi appelées variables instrumentales dans la littérature économétrique. On va noter par $\mathbf{X}^I$, la matrice qui contient les valeurs du vecteur $\mathbf{x}_k^I$, $k \in U$. L'équation (4.8) requiert de connaître les variables de calage $\mathbf{x}_k^I$ pour les deux échantillons. Par contre, les variables explicatives $\mathbf{x}_k^*$ peuvent n'être observées que pour les unités de l'échantillon non probabiliste. L'équation (4.8) produit des poids $w_k^{\text{PS}} = 1 / g(\mathbf{x}_k^*; \hat{\boldsymbol{\alpha}})$ qui satisfont l'équation de calage $\sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} \mathbf{x}_k^I = \sum_{k \in s_P} w_k \mathbf{x}_k^I$. Une équation similaire à (4.8) a été initialement proposée par Deville (1998) pour traiter la non-réponse (voir aussi Kott, 2006; et Haziza et Beaumont, 2017). L'équation (4.8) est sans biais, conditionnellement à $\mathbf{Y}$, $\mathbf{X}$ et $\mathbf{X}^I$, si les variables instrumentales $\mathbf{x}_k^I$ peuvent être choisies de telle sorte que l'hypothèse suivante est satisfaite :

Hypothèse 5 : $\boldsymbol{\delta}$ et $\mathbf{X}^I$ sont indépendants après avoir conditionné sur $\mathbf{Y}$ et $\mathbf{X}$.

L'hypothèse 3 n'est plus requise mais elle est remplacée par une autre hypothèse. Le choix de variables instrumentales $\mathbf{x}_k^I$ qui satisfont l'hypothèse 5 n'est pas toujours évident en pratique. Elles ne doivent pas être prédictives de δ_k après avoir conditionné sur $\mathbf{x}_k^*$. Idéalement, pour des raisons d'efficacité, les variables instrumentales sont choisies de telle sorte qu'elles soient prédictives de $\mathbf{x}_k^*$ tout en ne compromettant pas l'hypothèse 5. Contrairement aux équations (4.5) et (4.6), l'équation (4.8) ne permet pas la formation de classes homogènes parce que les probabilités de participation $\hat{p}_k = g(\mathbf{x}_k^*; \hat{\boldsymbol{\alpha}})$ ne peuvent pas être calculées pour les unités de l'échantillon probabiliste. On perd ainsi la propriété de robustesse qui vient avec l'utilisation de classes homogènes. En raison de ces inconvénients, l'équation (4.8) ne devrait être considérée que lorsqu'on a de fortes raisons de croire que l'hypothèse 3 n'est pas appropriée.

Une fois que des poids w_k^{PS} ont été calculés selon une des méthodes de cette section, il est toujours possible de les rajuster en faisant un calage. L'objectif de ce calage est d'améliorer la précision de l'estimateur $\hat{\theta}^{\text{PS}}$ et aussi d'obtenir une propriété de double robustesse (voir Chen et coll., 2019).

En général, on va observer la variable y pour tout l'échantillon non probabiliste et on peut utiliser l'estimateur pondéré par l'inverse du score de propension, $\hat{\theta}^{\text{PS}} = \sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} y_k$, ou un estimateur pondéré obtenu par calage ou par appariement statistique. Il peut arriver que l'échantillon non probabiliste soit trop volumineux et que la variable y ne puisse être recueillie que pour un sous-échantillon de s_{NP}. L'échantillonnage par quotas (ex. : Deville, 1991) est une méthode fréquemment utilisée pour tirer le sous-échantillon si des variables auxiliaires sont disponibles pour $k \in s_{\text{NP}}$. Une alternative à l'échantillonnage par quotas est de calculer les poids w_k^{PS} pour tout l'échantillon non probabiliste et ensuite de les utiliser pour sélectionner un sous-échantillon aléatoire avec probabilités proportionnelles aux poids. La variable y ne sera recueillie que pour le sous-échantillon et les estimations pourront être obtenues comme si le sous-échantillon avait été tiré de la population selon un plan à probabilités égales. Cette approche s'appelle l'échantillonnage inverse dans la littérature sur les sondages probabilistes (voir, par exemple, Hinkins, Oh et Scheuren, 1997; ou Rao, Scott et Benhin, 2003) et a été proposée par Kim et Wang (2019) dans le cas des échantillons non probabilistes.

4.4 Estimation sur petits domaines

Dans la plupart des enquêtes, on s'intéresse non seulement à estimer le total de la variable y pour toute la population U mais également pour différents sous-groupes de la population appelés domaines. Les enquêtes probabilistes menées par les agences nationales de statistique produisent généralement des estimations fiables pour des domaines qui contiennent suffisamment d'unités échantillonnées. Leur biais est contrôlé par les différentes procédures d'échantillonnage et de collecte et leur variance est typiquement assez petite pour être en mesure de tirer des conclusions justes. Lorsque le domaine d'intérêt contient peu d'unités échantillonnées, les estimations de l'enquête peuvent devenir si instables qu'elles en deviennent inutilisables même quand leur biais demeure contrôlé. Pour pallier à un manque de données dans un domaine d'intérêt, on peut considérer l'utilisation de méthodes d'estimation sur petits domaines. Ces méthodes compensent le manque de données observées dans un domaine par des hypothèses de modèle qui relient des données auxiliaires aux données de l'enquête. Deux types de modèle sont fréquemment utilisés : les modèles au niveau des unités et les modèles au niveau des domaines. Le modèle au niveau des domaines de Fay et Herriot (1979) est sans contredit le plus populaire en pratique. Il requiert la disponibilité de données auxiliaires uniquement au niveau des domaines, contrairement aux modèles au niveau des unités qui nécessitent d'observer les variables auxiliaires pour chaque unité de la population U. On réfère le lecteur à Rao et Molina (2015) pour une excellente couverture des différentes approches. Dans ce qui suit, on se concentre sur le modèle de Fay-Herriot.

Supposons qu'on veuille estimer D totaux, $\theta_d = \sum_{k \in U_d} y_k$, $d = 1, \ldots, D$, où U_d sont D sous-ensembles disjoints de la population. À partir d'une enquête probabiliste, on peut estimer θ_d par $\hat{\theta}_d = \sum_{k \in s_{P,d}} w_k y_k$, où $s_{P,d}$ est l'ensemble des unités échantillonnées qui tombent dans le domaine d. On appelle $\hat{\theta}_d$ l'estimateur direct de θ_d car il utilise des valeurs y_k seulement pour des unités appartenant au domaine d. Les techniques d'estimation sur petits domaines mènent généralement à des

estimateurs indirects qui combinent les valeurs échantillonnées y_k du domaine d avec des valeurs y_k pour des unités en-dehors du domaine d. On va supposer qu'on a accès à un vecteur de variables auxiliaires disponibles au niveau des domaines et qui proviennent de sources indépendantes de l'échantillon probabiliste. On va noter ce vecteur pour le domaine d par $\mathbf{x}_d$. Par exemple, on pourrait considérer le vecteur $\mathbf{x}_d' = \left(N_d,\ N_d\hat{\mu}_d^{\mathrm{NP}}\right)$, où N_d est la taille de population dans le domaine d, $\hat{\mu}_d^{\mathrm{NP}} = \sum_{k \in s_{\mathrm{NP},d}} y_k^* / n_d^{\mathrm{NP}}$ est la moyenne de la variable y^* dans un échantillon non probabiliste, $s_{\mathrm{NP},d}$ est l'ensemble des unités de l'échantillon non probabiliste qui tombent dans le domaine d et n_d^{NP} est la taille de l'échantillon non probabiliste dans le domaine d. Si la taille de population N_d n'est pas connue, on peut la remplacer par une estimation indépendante de l'enquête probabiliste. On va noter par $\mathbf{X}$, la matrice qui contient les valeurs du vecteur $\mathbf{x}_d$, $d = 1, \ldots, D$. Il est à noter que le vecteur $\boldsymbol{\delta}$ est caché dans la matrice $\mathbf{X}$ dans cette section.

Le modèle de Fay-Herriot a deux composantes : le modèle d'échantillonnage et le modèle de lien. Le modèle d'échantillonnage est fondé sur l'hypothèse que, conditionnellement à $\boldsymbol{\Omega}_P$, les estimateurs directs $\hat{\theta}_d$ sont indépendants et sans biais, c'est-à-dire que $E\left(\hat{\theta}_d \mid \boldsymbol{\Omega}_P\right) = \theta_d$. Leur variance par rapport au plan est notée par $\psi_d = \mathrm{var}\left(\hat{\theta}_d \mid \boldsymbol{\Omega}_P\right)$. Le modèle d'échantillonnage est habituellement écrit sous la forme :

$$\hat{\theta}_d = \theta_d + e_d, \tag{4.9}$$

où e_d est l'erreur d'échantillonnage telle que $E(e_d \mid \boldsymbol{\Omega}_P) = 0$ et $\mathrm{var}(e_d \mid \boldsymbol{\Omega}_P) = \psi_d$. L'hypothèse d'indépendance des estimateurs $\hat{\theta}_d$ (et donc des erreurs d'échantillonnage e_d) peut être mise en doute lorsque les strates ne coïncident pas avec les domaines d'intérêt. La section 8.2 de Rao et Molina (2015) discutent de méthodes pour tenir compte des corrélations dans les erreurs d'échantillonnage. En pratique, on suppose souvent que ces corrélations sont faibles et on les ignore.

Le modèle de lien suppose que, conditionnellement à $\mathbf{X}$, les totaux θ_d sont indépendants, $E(\theta_d \mid \mathbf{X}) = \mathbf{x}_d'\boldsymbol{\beta}$ et $\mathrm{var}(\theta_d \mid \mathbf{X}) = b_d^2\sigma_v^2$, où b_d sont des constantes connues utilisées pour contrôler l'hétéroscédasticité, et $\boldsymbol{\beta}$ et σ_v^2 sont des paramètres inconnus du modèle. Le modèle de lien est habituellement écrit sous la forme :

$$\theta_d = \mathbf{x}_d'\boldsymbol{\beta} + b_d v_d, \tag{4.10}$$

où v_d est l'erreur du modèle telle que $E(v_d \mid \mathbf{X}) = 0$ et $\mathrm{var}(v_d \mid \mathbf{X}) = \sigma_v^2$. Lorsque les paramètres d'intérêt, θ_d, sont des totaux, il est souvent approprié de poser $b_d = N_d$. À partir de (4.9) et (4.10), on obtient le modèle combiné :

$$\hat{\theta}_d = \mathbf{x}_d'\boldsymbol{\beta} + a_d, \tag{4.11}$$

où $a_d = b_d v_d + e_d$ est l'erreur combinée. Lorsqu'on utilise le modèle de Fay-Herriot (4.11), on choisit habituellement de faire les inférences conditionnellement à $\mathbf{X}$. On peut facilement montrer que $E(a_d \mid \mathbf{X}) = 0$ et $\mathrm{var}(a_d \mid \mathbf{X}) = b_d^2\sigma_v^2 + \tilde{\psi}_d$, où $\tilde{\psi}_d = E(\psi_d \mid \mathbf{X})$ est appelée la variance lissée (Beaumont et Bocci, 2016; et Hidiroglou, Beaumont et Yung, 2019).

Supposons maintenant qu'on considère prédire le total θ_d au moyen d'un prédicteur linéaire $\hat{\theta}_d^{\text{LIN}} = \sum_{i=1}^{D} \lambda_{di}\hat{\theta}_i$, où λ_{di} sont des constantes à déterminer. Un prédicteur linéaire utilise toutes les données de l'échantillon probabiliste pour prédire θ_d et pas seulement celles provenant du domaine d. C'est ainsi qu'il tire son efficacité. Cependant, les prédicteurs linéaires ne sont pas tous appropriés pour la prédiction de θ_d. Une stratégie souvent utilisée pour choisir les constantes λ_{di} consiste à minimiser la variance de l'erreur de prédiction, $\text{var}\left(\hat{\theta}_d^{\text{LIN}} - \theta_d \mid \mathbf{X}\right)$, sous la contrainte que le prédicteur doit être sans biais, $E\left(\hat{\theta}_d^{\text{LIN}} - \theta_d \mid \mathbf{X}\right) = 0$. Le prédicteur résultant, appelé meilleur prédicteur linéaire sans biais, est noté, $\hat{\theta}_d^{\text{BLUP}}$, et peut être écrit sous la forme (voir, par exemple, Rao et Molina, 2015) :

$$\hat{\theta}_d^{\text{BLUP}} = \gamma_d\hat{\theta}_d + (1 - \gamma_d)\,\mathbf{x}_d'\hat{\boldsymbol{\beta}}, \tag{4.12}$$

où $\gamma_d = b_d^2\sigma_v^2 \,/\, (b_d^2\sigma_v^2 + \tilde{\psi}_d)$ est compris entre 0 et 1 et

$$\hat{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \frac{\mathbf{x}_d\mathbf{x}_d'}{b_d^2\sigma_v^2 + \tilde{\psi}_d}\right)^{-1} \sum_{d=1}^{D} \frac{\mathbf{x}_d}{b_d^2\sigma_v^2 + \tilde{\psi}_d}\,\hat{\theta}_d.$$

Le prédicteur (4.12) est une moyenne pondérée de l'estimateur direct $\hat{\theta}_d$ et de la prédiction, $\mathbf{x}_d'\hat{\boldsymbol{\beta}}$, souvent appelée l'estimateur synthétique. On donne plus de poids à l'estimateur direct quand la variance lissée, $\tilde{\psi}_d$, est petite par rapport à la variance du modèle de lien, $b_d^2\sigma_v^2$. Le prédicteur $\hat{\theta}_d^{\text{BLUP}}$ est alors semblable à l'estimateur direct. Cette situation survient normalement quand la taille d'échantillon dans le domaine est grande. À l'opposé, si l'estimateur direct est instable et a une grande variance lissée alors on donne plus de poids à l'estimateur synthétique. Si le nombre de domaines est grand, la variance de prédiction de $\hat{\theta}_d^{\text{BLUP}}$, $\text{var}\left(\hat{\theta}_d^{\text{BLUP}} - \theta_d \mid \mathbf{X}\right)$, est approximativement égale à $\gamma_d\tilde{\psi}_d$. Puisque $\text{var}\,(\hat{\theta}_d - \theta_d \mid \mathbf{X}) = \tilde{\psi}_d$, la constante γ_d peut être interprétée comme étant un facteur de réduction de variance obtenu en utilisant $\hat{\theta}_d^{\text{BLUP}}$ plutôt que $\hat{\theta}_d$. La réduction de variance est donc plus importante quand γ_d est petit, c'est-à-dire quand l'estimateur direct n'est pas précis. En contrepartie, si le modèle de lien n'est pas correctement spécifié, le risque d'un biais important est plus grand quand γ_d est petit. Pour mieux comprendre ce point, supposons que le vrai modèle de lien est tel que $E(\theta_d \mid \mathbf{X}) = \mu(\mathbf{x}_d)$ pour une certaine fonction $\mu(\cdot)$. Sous ce modèle, on peut montrer que le biais du prédicteur $\hat{\theta}_d^{\text{BLUP}}$ est donné par

$$E\left(\hat{\theta}_d^{\text{BLUP}} - \theta_d \mid \mathbf{X}\right) = -(1 - \gamma_d)\left(\mu(\mathbf{x}_d) - \mathbf{x}_d'\boldsymbol{\beta}_0\right), \tag{4.13}$$

où

$$\boldsymbol{\beta}_0 = \left(\sum_{d=1}^{D} \frac{\mathbf{x}_d\mathbf{x}_d'}{b_d^2\sigma_v^2 + \tilde{\psi}_d}\right)^{-1} \sum_{d=1}^{D} \frac{\mathbf{x}_d}{b_d^2\sigma_v^2 + \tilde{\psi}_d}\,\mu(\mathbf{x}_d).$$

Si le modèle linéaire $\mu(\mathbf{x}_d) = \mathbf{x}_d'\boldsymbol{\beta}$ est valide alors le biais disparaît. Autrement, le biais n'est pas nul et augmente à mesure que γ_d diminue ou que l'erreur de spécification du modèle de lien,

$\mu(\mathbf{x}_d) - \mathbf{x}_d'\boldsymbol{\beta}_0$, augmente. Quand γ_d est près de 1, le biais est généralement négligeable mais la réduction de variance l'est aussi.

Remarque : Il est à noter que le prédicteur $\hat{\theta}_d^{\text{BLUP}}$ et le biais (4.13) dépendent de la variance σ_v^2. Si le modèle linéaire (4.10) n'est pas valide, les paramètres $\boldsymbol{\beta}$ et σ_v^2 n'existent plus. On peut toutefois continuer de postuler le modèle de lien (4.10) et estimer ses paramètres à partir des données observées comme si le modèle était valide. On peut voir σ_v^2, qui intervient dans le calcul du prédicteur $\hat{\theta}_d^{\text{BLUP}}$ et du biais (4.13), comme étant la valeur vers laquelle un estimateur de σ_v^2 converge.

Le prédicteur (4.12) ne peut pas être calculé parce qu'il dépend des variances inconnues σ_v^2 et $\tilde{\psi}_d$. Lorsqu'on remplace σ_v^2 et $\tilde{\psi}_d$ dans (4.12) par des estimateurs $\hat{\sigma}_v^2$ et $\hat{\tilde{\psi}}_d$, on obtient le meilleur prédicteur linéaire sans biais empirique, noté par $\hat{\theta}_d^{\text{EBLUP}}$. Il existe plusieurs méthodes pour estimer σ_v^2 (voir Rao et Molina, 2015). Une des méthodes les plus fréquemment utilisées est celle du maximum de vraisemblance restreint. Pour estimer $\tilde{\psi}_d$, on peut supposer qu'on dispose d'un estimateur sans biais sous le plan de ψ_d, noté par $\hat{\psi}_d$. Cette hypothèse s'écrit formellement : $E(\hat{\psi}_d \mid \boldsymbol{\Omega}_P) = \psi_d$. Il s'ensuit que $E(\hat{\psi}_d \mid \mathbf{X}) = \tilde{\psi}_d$. L'estimateur $\hat{\psi}_d$ est donc sans biais pour $\tilde{\psi}_d$ mais peut être très instable quand la taille d'échantillon dans le domaine est petite. Une approche plus efficace pour estimer $\tilde{\psi}_d$ consiste à modéliser $\hat{\psi}_d$ en fonction des variables auxiliaires $\mathbf{x}_d$. En pratique, on a souvent recours à un modèle linéaire pour $\log(\hat{\psi}_d)$ et on suppose que les erreurs de ce modèle suivent une loi normale (par exemple, Rivest et Belmonte, 2000). Beaumont et Bocci (2016), voir aussi Hidiroglou et coll. (2019), fournissent une méthode de moments pour estimer $\tilde{\psi}_d$ qui ne nécessite pas l'hypothèse de normalité.

Le modèle de Fay-Herriot requiert la disponibilité de données auxiliaires uniquement au niveau des domaines d'intérêt. La variable y doit être mesurée sans erreur dans l'enquête probabiliste mais il n'est pas essentiel que la source auxiliaire soit parfaite, ce qui ouvre la porte à toutes sortes de fichiers externes à l'enquête probabiliste tels que des fichiers de données massives. Kim, Wang, Zhu et Cruze (2018) est un exemple récent où on a utilisé une extension du modèle de Fay-Herriot avec des données auxiliaires provenant d'images satellite. Les méthodes d'estimation sur petits domaines permettent souvent d'obtenir des réductions de variance significatives, parfois impressionnantes (voir, par exemple, Hidiroglou et coll., 2019). Le prix à payer pour ces gains est l'introduction d'hypothèses de modèle et le risque que ces hypothèses ne soient pas appropriées. La validation du modèle est donc une étape critique de la production d'estimations sur petits domaines tout comme c'est le cas de toute approche fondée sur un modèle.

Les méthodes d'estimation sur petits domaines sont généralement utilisées pour améliorer l'efficacité d'estimateurs pour des domaines dont la taille d'échantillon est petite. Elles pourraient également être utilisées pour réduire les coûts de collecte et le fardeau sur les répondants en réduisant la taille d'échantillon globale d'une enquête probabiliste pour quelques variables de l'enquête si ce n'est pas toutes les variables. Les estimations obtenues à partir de l'échantillon réduit et du modèle de Fay-Herriot, par exemple, pourraient ainsi avoir une précision similaire aux estimations directes de l'enquête probabiliste obtenues à partir de l'échantillon complet. Dans ce contexte, les méthodes d'estimation sur petits domaines ne seraient pas utilisées pour améliorer la précision pour les domaines contenant peu d'unités mais plutôt pour diminuer l'effort global de collecte tout en préservant la qualité des estimations.

5 Conclusion

Dans cet article, nous avons présenté quelques méthodes qui utilisent des données d'une source non probabiliste tout en conservant un cadre statistique qui permet de faire des inférences valides. Ceci est à notre avis essentiel pour les agences nationales de statistique car sans ce cadre les mesures habituelles de la qualité des estimations, telles que les estimations de la variance ou de l'erreur quadratique moyenne, disparaissent et il devient difficile de tirer des conclusions justes. L'utilisation de données d'une source non probabiliste ne vient pas sans risques. Pour les approches fondées sur un modèle, il nous semble incontournable de planifier suffisamment de temps et de ressources à la modélisation. La littérature sur la statistique classique regorge d'outils permettant de valider les hypothèses d'un modèle. Bien que ce sujet n'ait pas été abordé adéquatement aux sections précédentes, une validation minutieuse des hypothèses demeure néanmoins une étape primordiale dans le succès de ces approches (Chambers, 2014) et constitue une des recommandations formulées par Baker et coll. (2013).

L'estimation de la variance ou de l'erreur quadratique moyenne des estimateurs décrits dans les sections précédentes est également un sujet important que nous avons omis. Ce problème ne pose toutefois pas de difficultés particulières, en général, et plusieurs méthodes existent pour estimer la variance ou l'erreur quadratique moyenne. Pour les approches fondées sur le plan de sondage, le sujet a été amplement couvert dans la littérature (voir, par exemple, Wolter, 2007). Cela est également vrai pour les méthodes d'estimation sur petits domaines (voir Rao et Molina, 2015) ou pour l'approche par calage (voir Valliant, Dorfman et Royall, 2000). Néanmoins, il pourrait être utile que des travaux de recherche soient entrepris pour traiter adéquatement cette question dans certains cas particuliers, notamment lorsqu'on pondère par l'inverse du score de propension ou lorsqu'on effectue un appariement statistique par donneur le plus proche.

Nous avons supposé que la source non probabiliste était un sous-ensemble de la population d'intérêt et qu'elle pouvait être sujette aux erreurs de mesure. Il existe toutefois d'autres défauts possibles des sources non probabilistes. Par exemple, elles pourraient contenir des doublons ou des unités en dehors de la population. Cela pourrait rendre inutilisables certaines des méthodes discutées dans cet article, particulièrement les méthodes fondées sur le plan de sondage. Il pourrait donc être utile de s'attaquer à ces problèmes dans le futur.

Nous nous sommes principalement limités à décrire quelques méthodes qui permettent d'utiliser les données d'un échantillon non probabiliste, combinées ou non à des données d'une enquête probabiliste, une fois que toutes ces données ont été recueillies et traitées. Il existe plusieurs autres méthodes qui utilisent des données de sources non probabilistes lors des différentes étapes d'une enquête probabiliste. Par exemple, on peut utiliser une ou plusieurs sources non probabilistes pour créer une base de sondage ou améliorer sa couverture. On peut aussi utiliser de telles sources dans un contexte d'échantillonnage à partir de bases multiples ou encore les utiliser pour remplacer la collecte de certaines variables ou imputer les valeurs manquantes d'une enquête probabiliste. Ces sujets n'ont pas été traités dans cet article mais sont passés en revue dans celui de Lohr et Raghunathan (2017).

La littérature sur l'intégration de données d'un échantillon probabiliste et non probabiliste est plutôt récente. Il existe cependant plusieurs méthodes qui combinent des données de deux enquêtes probabilistes (ex. : Hidiroglou, 2001; Merkouris, 2004; Ybarra et Lohr, 2008; Merkouris, 2010; et Kim et Rao, 2012). De telles méthodes pourraient être utilisées pour d'abord combiner deux enquêtes probabilistes avant de les intégrer à une source non probabiliste au moyen d'une des méthodes de la section 4. Par exemple, si le total $\mathbf{T}_x$ n'est pas connu, on pourrait envisager l'estimer au moyen de plus d'une enquête probabiliste et ensuite utiliser ce total estimé dans l'approche par calage. Il reste à évaluer si une telle stratégie permettrait d'obtenir des gains d'efficacité significatifs.

Est-ce que les enquêtes probabilistes sont vouées à disparaître pour la production de statistiques officielles ? La question est pertinente dans le contexte actuel des enquêtes menées par les agences nationales de statistique où on observe des coûts de collecte élevés et des taux de réponse de plus en plus faibles. À notre avis, le moment n'est pas encore venu car les alternatives ne sont pas assez fiables et générales pour éliminer le recours aux enquêtes probabilistes sans sacrifier sévèrement la qualité des estimations. À la section 4, nous avons mentionné que le calage et la pondération par l'inverse du score de propension peuvent permettre d'éliminer l'utilisation d'une enquête probabiliste à condition qu'un vecteur de totaux de population $\mathbf{T}_x$ soit accessible à partir d'un recensement ou d'une source administrative exhaustive. En général, ces totaux connus ne seront pas assez nombreux et efficaces pour réduire suffisamment le biais de sélection d'un échantillon non probabiliste. Pour contourner ce problème, on a suggéré dans la littérature de compléter $\mathbf{T}_x$ par d'autres totaux estimés par une enquête probabiliste de bonne qualité. Il nous semble que c'est de cette façon qu'on pourra réduire le biais de façon significative et vraiment tirer avantage des méthodes de calage et de pondération par l'inverse du score de propension présentées à la section 4. Bien sûr, certaines enquêtes probabilistes dont les taux de réponse sont très faibles et/ou pour lesquelles la qualité des données recueillies est douteuse pourront occasionnellement être éliminées au profit de données provenant de sources non probabilistes. À notre avis, la plupart des enquêtes menées par Statistique Canada ne tombent pas dans cette catégorie et, même si elles ne sont pas parfaites, continuent de fournir des informations fiables pour répondre aux besoins des utilisateurs et prendre des décisions éclairées. L'élimination complète des enquêtes probabilistes semble grandement improbable à court ou moyen terme. On peut toutefois s'attendre à une réduction de leur utilisation dans le futur afin de contrôler les coûts et le fardeau sur les répondants.

Bibliographie

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

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

Beaumont, J.-F., et Bissonnette, J. (2011). Estimation de la variance sous imputation composite : méthodologie programmée dans le SEVANI. *Techniques d'enquête*, 37, 2, 183-192. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011002/article/11605-fra.pdf.

Beaumont, J.-F., et Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37, 400-416.

Beaumont, J.-F., et Bocci, C. (2016). *Small Area Estimation in the Labour Force Survey*. Document présenté au comité consultatif des méthodes statistiques, mai 2016, Statistique Canada.

Beaumont, J.-F., Bocci, C. et Hidiroglou, M. (2014). On weighting late respondents when a follow-up subsample of nonrespondents is taken. Document présenté au comité consultatif des méthodes statistiques, mai 2014, Statistique Canada.

Beaumont, J.-F., Haziza, D. et Bocci, C. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, 30, 607-621.

Berg, E., Kim, J.-K. et Skinner, C. (2016). Imputation under informative sampling. *Journal of Survey Statistics and Methodology*, 4, 436-462.

Bethlehem, J. (2009). The rise of survey sampling. Document de discussion (09015), Statistics Netherlands, La Haye.

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching. *Social Science Computer Review*, 34, 59-77.

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

Chambers, R. (2014). Survey sampling in official statistics – Some thoughts on directions. *Proceedings of the 2014 International Methodology Symposium*, Statistique Canada, Ottawa, Canada.

Chen, Y., Li, P. et Wu, C. (2019). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association* (publié en ligne).

Chen, J.K.T., Valliant, R.L. et Elliott, M.R. (2018). Calage assisté par un modèle pour des données de sondage non probabiliste en utilisant le LASSO adaptatif. *Techniques d'enquête*, 44, 1, 125-155. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2018001/article/54963-fra.pdf.

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 https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2014002/article/14128-fra.pdf.

Couper, M.P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64, 464-494.

Couper, M.P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods*, 7, 145-156.

Deville, J.-C. (1991). Une théorie des enquêtes par quotas. *Techniques d'enquête*, 17, 2, 177-195. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1991002/article/14504-fra.pdf.

Deville, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, Sherbrooke, Canada.

Deville, J.-C., et Dupont, F. (1993). Non-réponse : principes et méthodes. *Actes des Journées de Méthodologie Statistique*, 53-69, 15 et 16 décembre 1993, INSEE, Paris.

Deville, J.-C., et Lavallée, P. (2006). Sondage indirect : les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*, 32, 2, 185-196. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006002/article/9551-fra.pdf.

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

D'Orazio, M., Di Zio, M. et Scanu, M. (2006). *Statistical Matching: Theory and Practice*. New York: John Wiley & Sons, Inc.

Dutwin, D., et Buskirk, T.D. (2017). Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly*, 81, 213-249.

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

Eltinge, J.L., et Yansaneh, I.S. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec application à la non-réponse aux questions sur le revenu de la *U.S. Consumer Expenditure Survey*. *Techniques d'enquête*, 23, 1, 37-45. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1997001/article/3103-fra.pdf.

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

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

Haziza, D., et Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *Revue Internationale de Statistique*, 75, 25-43.

Haziza, D., et Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32, 206-226.

Haziza, D., et Lesage, É. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32, 129-145.

Hidiroglou, M.A. (2001). L'échantillonnage double. *Techniques d'enquête*, 27, 2, 157-169. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2001002/article/6091-fra.pdf.

Hidiroglou, M.A., Beaumont, J.-F. et Yung, W. (2019). Élaboration d'un système d'estimation sur petits domaines à Statistique Canada. *Techniques d'enquête*, 45, 1, 107-133. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2019001/article/00009-fra.pdf.

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

Iannacchione, V.G., Milne, J.G. et Folsom, R.E. (1991). Response probability weight adjustments using logistic regression. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 637-642, Alexandria, VA.

Kalton, G. (2019). Developments in survey research over the past 60 years: A personal perspective. *Revue Internationale de Statistique*, 87, S10-S30.

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

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

Kim, J.K., et Tam, S.M. (2020). Data integration by combining big data and survey data for finite population inference. Manuscrit non publié.

Kim, J.K., et Wang, Z. (2019). Sampling techniques for big data analysis. *Revue Internationale de Statistique*, 87, S177-S191.

Kim, J.K., Wang, Z., Zhu, Z. et Cruze, N.B. (2018). Combining survey and non-survey data for improved sub-area prediction using a multi-level model. *Journal of Agricultural, Biological and Environmental Statistics*, 23, 175-189.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 2, 149-160. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006002/article/9547-fra.pdf.

Kott, P.S. (2019). A partially successful attempt to integrate a Web-recruited cohort into an address-based sample. *Survey Research Methods*, 13, 95-101.

Laflamme, F., et Karaganis, M. (2010). Development and implementation of responsive design for CATI surveys at Statistics Canada. *Proceedings of the European Conference on Quality in Official Statistics*, Helsinki, Finland, mai 2010.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer.

Lavallée, P., et Brisbane, J. (2016). Sample matching: Towards a probabilistic approach for web surveys and big data? Document présenté au comité consultatif des méthodes statistiques, mai 2016, Statistique Canada.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel Web survey. *Journal of Official Statistics*, 22, 329-349.

Lesage, É. (2017). Combiner des données d'enquêtes probabilistes et des données massives non probabilistes pour estimer des paramètres de population finie. Manuscrit non publié.

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

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

Lundquist, P., et Särndal, C.-E. (2013). Aspects of responsive design with applications to the swedish living conditions survey. *Journal of Official Statistics*, 29, 557-582.

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

Mercer, A.W., Kreuter, F., Keeter, S. et Stuart, E.A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly*, 81, 250-271.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society: Series B*, 72, 27-48.

Miller, P.V. (2017). Is there a future for surveys? *Public Opinion Quarterly*, 81, 205-212.

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.

Phipps, P., et Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *Annals of Applied Statistics*, 6, 772-794.

Rancourt, E. (2019). Admin-first as a statistical paradigm for Canadian statistics: Meaning, challenges and opportunities. *Proceedings of Statistics Canada's 2018 International Methodology Symposium* (à paraitre).

Rao, J.N.K. (2005). Évaluation de l'interaction entre la théorie et la pratique des enquête par sondage. *Techniques d'enquête*, 31, 2, 127-151. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2005002/article/9040-fra.pdf.

Rao, J.N.K. (2020). Making inference by combining data from multiple sources: An appraisal. *Sankhyā* (à l'étude).

Rao, J.N.K., et Fuller, W. (2017). Théorie et méthodologie des enquêtes par sondage : orientations passées, présentes et futures. *Techniques d'enquête*, 43, 2, 159-176. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2017002/article/54888-fra.pdf.

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

Rao, J.N.K., Scott, A.J. et Benhin, E. (2003). Défaire les structures des données d'enquête complexes : théorie élémentaire et applications de l'échantillonnage inverse (avec discussion). *Techniques d'enquête*, 29, 2, 119-143. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2003002/article/6787-fra.pdf.

Rässler, S. (2012). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Notes de cours en statistiques, New York: Springer, 168.

Rivers, D. (2007). Sampling from web surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.

Rivest, L.-P., et Belmonte, E. (2000). Une erreur quadratique moyenne conditionnelle des estimateurs régionaux. *Techniques d'enquête*, 26, 1, 79-90. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2000001/article/5179-fra.pdf.

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

Royall, R.M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.

Särndal, C.-E., Lumiste, K. et Traat, I. (2016). Est-ce que la réduction du déséquilibre de la réponse accroît l'exactitude des estimations de l'enquête ? *Techniques d'enquête*, 42, 2, 233-253. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2016002/article/14663-fra.pdf.

Schonlau, M., et Couper, M.P. (2017). Options for conducting Web surveys. *Statistical Science*, 32, 279-292.

Schouten, B., Calinescu, M. et Luiten, A. (2013). Optimiser la qualité de la réponse au moyen de plans de collecte adaptatifs. *Techniques d'enquête*, 39, 1, 33-66. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2013001/article/11824-fra.pdf.

Squire, P. (1988). Why the 1936 *Literary Digest* Poll failed. *Public Opinion Quarterly*, 52, 125-133.

Tourangeau, R., Brick, J.M., Lohr, S. et Li, J. (2017). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society*, 180, 201-223.

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

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

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

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Ybarra, L.M., et Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95, 919-931.

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