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

Techniques d'enquête

Ajustements pour la non-réponse dans les plans stratifiés assortis de modèles aux spécifications erronées

par Ismael Flores Cervantes et J. Michael Brick

Date de diffusion : le 22 juin 2016

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

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

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

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

Programme des services de dépôt

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Signes conventionnels dans les tableaux

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

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

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, juin 2016
Vol. 42, nº 1, pp. 173-190
Statistique Canada, nº 12-001-X au catalogue

Ajustements pour la non-réponse dans les plans stratifiés assortis de modèles aux spécifications erronées

Ismael Flores Cervantes et J. Michael Brick[1]

Résumé

L'ajustement des poids de base au moyen de classes de pondération est une méthode communément employée pour composer avec la non-réponse totale. Une approche courante consiste en l'application d'ajustements pour la non-réponse pondérés selon l'inverse de la propension à répondre supposée des répondants dans les classes de pondération en vertu d'une méthode de quasi-randomisation. Little et Vartivarian (2003) ont remis en question l'utilité de la pondération du facteur d'ajustement. Dans la pratique, les modèles utilisés sont mal spécifiés; il est donc essentiel de comprendre l'incidence que peut avoir la pondération dans un tel cas. Le présent article décrit les effets, sur les estimations corrigées pour la non-réponse de moyennes et de totaux pour l'ensemble de la population et pour certains domaines qui ont été calculés selon l'inverse pondéré et non pondéré de la propension à répondre en vertu de plans d'échantillonnage aléatoires simples stratifiés. Le rendement de ces estimateurs est évalué dans différentes conditions, par exemple selon des répartitions différentes de l'échantillon, le mécanisme de réponse et la structure de population. Les résultats montrent que pour les scénarios étudiés, l'ajustement pondéré présente des avantages considérables pour l'estimation des totaux, et que le recours à un ajustement non pondéré peut donner lieu à des biais importants, sauf dans des cas très limités. En outre, contrairement aux estimations non pondérées, les estimations pondérées ne sont pas sensibles à la façon dont la répartition de l'échantillon est faite.

Mots-clés : Non-réponse; stratification; poids d'échantillonnage; repondération des classes de pondération.

1 Introduction

L'ajustement des poids de base au moyen de classes de pondération pour tenir compte de la non-réponse totale est une méthode couramment employée pour pondérer les données d'enquête, mais les chercheurs et les organismes d'enquête ne font pas tous ces ajustements de la même manière. Little et Vartivarian (2003), ci-après désignés « L et V », constatent que le recours à un facteur d'ajustement pour la non-réponse pondéré en fonction de l'inverse de la probabilité de sélection semble être l'approche la plus courante. Ils soulignent aussi que le fait d'utiliser des poids de sondage pour calculer un ajustement pondéré pour la non-réponse n'élimine pas le biais de non-réponse dans les estimations de la moyenne de population lorsque le mécanisme de réponse n'est pas précisé correctement dans le modèle d'ajustement de la pondération. L et V ont donc réalisé une étude par simulation à l'aide d'un plan d'échantillonnage simple stratifié afin d'examiner l'effet de la pondération des facteurs d'ajustement pour la non-réponse. Ils ont conclu que la pondération de l'ajustement pour la non-réponse est peu utile, voire inutile.

Afin d'éliminer le biais de non-réponse, les justifications théoriques pour l'ajustement pour la non-réponse exigent une modélisation exacte soit du mécanisme de réponse, soit de la variable cible; nous ne connaissons aucune théorie stipulant que la pondération selon l'inverse de la probabilité de sélection élimine complètement le biais lorsque les spécifications du modèle sont erronées (par exemple Kalton 1983; Little 1986; Little et Rubin 2002; Särndal et Lundström 2005). C'est pourquoi l'intégration dans la modélisation de l'ajustement pour la non-réponse que préconisent L et V est essentielle à une bonne pratique

1. Ismael Flores Cervantes et J. Michael Brick, Westat, 1600 Research Blvd, Rockville, Maryland, États-Unis, 20850. Courriel : ismaelflorescervantes@westat.com.

statistique. Toutefois, la spécification exacte d'un modèle hautement prédictif est un objectif qu'il n'est pas possible d'atteindre dans la plupart des enquêtes à cause de la complexité du phénomène et du fait qu'il existe rarement des variables auxiliaires suffisamment puissantes. Les recherches visant à trouver de meilleures données auxiliaires pour cette modélisation ont mené à l'exploration des paradonnées, mais les modèles qui font appel à ces données sont toujours associés à de faibles corrélations avec la propension à répondre (Kreuter, Olson, Wagner, Yan, Ezzati-Rice, Casas-Cordero, Lemay, Peytchev, Groves et Raghunathan 2010). Dans la pratique, on a recours à des modèles imparfaits et le biais de non-réponse n'est jamais complètement éliminé.

En conséquence, il importe de comprendre les effets des méthodes d'ajustement pour la non-réponse et de déterminer s'il est utile de pondérer l'ajustement pour la non-réponse lorsque les spécifications du modèle de réponse sont erronées. Bien que L et V insistent entre autres sur la nécessité d'inclure les variables de plan dans la modélisation de la non-réponse, certains chercheurs semblent avoir conclu que la pondération de l'ajustement est inutile (par exemple Chadborn, Baster, Delpech, Sabin, Sinka, Rice et Evans 2005; Haukoos et Newgard 2007). Cependant, la conclusion de L et V, selon laquelle la pondération du facteur d'ajustement pour la non-réponse est incorrecte ou inefficace, est fondée sur des comparaisons avec des modèles correctement spécifiés qui produisent toujours des estimations non biaisées. Leur suggestion de conditionner le modèle sur les variables de plan (dans le scénario de L et V, la variable de plan correspondait à la strate) a donné lieu à des estimateurs avec et sans pondération identiques. Leurs simulations étaient aussi axées sur un plan d'échantillonnage stratifié spécifique et ils n'ont tenu compte que de l'estimation des moyennes. Comme il est expliqué plus loin, ces limitations sont considérables et il convient de revoir les conclusions de certains quant à l'inutilité de pondérer l'ajustement.

Après L et V, des chercheurs ont examiné les effets de la pondération dans d'autres cas. Sukasih, Jang, Vartivarian, Cohen et Zhang (2009) ont comparé les ajustements pour la non-réponse avec et sans pondération à l'aide de simulations dans le contexte d'une enquête particulière. West (2009) a utilisé une simulation pour étudier les estimations des moyennes de population en vertu de plans d'échantillonnage plus complexes comprenant des grappes et des taux d'échantillonnage différentiels. Ces deux études ont conclu que la pondération des ajustements pour la non-réponse à l'aide des poids de sondage était utile comparativement à une approche de non-pondération, même si les différences obtenues après pondération n'étaient pas importantes. Après avoir évalué la robustesse des ajustements sur le plan théorique et décrit les conditions en vertu desquelles les divers estimateurs des moyennes de population étaient le moins influencés par le biais de non-réponse, Kott (2012) recommande une approche de pondération. D'autres recherches ont été menées sur la nécessité de pondérer pour estimer les coefficients des modèles de la propension à répondre (Wun, Ezzati-Rice, Diaz-Tena et Greenblatt 2007; Grau, Potter, Williams et Diaz-Tena 2006), mais cette piste de recherche est assez éloignée de la nôtre et nous ne l'abordons pas ici.

Dans le présent article, nous explorons l'effet de la pondération des ajustements pour la non-réponse lorsque le modèle de non-réponse est imparfait. Dans la section 2, nous prenons les résultats de L et V comme point de départ, pour aller plus loin et examiner les estimateurs pour les totaux et pour les moyennes et totaux de domaine; L et V n'ont tenu compte que des moyennes globales. À l'aide de la même population et du même scénario de simulation de base que L et V, nous examinons aussi l'effet de différentes répartitions de l'échantillon dans les strates, tandis que L et V n'ont utilisé qu'une seule répartition de

l'échantillon. Les résultats des simulations présentés à la section 3 révèlent des différences importantes des propriétés des estimateurs avec et sans pondération, qui varient selon la répartition de l'échantillon. Nous expliquons les comportements des estimateurs à l'aide d'approximations simples afin d'illustrer pourquoi ils sont différents. Bien que la pondération des facteurs d'ajustement ne donne pas toujours des estimations assorties d'un biais et d'une racine de l'erreur quadratique moyenne (reqm) plus faibles que ceux des estimations obtenues sans pondération, elle présente des avantages substantiels pour les estimations des totaux et fournit une protection contre les erreurs importantes qui pourraient découler d'une approche sans pondération. En conséquence, nous recommandons de pondérer lorsque le véritable mécanisme de réponse n'est pas entièrement connu. La section 4 donne les conclusions.

2 Scénario

Les poids de sondage compensent pour différents types de données manquantes : les poids d'échantillonnage ou de base compensent les unités non échantillonnées; les poids d'ajustement de non-couverture tiennent compte des unités qui ne font pas partie de la base de sondage; et les poids d'ajustement pour la non-réponse compensent les unités qui font partie de l'échantillon, mais qui ne répondent pas. Nous nous concentrons ici sur les poids d'ajustement pour la non-réponse et sur l'effet du recours aux poids de base pour établir les ajustements pour la non-réponse.

Commençons par l'estimateur de Horvitz-Thompson non ajusté pour le total :

$$\hat{y}_{na} = \sum_s R_i d_i y_i, \tag{2.1}$$

où d_i est l'inverse de la probabilité de sélection de l'unité $i, R_i = 1$ si l'unité i répond et $= 0$ autrement; la somme est calculée pour l'ensemble des unités de l'échantillon s. La moyenne du ratio est $\hat{\bar{y}}_{na} = \hat{y}_{na} \big/ \sum_s R_i d_i$. Si toutes les données de l'échantillon sont observées et que la base de sondage est complète, alors $E(\hat{y}_{na}) = Y$, et la moyenne du ratio est convergente pour $\bar{Y}$.

Lorsqu'il y a non-réponse totale, on présume que la réponse est une variable aléatoire et que la probabilité de réponse ou la propension à répondre $(\phi_i = \Pr(R_i = 1))$ correspond à la probabilité pour une phase supplémentaire d'échantillonnage (Särndal, Swensson et Wretman 1992). Si on suppose que $\phi_i > 0$ pour toutes les valeurs de i, alors le biais de non-réponse d'une moyenne de ratio estimée en vertu du modèle stochastique correspond à :

$$\text{biais}(\hat{\bar{y}}_{na}) \approx \bar{\phi}^{-1}\sigma_\phi\sigma_y\rho_{\phi,y}, \tag{2.2}$$

où $\bar{\phi}$ correspond à la moyenne de population des propensions à répondre, σ_ϕ est l'écart type de ϕ, σ_y est l'écart type de y et $\rho_{\phi,y}$ est la corrélation entre ϕ et y (Bethlehem 1988). La moyenne estimée pour les répondants est non biaisée si ϕ et y ne sont pas corrélés. Brick et Jones (2008) élargissent ces résultats à d'autres types de statistiques et d'estimateurs.

Pour réduire le biais de non-réponse, on peut utiliser les variables auxiliaires associées à l'échantillon pour étayer les ajustements pour la non-réponse en fonction des poids de base. Les ajustements peuvent être mis en œuvre par modélisation de la répartition de ϕ ou de y, ou encore des deux à l'aide des variables auxiliaires. Nous nous intéressons particulièrement à la modélisation du mécanisme de réponse.

Les propensions à répondre estimées sont appliquées comme si elles correspondaient aux probabilités réelles de réponse. En d'autres termes, le facteur d'ajustement pour la non-réponse correspond à l'inverse de la propension à répondre estimée pour l'unité échantillonnée $i\left(\hat{\phi}_i\right)$. La propension à répondre peut être estimée de différentes manières, par exemple par régression logistique, mais pour la plupart des enquêtes, on établit des groupes mutuellement exclusifs appelés classes de pondération ou groupes de réponse homogènes qui renferment des unités ayant des propensions estimées similaires et on ajuste les poids dans chaque groupe ou classe en fonction d'un facteur commun, par exemple $\hat{f}_c = \hat{\phi}_c^{-1}$ pour toutes les valeurs de $i \in c$ (Särndal et coll. 1992; Little 1986). En vertu de cette approche, l'estimateur ajusté est appelé un estimateur de classe de pondération et s'écrit comme suit :

$$\hat{y}_{cp} = \sum_c \sum_{i \in s_c} R_{ci} d_{ci} \hat{f}_c y_{ci}, \tag{2.3}$$

où $c = 1, 2, \ldots, C$ correspond aux classes d'ajustement pour la non-réponse et $i \in s_c$ est une unité échantillonnée de la classe c.

La question spécifique à laquelle nous nous intéressons ici est l'effet de la pondération du facteur d'ajustement. Le facteur non pondéré s'écrit

$$\hat{f}_c^{np} = \frac{\sum_{i \in s_c} \delta_{ci}}{\sum_{i \in s_c} R_{ci} \delta_{ci}} = \frac{n_{c+}}{r_{c+}}$$

où $\delta_{ci} = 1$ si $i \in c$ et $\delta_{ci} = 0$ si $i \notin c$, et n_{c+} et r_{c+} correspondent au nombre d'unités échantillonnées et répondantes de la classe c. Le facteur d'ajustement pondéré s'écrit

$$\hat{f}_c^{p} = \frac{\sum_{i \in s_c} d_{ci}}{\sum_{i \in s_c} R_{ci} d_{ci}} = \frac{\hat{N}_c}{\hat{N}'_c},$$

où $\hat{N}_c = \sum_{i \in s_c} d_{ci}$ et $\hat{N}'_c = \sum_{i \in s_c} R_{ci} d_{ci}$. Les facteurs correspondent aux taux de réponse non pondéré et pondéré, respectivement. En substituant les facteurs dans l'estimateur (2.3), on obtient deux nouveaux estimateurs (2.4) et (2.5) de la population totale. Il s'agit de deux estimateurs de classes de pondération, dont la notation a été modifiée pour mettre en évidence le taux de réponse utilisé (pondéré ou non pondéré).

$$\hat{y}_{trnp} = \sum_c \hat{f}_c^{np} \sum_{i \in r_c} d_{ci} y_{ci} = \sum_c \frac{n_{c+}}{r_{c+}} \sum_{i \in r_c} d_{ci} y_{ci}, \tag{2.4}$$

$$\hat{y}_{trp} = \sum_c \hat{f}_c^{p} \sum_{i \in r_c} d_{ci} y_{ci} = \sum_c \frac{\hat{N}_c}{\hat{N}'_c} \sum_{i \in r_c} d_{ci} y_{ci}. \tag{2.5}$$

Ces deux estimateurs constituent les éléments de base pour tous les types de statistiques que nous examinons dans l'étude par simulation. Par exemple, les estimateurs des moyennes, des moyennes de domaine et des ratios sont de simples fonctions des estimateurs (2.4) et (2.5).

Pour respecter la structure, la notation et les simulations de L et V, la présente étude est restreinte à la même population et à un échantillon aléatoire simple stratifié où deux strates sont définies par la variable de plan binaire Z, et où deux classes d'ajustement pour la non-réponse sont définies par une variable

auxiliaire binaire C, qui recoupe les strates comme indiqué dans le tableau 2.1. Nous avons remplacé la lettre X utilisée par L et V par la lettre C dans la cellule de pondération introduite ci-dessus afin de faciliter l'identification de la cellule d'ajustement pour la non-réponse. Comme dans l'étude de L et V, la taille de la population est fixée à $N = 10\,000$.

Tableau 2.1
Chiffres de population par strate *Z* et par cellule d'ajustement pour la non-réponse *C*

Strate d'échantillonnage	Cellule d'ajustement pour la non-réponse	
	***C* = 0**	***C* = 1**
Z = 0	3 064	3 931
Z = 1	2 079	926

Source : Little et Vartivarian (2003), qui ont utilisé *X* au lieu de *C*.

La variable d'intérêt, Y, est une variable binaire pour laquelle la probabilité que $Y = 1$ est définie par un modèle logistique où $\text{logit}(Y = 1|C, Z) = 0{,}5 + \gamma_C (C - \bar{C}) + \gamma_Z (Z - \bar{Z}) + \gamma_{CZ} (C - \bar{C})(Z - \bar{Z})$. La variable de réponse R est aussi binaire, et la probabilité que $R = 1$ est générée à partir d'un modèle logistique où $\text{logit}(R|C, Z) = 0{,}5 + \beta_C (C - \bar{C}) + \beta_Z (Z - \bar{Z}) + \beta_{CZ} (C - \bar{C})(Z - \bar{Z})$. Différentes populations et propensions à répondre sont générées en fonction des valeurs de $\gamma_C, \gamma_Z, \gamma_{CZ}, \beta_C, \beta_Z$ et β_{CZ} indiquées dans le tableau 2.2. Nous avons adopté la notation de L et V pour les modèles linéaires généralisés afin de faciliter la comparaison avec leurs travaux. Les valeurs indiquées dans le tableau sont les mêmes variables de population et de réponse que celles que L et V ont produites en affectant des valeurs à $(\gamma_C, \gamma_Z, \gamma_{CZ}, \beta_C, \beta_Z, \beta_{CZ})$. Dans la notation $[A]^B$ présentée au tableau 2.2, la population (Y) ou la propension à répondre (R) sont indiquées par l'exposant B, alors que les paramètres et les interactions du modèle pour la répartition de la population ou de la réponse sont indiqués par la lettre A entre crochets. Par exemple, le modèle logistique additif qui génère la répartition de Y dans la strate d'échantillonnage Z et la cellule de non-réponse C est indiqué comme suit : $[C + Z]^Y$. De même, les modèles où R dépend de C seulement, de Z seulement ou ni de C ni de Z sont indiqués respectivement par $[C]^R$, $[Z]^R$ et $[C + Z]^R$. L et V donnent plus de détails sur les motifs justifiant le choix de ces populations et modèles de réponse en particulier.

Tableau 2.2
Modèles pour la variable de résultat *Y* et la probabilité de réponse *R*

Modèle pour *Y* (variable d'intérêt)	Modèle pour *R* (propension à répondre)	Paramètres		
		γ_C, β_C	γ_Z, β_Z	γ_{CZ}, β_{CZ}
$[CZ]^Y$	$[CZ]^R$	2	2	2
$[C + Z]^Y$	$[C + Z]^R$	2	2	0
$[C]^Y$	$[C]^R$	2	0	0
$[Z]^Y$	$[Z]^R$	0	2	0
$[\phi]^Y$	$[\phi]^R$	0	0	0

Source : Little et Vartivarian (2003).

L et V ont calculé des estimations des moyennes comme suit, selon notre notation :

$$\hat{\bar{y}}_{trnp} = \frac{\hat{y}_{trnp}}{\sum_c \hat{f}_c^{np} \sum_{i \in s_c} R_{ci} d_{ci}} = \frac{\hat{y}_{trnp}}{\sum_c \hat{f}_c^{np} \hat{N}'_c}, \tag{2.6}$$

et

$$\hat{\bar{y}}_{trp} = \frac{\hat{y}_{trp}}{\sum_c \hat{f}_c^{p} \sum_{i \in s_c} R_{ci} d_{ci}} = \frac{\hat{y}_{trp}}{\sum_c \hat{N}_c}. \tag{2.7}$$

Les dénominateurs des moyennes sont des estimations de la taille de population N. Dans l'estimateur (2.7), le dénominateur est une constante égale à N, mais dans l'estimateur (2.6), le dénominateur est une variable aléatoire. Dans le scénario de simulation comprenant le plan d'échantillonnage aléatoire simple stratifié décrit ci-dessous, ou tout plan où $\sum_{i \in s} d_i = N$ pour chaque valeur de s, l'estimateur (2.7) se réduit à l'estimateur linéaire $\hat{\bar{y}}_{trp} = N^{-1} \hat{y}_{trp}$, tandis que l'estimateur (2.6) est un estimateur par le ratio. Il s'agit là d'un point important sur lequel nous reviendrons.

Les moyennes de domaine peuvent avoir des propriétés différentes des moyennes globales parce que les dénominateurs des moyennes de domaine pondérées et non pondérées sont des variables aléatoires, sauf quand les domaines concordent avec la strate d'échantillonnage et que les tailles des domaines et les tailles des strates sont connues. L et V n'abordent pas la question des domaines et n'ont donc pas examiné ces estimations dans le cadre de leur simulation. Nous avons établi des domaines en générant au hasard une variable aléatoire ν_i à partir d'une distribution uniforme (0, 1) et en définissant la fonction d'appartenance $\tau(a) = 1$ si $a < 0$ et $\tau(a) = 0$ si $a \geq 0$. Des moyennes de domaine de 50 % ont été créées en substituant $d_{ci}^* = \tau(\nu_i - 0{,}5)\, d_{ci}$ dans les expressions (2.6) et (2.7) afin de produire les estimateurs $\hat{\bar{y}}_{trnp;0,5}$ et $\hat{\bar{y}}_{trp;0,5}$, respectivement. Les estimateurs pondérés et non pondérés des totaux de domaine $\hat{y}_{trnp;0,5}$ et $\hat{y}_{trp;0,5}$ ont été établis de la même manière. Nous avons utilisé la même méthode pour créer des moyennes de domaine de 25 % et des totaux de domaine de 25 %. Comme nous nous intéressons à l'effet des ajustements pour la non-réponse sur les moyennes calculées sous forme d'estimateurs par ratio, d'autres domaines comme ceux qui correspondent à près de 100 % de la population ont été exclus de l'analyse parce que le dénominateur des moyennes de domaine est proche de la population totale constante N et qu'alors la moyenne devient un estimateur linéaire. Les domaines plus proches de 0 % ont été exclus à cause de la petite taille des échantillons.

3 Résultats

La simulation a été effectuée dans le logiciel R (R Development Core Team 2011) à partir de 10 000 tirages (L et V en ont utilisé 1 000). Nous avons évalué les estimateurs en calculant la racine de l'erreur quadratique moyenne (reqm) et le biais des estimations, le biais et la reqm étant mesurés par les écarts par rapport aux quantités de population comme l'ont fait L et V. Nous avons utilisé la même taille d'échantillon total (312) que dans la simulation, mais avec différentes répartitions de l'échantillon ou

différents taux d'échantillonnage relatifs entre les strates. Nous avons reproduit l'ensemble des 25 configurations de L et V; les résultats sont présentés dans le tableau S-1 des documents supplémentaires. Le tableau S-2 des documents supplémentaires comprend aussi les 25 configurations, mais présente le biais relatif des moyennes et des totaux avec et sans pondération, ainsi que les ratios des variances et des reqm des estimations non pondérées à ceux des estimations pondérées. Le biais relatif et les ratios des variances et des reqm facilitent les comparaisons entre les estimations. Les documents supplémentaires comprennent les erreurs de simulation estimées, qui sont toutes relativement petites. Pour les estimateurs et les taux d'échantillonnage donnés par L et V, nos résultats correspondent aux valeurs publiées, compte tenu des erreurs de simulation. Commençons par examiner le biais des estimateurs.

3.1 Biais

Il y a deux situations pour lesquelles il existe des résultats théoriques bien connus (Little et Rubin 2002). La première est lorsque la propension à répondre est la même dans toutes les cellules – les données manquent complètement au hasard (MCAR, de l'anglais *missing completely at random*); ces données de type MCAR correspondent au modèle $[\phi]^R = (\beta_c = 0, \beta_z = 0, \beta_{cz} = 0)$ de la dernière ligne du tableau 2.2. Lorsqu'on a des données de type MCAR, les facteurs d'ajustement non pondéré et pondéré ont la même espérance mathématique, et tous deux produisent des estimations non biaisées. Les résultats de la simulation présentés dans le tableau V de l'article de L et V (lignes 5, 10, 15, 20 et 25) confirment cette observation. La deuxième situation est lorsque la propension à répondre est indépendante de la strate, ce qui correspond à des données qui manquent au hasard (MAR, de l'anglais *missing at random*) selon le modèle de réponse $[\phi]^C = (\beta_c = 2, \beta_z = 0, \beta_{cz} = 0)$ de la troisième ligne du tableau 2.2. Nous considérons ces situations comme étant de type MAR parce que le biais de l'estimateur ne dépend pas de l'utilisation de données à propos de Z dans le modèle. Encore une fois, les estimations avec et sans pondération sont toutes deux sans biais, et les ajustements ont la même espérance mathématique. Les résultats de la simulation présentés dans le tableau V de L et V (lignes 3, 8, 13, 18 et 23) confirment cette observation de façon empirique.

Afin de nous concentrer sur la situation dans laquelle les spécifications du modèle sont erronées, nous ne présentons pas les résultats des simulations pour les situations de type MCAR et MAR dans le présent article; ces résultats sont toutefois présentés dans les documents supplémentaires. Il importe de souligner que même si les ajustements avec et sans pondération pour les modèles de type MCAR et MAR ont la même espérance mathématique, ils ne sont pas identiques. Après avoir simulé les deux approches en vertu de modèles de type MAR, Sukasih et coll. (2009) se sont prononcés en faveur d'une approche de pondération, principalement en raison de la variabilité moindre des estimations des totaux pour l'ensemble des simulations, même si les deux approches donnent des résultats non biaisés.

Comme il est précisé plus haut, les taux d'échantillonnage varient dans le cadre de nos simulations, tandis que la taille globale de l'échantillon est fixée à 312; L et V ont utilisé un taux d'échantillonnage unique. Quand les taux d'échantillonnage sont les mêmes dans toutes les strates (c'est-à-dire que l'échantillon est réparti proportionnellement dans toutes les strates), les poids d'échantillonnage sont les mêmes pour chaque strate et, en conséquence, les estimateurs avec et sans pondération sont identiques. Le taux d'échantillonnage selon une répartition proportionnelle joue un rôle important dans notre présentation, parce que les deux estimations doivent converger à cette étape.

Le graphique présenté à la figure 3.1 (à gauche) illustre les résultats de la simulation pour le biais des estimateurs avec et sans pondération du total pour $[CZ]^Y$ et $[C+Z]^R$. Nous avons choisi cette configuration (ligne 2 dans les tableaux de L et V) parce que les simulations de L et V montrent que la moyenne non pondérée est assortie d'un biais et d'une reqm plus faibles que la moyenne pondérée dans ce cas particulier. L'axe horizontal indique le taux d'échantillonnage relatif calculé comme étant le ratio du taux d'échantillonnage de $Z = 0$ à $Z = 1$ ou $N_0 n_0^{-1} / \left(N_1 n_1^{-1}\right)$. Le taux d'échantillonnage relatif employé par L et V était d'environ 2,25. On voit tout de suite que le biais de l'estimateur pondéré est pratiquement constant pour les différents taux d'échantillonnage, alors que le biais de l'estimateur non pondéré varie considérablement selon le taux d'échantillonnage relatif. Pour certains taux d'échantillonnage, le biais des estimateurs non pondérés du total peut être plus de deux fois celui de l'estimateur pondéré. Les deux types d'estimateur sont biaisés pour presque tous les taux d'échantillonnage relatifs, et l'estimateur qui a le biais le plus faible dépend du taux d'échantillonnage relatif. Lorsque les taux d'échantillonnage relatifs sont égaux (répartition proportionnelle), les estimateurs sans pondération et avec pondération ont le même biais, comme prévu. Cependant, dans la pratique, il n'est généralement pas possible de reconnaître l'effet du taux d'échantillonnage sur le biais et de choisir à l'avance la méthode d'ajustement qui permet de réduire le biais pour un échantillon particulier.

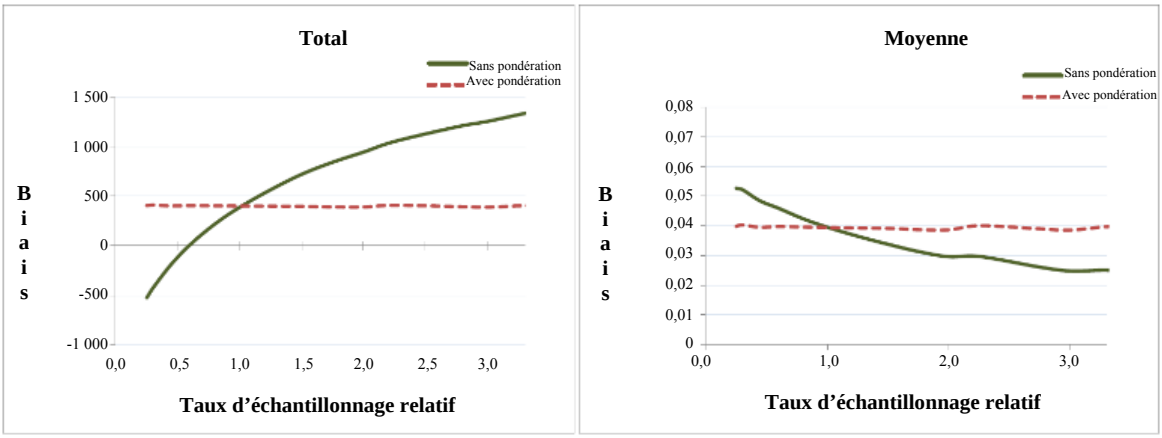

Figure 3.1 Biais des estimateurs avec et sans pondération pour le modèle de population $[CZ]^Y$ et le modèle de propension à répondre $[C+Z]^R$; le graphique de gauche correspond au total, et celui de droite, à la moyenne.

Pour comprendre ces résultats, nous avons appliqué des approximations standard qui se vérifient raisonnablement bien dans ce cas (c'est-à-dire $E\left(\eta^{-1}\right) \approx E^{-1}(\eta)\big)$. La valeur prévue approximative pour l'estimateur pondéré est

$$E\hat{y}_{trp} \approx \sum_z \sum_c \frac{N_c}{\left(\sum_z \phi_{cz} N_{cz}\right)} \phi_{cz} Y_{cz}, \tag{3.1}$$

où Y_{cz} est le total de population de la cellule cz. De même, la valeur prévue approximative pour l'estimateur non pondéré est

$$E\hat{y}_{trnp} \approx \sum_z \sum_c \frac{\left(\sum_z N_z n_z^{-1} N_{cz}\right)}{\left(\sum_z \phi_{cz} N_z n_z^{-1} N_{cz}\right)} \phi_{cz} Y_{cz}. \quad (3.2)$$

Si ϕ_{cz} est une constante (MCAR) ou ϕ_{cz} est une constante dans les cellules de pondération (MAR), alors les deux estimateurs ne sont pas biaisés à cet ordre d'approximation et concordent avec la théorie connue. Lorsque les taux d'échantillonnage sont les mêmes dans toutes les strates, les deux estimateurs ont la même valeur prévue (comme il est précisé plus haut, ils sont identiques dans ce cas). Surtout, ces approximations montrent que l'espérance mathématique de l'estimateur pondéré ne dépend pas du taux d'échantillonnage, mais que celle de l'estimateur non pondéré, elle, en dépend. Cela explique les courbes illustrées à la figure 3.1.

Quelques détails des estimations de la simulation pour cette configuration sont présentés dans le tableau 3.1 pour certains taux d'échantillonnage. Comme il est indiqué ci-dessus, les résultats complets de la simulation pour toutes les configurations et tous les taux d'échantillonnage utilisés pour dessiner les graphiques se trouvent dans les documents supplémentaires. Ces documents comprennent les biais relatifs, les ratios des variances et les ratios des reqm, qui constituent de meilleurs indicateurs pour évaluer l'incidence des ajustements sur les estimations. Nous avons constaté que pour toutes les configurations dont les estimations des totaux sont biaisées, les biais pour l'estimateur pondéré sont inférieurs d'un côté du taux d'échantillonnage relatif de 1, et supérieurs de l'autre côté. Toutes les configurations sont assorties d'un biais à peu près constant pour l'estimateur pondéré du total pour tous les taux d'échantillonnage relatifs, mais le biais de l'estimateur non pondéré varie en fonction du taux d'échantillonnage relatif.

Examinons maintenant les moyennes estimées – les seuls estimateurs examinés par L et V. Le graphique de droite de la figure 3.1 montre que le biais pour l'estimateur pondéré est encore une fois indépendant du taux d'échantillonnage relatif, alors que le biais de l'estimateur non pondéré varie en fonction du taux d'échantillonnage. L et V ont utilisé un taux d'échantillonnage de 2,25, ce qui explique pourquoi ils ont trouvé que l'estimateur non pondéré était associé à un biais inférieur pour la moyenne dans le cadre de leur exercice de simulation. Il importe de souligner deux choses à cet égard. D'une part, les biais pour les moyennes pour les deux méthodes d'ajustement sont tous relativement faibles, particulièrement par rapport aux biais relatifs potentiels des totaux obtenus à l'aide de l'estimateur non pondéré (graphique de gauche). D'autre part, il n'y a aucun moyen de déterminer si une estimation particulière tomberait du côté gauche ou du côté droit du taux d'échantillonnage relatif de 1. Le tableau 3.1 montre les biais estimés pour cette configuration.

Les graphiques illustrent aussi une relation quelque peu étonnante : les taux d'échantillonnage relatifs pour lesquels l'estimateur non pondéré du total est assorti d'un biais inférieur sont ceux pour lesquels l'estimateur non pondéré de la moyenne est assorti d'un biais supérieur. En d'autres termes, les moyennes se comportent différemment des totaux parce que la moyenne non pondérée est un ratio alors que la moyenne pondérée n'en est pas un. En conséquence, le biais relatif (br = biais/estimation) de l'estimateur non pondéré de la moyenne n'est pas égal au biais relatif de l'estimateur non pondéré du total (la relation est vérifiée pour l'estimateur pondéré). On peut approximer le biais relatif comme suit :

$$br\left(\hat{\bar{y}}_{trnp}\right) \approx \frac{1 + br\left(\hat{y}_{trnp}\right)}{1 + br\left(\hat{N}_{trnp}\right)},$$

où $\hat{N}_{trnp}$ est l'estimateur non pondéré du total (où $y_i = 1$ pour toutes les valeurs de i). Cette approximation se vérifie raisonnablement bien dans cette situation, puisque $\text{cov}\left(\hat{\bar{y}}_{trnp}, \hat{N}_{trnp}\right) / E\left(\hat{N}_{trnp}\right) \approx 0$. Le biais relatif de la moyenne non pondérée diminue donc quand les biais du numérateur et du dénominateur sont positivement corrélés.

Examinons maintenant les estimations de domaine – que L et V n'ont pas étudiées. Les biais pour les estimateurs du total de domaine avec et sans pondération et la relation avec les biais des estimateurs non pondérés qui varient en fonction du taux d'échantillonnage relatif sont les mêmes que ceux qui ont été observés pour les totaux globaux (voir le tableau 3.1), parce que les totaux de domaine demeurent des totaux et que les approximations (3.1) et (3.2) continuent de s'appliquer. Les moyennes de domaine sont aussi présentées dans le tableau, et elles aussi suivent la tendance des biais illustrée à la figure 3.1 pour la moyenne de l'échantillon complet. Il importe de souligner que les biais relatifs pour les estimations de la moyenne (globale et pour chaque domaine) ne varient pas beaucoup, la plupart d'entre eux se trouvant entre 5 % et 7 %.

Tableau 3.1
Biais (facteur 10 000), racine de l'erreur quadratique moyenne (facteur 10 000) et variance des estimateurs avec et sans pondération des moyennes et du total de l'échantillon complet et des domaines, configuration $[CZ]^Y$, $[C+Z]^R$ selon divers taux d'échantillonnage

				Taux d'échantillonnage relatif				
	Caractéristique	**Domaine**	**Ajustement**	**0,30**	**0,44**	**1,00**	**2,25**	**3,30**
Biais	Moyenne	Complet	trnp	515	491	404	301	248
			trp	398	403	404	404	394
		50 %	trnp	513	501	411	307	257
			trp	397	414	410	410	401
		25 %	trnp	523	498	407	298	252
			trp	408	411	407	400	395
	Total	Complet	trnp	-419	-184	401	1 058	1 335
			trp	398	403	404	404	394
		50 %	trnp	-214	-89	205	535	673
			trp	194	205	206	207	200
		25 %	trnp	-107	-48	101	264	335
			trp	97	98	102	101	100
reqm	Moyenne	Complet	trnp	643	614	546	536	566
			trp	553	547	545	587	616
		50 %	trnp	758	726	669	699	778
			trp	687	671	669	728	794
		25 %	trnp	949	898	863	952	1 062
			trp	895	859	863	955	1 041
	Total	Complet	trnp	537	376	543	1 183	1 485
			trp	553	547	545	587	616
		50 %	trnp	371	311	393	714	888
			trp	399	392	394	449	494
		25 %	trnp	255	233	282	451	553
			trp	285	273	283	328	365
Variance	Moyenne	Complet	trnp	15	14	14	20	26
			trp	15	14	14	18	22
		50 %	trnp	32	28	28	40	54
			trp	32	28	28	37	47
		25 %	trnp	64	57	59	83	107
			trp	64	58	59	76	93
	Total	Complet	trnp	11	11	14	28	43
			trp	15	14	14	18	22
		50 %	trnp	9	9	11	23	34
			trp	12	11	11	16	21
		25 %	trnp	5	5	7	14	20
			trp	7	7	7	10	12

3.2 Racine de l'erreur quadratique moyenne (reqm)

Malgré la petite taille de l'échantillon utilisé pour les simulations (312 avant la non-réponse) et le biais relatif plutôt modeste des estimations pour les moyennes, le biais demeure une composante importante de la reqm. Par exemple, le biais représente 56 % (sans pondération) à 69 % (avec pondération) de la reqm pour l'estimation de la moyenne selon la configuration $[CZ]^Y$ et $[C+Z]^R$ et le même taux d'échantillonnage que L et V. Lorsque l'échantillon est plus important, comme c'est généralement le cas pour les grandes enquêtes par sondage, le biais est souvent la composante dominante de la reqm (Brick 2013).

La figure 3.2 montre la reqm pour le total estimé (graphique de gauche) et pour la moyenne (graphique de droite) selon la même configuration que pour la figure précédente. La reqm pour le total pour l'estimateur pondéré est approximativement constante et inférieure à la reqm pour l'estimateur non pondéré, sauf lorsque le taux d'échantillonnage relatif est d'environ 0,5, ce qui correspond à la région où le biais est très faible pour l'estimateur non pondéré (voir la figure 3.1). Toutefois, lorsque le taux d'échantillonnage relatif est supérieur à un, la reqm pour l'estimateur non pondéré du total est beaucoup plus grande que la reqm pour l'estimateur pondéré (jusqu'à deux fois plus élevée pour certains taux d'échantillonnage). En revanche, pour les estimations de la moyenne illustrées à la figure 3.2 (graphique de droite), les reqm des estimateurs avec et sans pondération sont du même ordre de grandeur, et la symétrie autour du taux de répartition proportionnelle demeure. Même si L et V soulignent que l'estimateur non pondéré a une reqm inférieure (au taux d'échantillonnage relatif de 2,25), nous considérons les reqm des deux estimateurs comme étant approximativement égales pour tous les taux d'échantillonnage relatifs.

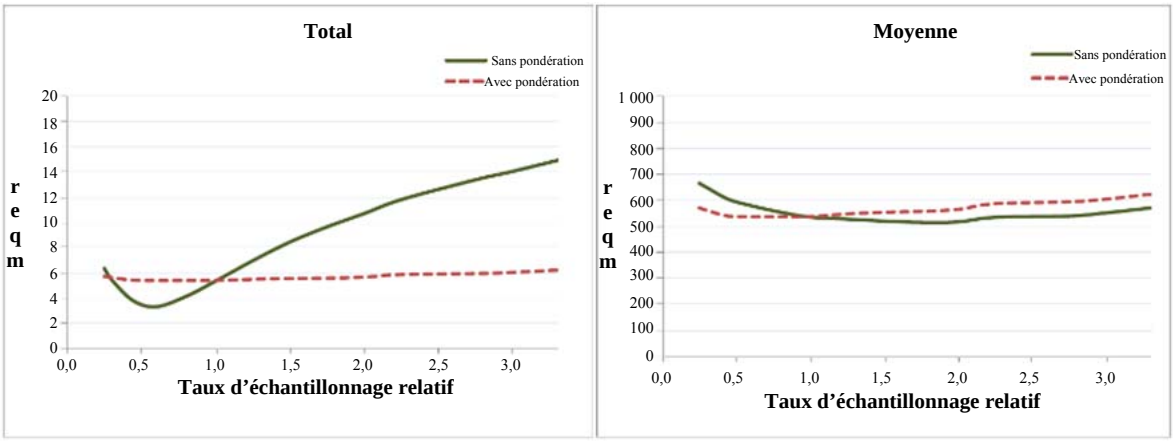

Figure 3.2 Racine de l'erreur quadratique moyenne pour les estimateurs avec et sans pondération pour $[CZ]^Y$ et $[C+Z]^R$; le graphique de gauche correspond au total (reqm en millions) et celui de droite, à la moyenne.

La figure 3.3 indique la reqm pour la moyenne estimée pour un domaine de 50 % (graphique de gauche) et un domaine de 25 % (graphique de droite), encore une fois pour $[CZ]^Y$ et $[C+Z]^R$. L'examen des trois graphiques de la reqm (pour la moyenne globale, la moyenne pour un domaine de 50 % et la moyenne pour un domaine de 25 %) révèle l'effet de l'estimateur par ratio. À mesure que la taille du domaine passe de

100 % à 25 %, l'estimateur pondéré ressemble de plus en plus à un estimateur par ratio inconditionnel et la corrélation entre le numérateur et le dénominateur réduit la reqm de l'estimation. En conséquence, les reqm des estimateurs de domaine avec et sans pondération sont très semblables. Même si l'estimateur pondéré est assorti d'une reqm inférieure à chacun des taux d'échantillonnage relatifs comparativement à l'estimateur non pondéré pour la moyenne pour un domaine de 25 %, les deux estimateurs sont essentiellement équivalents en termes de reqm. Le léger avantage de l'estimateur non pondéré qu'ont souligné L et V pour la moyenne pour l'ensemble de la population selon cette configuration disparaît pour les moyennes de domaine où l'estimateur pondéré est aussi un estimateur par ratio.

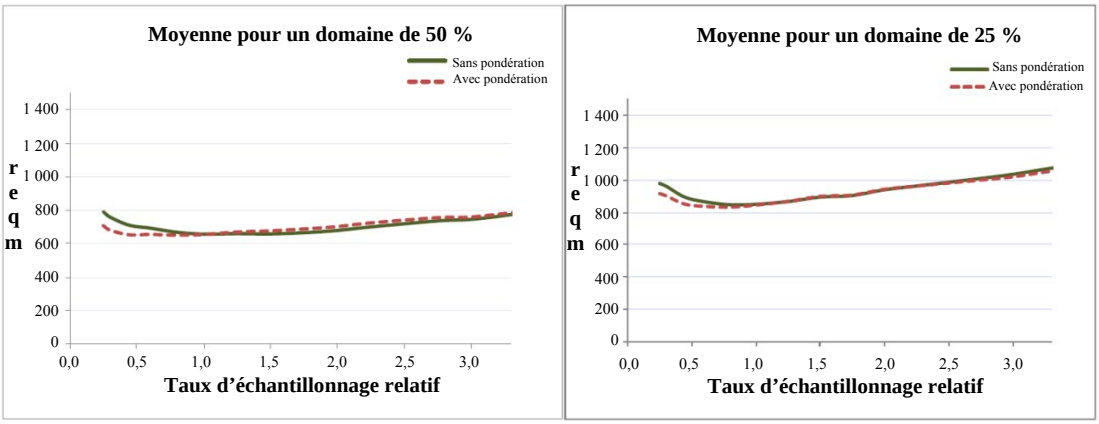

Figure 3.3 Racine de l'erreur quadratique moyenne pour les estimateurs avec et sans pondération pour $[CZ]^Y$ et $[C+Z]^R$; le graphique de gauche correspond à la moyenne pour un domaine de 50 % et celui de droite, à la moyenne pour un domaine de 25 %.

3.3 Variance

Quand les facteurs d'ajustement pour la non-réponse sont fondés sur un petit nombre de répondants, il est possible qu'ils accroissent la variance des estimations (Kalton 1983; Tremblay 1986). L et V sont d'avis que la pondération des facteurs d'ajustement pour la non-réponse pourrait entraîner une inflation de la variance supérieure à celle que l'on obtient lorsqu'on utilise des facteurs non pondérés. Les figures ci-dessus montrent que cela ne s'est pas produit dans le cadre de notre exercice de simulation. La figure 3.4 illustre le ratio de la variance de l'estimateur non pondéré à la variance de l'estimation pondérée pour la moyenne et le total pour l'ensemble de la population et pour le total du domaine de 50 % selon la configuration $[CZ]^Y$ et $[C+Z]^R$. Pour la moyenne, le ratio des variances est presque égal à un pour tous les taux d'échantillonnage relatifs; il n'y a pas d'inflation de la variance pour l'estimateur pondéré comparativement à l'estimateur non pondéré. En ce qui concerne les totaux, le ratio est inférieur à un pour les taux d'échantillonnage relatifs de moins de 1, et supérieur à un pour les taux d'échantillonnage relatifs de plus de 1. Cette relation se vérifie aussi pour le total du domaine de 50 %. Ces résultats semblent indiquer que la pondération de l'ajustement n'est pas une source de facteurs importants susceptibles de faire augmenter la variance des estimations. Par mesure de prudence, il convient d'examiner l'importance des facteurs de non-réponse, qu'ils soient ou non pondérés.

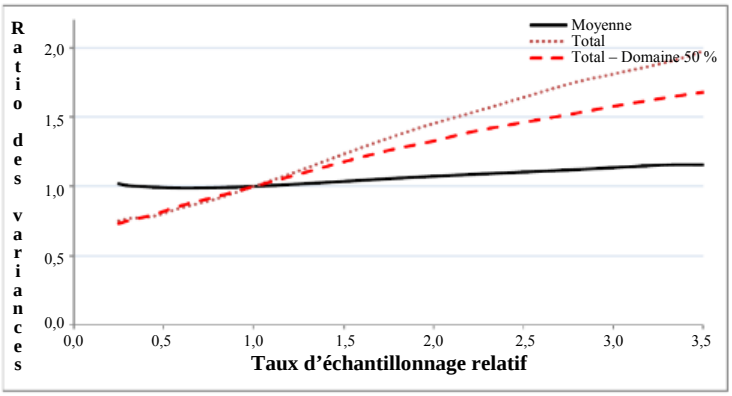

Figure 3.4 Ratio des variances des estimations non pondérées aux estimations pondérées de la moyenne, du total global et du total pour un domaine de 50 % selon $[CZ]^Y$ et $[C+Z]^R$.

Le tableau 3.2 présente les résultats de simulation pour une autre configuration, $[CZ]^Y$ et $[CZ]^R$, qui était favorable à l'ajustement non pondéré dans le cadre de l'étude de L et V (première ligne de leurs tableaux), alors que le tableau 3.3 présente les résultats de simulation pour la configuration $[C + Z]^Y$ et $[C + Z]^R$, qui était favorable à l'ajustement pondéré. Les résultats pour ces deux configurations montrent les mêmes tendances générales présentées ci-dessus pour $[CZ]^Y$ et $[C + Z]^R$.

Tableau 3.2
Biais (facteur 10 000), racine de l'erreur quadratique moyenne (facteur 10 000) et variance des estimateurs pondérés et non pondérés des moyennes et du total pour l'échantillon complet et pour les domaines, configuration $[CZ]^Y$, $[CZ]^R$ selon divers taux d'échantillonnage

				Taux d'échantillonnage relatif				
	Caractéristique	Domaine	Ajustement	0,30	0,44	1,00	2,25	3,30
Biais	Moyenne	Complet	trnp	329	329	289	255	237
			trp	294	299	289	298	298
		50 %	trnp	334	341	293	251	238
			trp	299	311	293	294	298
		25 %	trnp	336	344	306	257	247
			trp	302	314	306	299	307
	Total	Complet	trnp	-412	-187	287	732	901
			trp	294	299	289	298	298
		50 %	trnp	-209	-91	145	367	455
			trp	143	152	146	149	154
		25 %	trnp	-103	-46	72	184	230
			trp	74	76	73	75	79
reqm	Moyenne	Complet	trnp	530	507	476	501	533
			trp	505	487	476	520	554
		50 %	trnp	684	653	616	664	732
			trp	666	638	616	674	740
		25 %	trnp	911	859	832	920	1 016
			trp	900	849	832	920	1 011
	Total	Complet	trnp	550	395	474	886	1 078
			trp	505	487	476	520	554
		50 %	trnp	385	326	373	575	696
			trp	394	375	373	425	475
		25 %	trnp	263	244	278	390	464
			trp	285	274	278	321	361
Variance	Moyenne	Complet	trnp	17	15	14	19	23
			trp	17	15	14	18	22
		50 %	trnp	36	31	30	38	48
			trp	36	31	30	37	46
		25 %	trnp	73	63	61	79	98
			trp	73	63	61	76	94
	Total	Complet	trnp	14	12	14	25	35
			trp	17	15	14	18	22
		50 %	trnp	11	10	12	20	28
			trp	14	12	12	16	20
		25 %	trnp	6	6	7	12	16
			trp	8	7	7	10	13

Tableau 3.3
Biais (facteur 10 000), racine de l'erreur quadratique moyenne (facteur 10 000) et variance des estimateurs pondérés et non pondérés des moyennes et du total pour l'échantillon complet et pour les domaines, configuration $[C+Z]^Y$, $[C+Z]^R$ selon divers taux d'échantillonnage

				Taux d'échantillonnage relatif				
	Caractéristique	**Domaine**	**Ajustement**	**0,30**	**0,44**	**1,00**	**2,25**	**3,30**
Biais	Moyenne	Complet	trnp	763	735	654	566	529
			trp	665	661	654	654	652
		50 %	trnp	773	737	653	564	532
			trp	677	664	653	651	656
		25 %	trnp	773	739	659	574	513
			trp	679	668	659	660	636
	Total	Complet	trnp	-272	-8	651	1 411	1 744
			trp	665	661	654	654	652
		50 %	trnp	-133	-6	326	711	875
			trp	336	328	328	332	328
		25 %	trnp	-69	-2	157	359	438
			trp	165	166	158	168	165
reqm	Moyenne	Complet	trnp	854	818	745	699	711
			trp	767	753	745	764	790
		50 %	trnp	951	901	827	816	863
			trp	877	845	826	863	912
		25 %	trnp	1 101	1 046	981	1 023	1 098
			trp	1 044	1 004	981	1 045	1 107
	Total	Complet	trnp	426	313	741	1 503	1 868
			trp	767	753	745	764	790
		50 %	trnp	334	300	475	867	1 071
			trp	489	470	476	529	575
		25 %	trnp	246	240	314	530	649
			trp	320	316	314	372	409
Variance	Moyenne	Complet	trnp	15	13	13	17	23
			trp	15	13	13	16	20
		50 %	trnp	31	27	26	35	46
			trp	31	28	26	32	40
		25 %	trnp	62	56	54	73	95
			trp	63	57	54	67	83
	Total	Complet	trnp	11	10	13	27	45
			trp	15	13	13	16	20
		50 %	trnp	10	9	12	25	39
			trp	13	12	12	17	22
		25 %	trnp	6	6	7	15	23
			trp	8	7	8	11	14

3.4 Estimation de la taille de population

Sukasih et coll. (2009) ont étudié un type particulier d'estimation, soit l'estimation du nombre d'unités d'une population. On parle alors d'une estimation de la taille de population où la taille de population n'est qu'une estimation d'un total où $y_i = 1$ pour toutes les valeurs de i. Elle peut être estimée pour un domaine en affectant à toutes les unités en dehors du domaine la valeur $y_i = 0$. Dans le plan d'échantillonnage simple stratifié étudié ici, l'estimateur pondéré reproduit toujours la taille de population totale, $N = 10\,000$, mais pas l'estimateur non pondéré. Comme cette situation favorise clairement l'estimateur pondéré, nous examinons plutôt l'estimation de la taille de population d'un domaine.

Supposons que nous voulions estimer le nombre d'unités d'un domaine ou d'un sous-groupe qui ont une valeur en dessous d'un centile défini par une caractéristique pour la population totale (par exemple le revenu médian national). Ce type de statistique est extrêmement important dans les enquêtes, parce que les estimations de la taille de population pour les domaines sont souvent des statistiques clés. Ce type d'estimation peut être, par exemple, le nombre total de personnes ayant un revenu sous le seuil de pauvreté ou de faible revenu (Kovačević et Yung 1997).

Comme l'analyse de L et V ne tenait pas compte des estimations pour les tailles ou les moyennes de domaine, il n'existe pas de variable explicite qui pourrait servir à définir une sous-population. Pour ne pas

compliquer l'analyse, nous illustrons le rendement des deux estimateurs à l'aide d'un domaine artificiel créé par la sélection aléatoire de la moitié de la population (c'est-à-dire un domaine de 50 %). Selon une analyse semblable à celle dont il est question dans les sections précédentes, nous avons calculé les totaux et les moyennes pondérés et non pondérés pour le domaine de 50 %. Même si nous connaissons déjà la taille du domaine de l'exemple (c'est-à-dire 50 % de la population totale), l'analyse demeure valide. Dans la pratique, la taille du domaine n'est pas connue.

Quand on estime une statistique comme la taille de population d'un domaine, les deux estimateurs, pondéré et non pondéré, de la taille de population du domaine ne sont pas biaisés lorsque les données sont de type MCAR ou MAR, comme le soulignent Sukasih et coll. (2009). En outre, les reqm des estimateurs avec et sans pondération sont approximativement égales dans ce cas, comme le confirment les simulations.

Si les données ne sont pas de type MAR, la situation peut être très différente. L'estimateur pondéré d'une taille de population de domaine est à peu près non biaisé pour tous les taux d'échantillonnage relatifs et toutes les configurations, alors que l'estimateur non pondéré est toujours biaisé, sauf lorsqu'il est identique à l'estimateur pondéré (à un taux d'échantillonnage relatif de 1). En conséquence, la reqm de l'estimateur non pondéré pour la taille de domaine est souvent considérablement plus élevée que celle de l'estimateur pondéré. La figure 3.5 montre que la reqm de l'estimateur non pondéré de la taille de domaine de 50 % pour $[CZ]^Y$ et $[C + Z]^R$ est beaucoup plus grande que celle de l'estimateur pondéré pour la plupart des taux d'échantillonnage relatifs (jusqu'à deux fois la reqm de l'estimateur pondéré). La seule exception, c'est lorsque deux estimateurs sont à peu près égaux (répartition presque proportionnelle).

L'estimateur pondéré des tailles de domaine présente donc un avantage considérable par rapport à l'estimateur non pondéré pour tous les mécanismes de données manquantes présentés par L et V qui ne sont pas de type MCAR ou MAR.

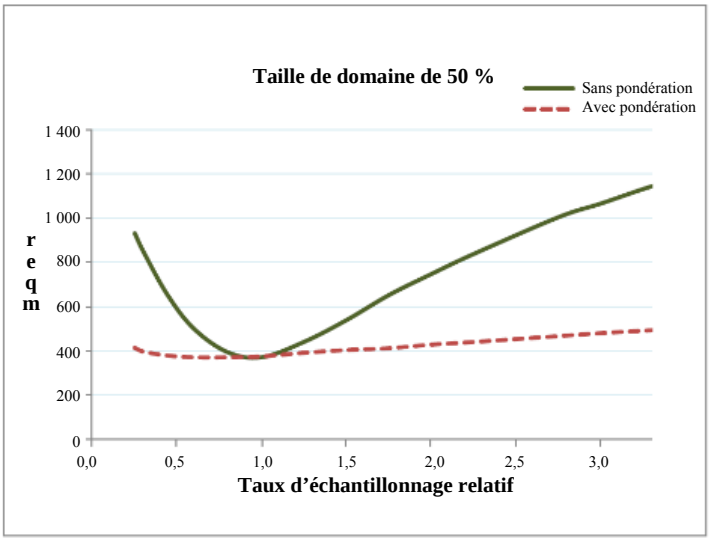

Figure 3.5 Racine de l'erreur quadratique moyenne (reqm) pour les estimateurs avec et sans pondération de la taille de domaine de 50 % pour $[CZ]^Y$ et $[C+Z]^R$.

4 Conclusions

Presque toutes les enquêtes sont touchées par la non-réponse; la méthode utilisée pour ajuster les poids de base pour la non-réponse totale est donc une question importante. L et V ont souligné à juste titre que l'utilisation de poids de sondage pour calculer un ajustement pondéré pour la non-réponse n'élimine pas le biais de non-réponse lorsque le mécanisme de réponse n'est pas spécifié correctement dans le modèle d'ajustement de la pondération. Toutefois, leur étude par simulation a porté au moins certains chercheurs à penser qu'un ajustement non pondéré pourrait mieux convenir qu'un ajustement pondéré dans la plupart des cas. Les résultats de notre évaluation, fondée sur le même scénario que celui de L et V, contredit cette perception. Nous avons examiné de façon plus approfondie les différences entre les estimateurs avec et sans pondération lorsque le modèle d'ajustement est inexact en utilisant le même scénario que L et V et en incluant différents taux d'échantillonnage et estimations des totaux et des domaines, en plus des moyennes étudiées par L et V.

Ces simulations élargies montrent que les ajustements avec et sans pondération ont effectivement des propriétés différentes. Le biais de l'estimateur pondéré des moyennes des totaux de plans d'échantillonnage aléatoires simples stratifiés est à peu près constant, quel que soit le taux d'échantillonnage, tandis que le biais de l'estimateur non pondéré dépend du taux d'échantillonnage. En revanche, le biais de l'estimateur non pondéré du total est considérablement plus important que celui de l'estimateur pondéré pour certains taux d'échantillonnage. Pour les moyennes, le biais et la reqm des deux estimateurs ne sont pas très différents, y compris pour les configurations que L et V ont décrites comme étant favorables à l'estimateur non pondéré. Les mêmes conclusions générales se vérifient pour les estimations des moyennes et des totaux de domaines à mesure que la moyenne pondérée se rapproche de plus en plus d'une estimation par ratio pour les domaines, ce qui finit par influencer quelque peu son comportement.

Nous avons aussi examiné l'estimation des tailles de domaine. Pour ce type de statistique, la reqm de l'estimateur pondéré est presque systématiquement inférieure à celle de l'estimateur non pondéré lorsque les données du scénario de simulation ne sont pas de type MAR. Les différences sont attribuables au biais de l'estimateur non pondéré de la taille de domaine; à cause de ce biais, l'estimateur non pondéré est assorti d'une reqm beaucoup plus grande que celle de l'estimateur pondéré pour certains taux d'échantillonnage.

Les modèles utilisés dans la plupart des enquêtes sont imparfaits; il est donc important de choisir la bonne méthode d'ajustement pour la non-réponse. Les résultats de la simulation élargie que nous présentons montrent que l'ajustement pondéré présente des avantages considérables pour certaines estimations et certains taux d'échantillonnage, comparativement à l'ajustement non pondéré. Plus particulièrement, une enquête selon le même plan qui produit des estimations des totaux et des statistiques autres que de simples moyennes semble bénéficier d'une pondération de l'ajustement. Bien sûr, la pondération de l'ajustement n'élimine pas le biais; elle diminue toutefois l'ampleur du biais dans bon nombre de situations et pour beaucoup des estimateurs que nous avons examinés. En outre, le biais de l'estimateur pondéré n'est pas sensible au taux d'échantillonnage relatif, alors que le biais de l'estimateur non pondéré l'est. L'inconvénient possible d'une augmentation de la variance de l'estimation lorsqu'on utilise un ajustement pondéré ne s'est pas concrétisé durant les simulations, et on peut l'éviter en examinant les facteurs d'ajustement, ce qu'il convient également de faire lorsqu'on utilise un ajustement non pondéré. Enfin, les

résultats de l'étude font ressortir le problème potentiel de la généralisation à partir de simulations. Même si les simulations constituent un outil précieux pour vérifier un point précis, une généralisation à plus grande échelle des résultats d'une simulation peut être trompeuse, particulièrement lorsque les résultats dépendent fortement des conditions du modèle utilisé pour la simulation.

Bibliographie

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics,* 29(2), 329-353.

Brick, J., et Jones, M. (2008). Propensity to respond and nonresponse bias. *Metron-International Journal of Statistics*, LXVI, 51-73.

Chadborn, T.R., Baster, K., Delpech, V., Sabin, C.A., Sinka, K., Rice, B.D. et Evans, B. (2005). No time to wait: How many HIV-infected homosexual men are diagnosed late and consequently die? (England and Wales, 1993-2002). *Aids*, 19(5), 513-520.

Grau, E., Potter, F., Williams, S. et Diaz-Tena, N. (2006). Nonresponse adjustment using logistic regression: To weight or not to weight? *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3073-3080.

Haukoos, J.S., et Newgard, C.D. (2007). Advanced statistics: Missing data in clinical research - part 1: An introduction and conceptual framework. *Academic Emergency Medicine*, 14(7), 662-668.

Kalton, G. (1983). *Introduction to Survey Sampling, SAGE University Paper 35.* Thousand Oaks, CA: SAGE Publications.

Kott, P. (2012). Pourquoi les poids de sondage devraient être intégrés dans la correction de la non-réponse totale fondée sur des groupes de réponse homogènes. *Techniques d'enquête*, 38, 1, 103-107.

Kovačević, M., et Yung, W. (1997). Estimation de la variance des mesures de l'inégalité et de la polarisation du revenu – Étude empirique. *Techniques d'enquête*, 23, 1, 47-59.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M. et Raghunathan, T.E. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response: Examples from multiple surveys. *Journal of the Royal Statistical Society Series A*, Royal Statistical Society, 173(2), 389-407.

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

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

Little, R., et Vartivarian, S. (2003). On weighting the rates in nonresponse weights. *Statistics in Medicine*, 22, 1589-1599.

R Development Core Team (2011). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. doi: http://www.R-project.org.

Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* Chichester, England : John Wiley & Sons, Inc.

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

Sukasih, A., Jang, D., Vartivarian, S., Cohen, S. et Zhang, F. (2009). A simulation study to compare weighting methods for nonresponses in the National Survey of Recent College Graduates. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Récupéré le 21 octobre 2013, à partir de www.amstat.org/sections/srms/proceedings/y2009/Files/304345.pdf.

Tremblay, V. (1986). Critères pratiques pour la définition des classes de pondération. *Techniques d'enquête,* 12, 1, 91-103.

West, B.T. (2009). A simulation study of alternative weighting class adjustments for nonresponse when estimating a population mean from complex sample survey data. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Récupéré le 21 octobre 2013, à partir de www.amstat.org/sections/srms/proceedings/y2009/Files/305394.pdf.

Wun, L.-M., Ezzati-Rice, T.M., Diaz-Tena, N. et Greenblatt, J. (2007). On modelling response propensity for dwelling unit (DU) level non-response adjustment in the Medical Expenditure Panel Survey (MEPS). *Statistics in Medicine*, 26(8), 1875-1884.