Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Traitements de la non-réponse de lien dans l'échantillonnage indirect

par Xiaojian Xu et Pierre Lavallée

Décembre 2009

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2009
Vol. 35, N° 2, pp. 165-177
Statistique Canada, N° 12-001-X au catalogue

Traitements de la non-réponse de lien dans l'échantillonnage indirect

Xiaojian Xu et Pierre Lavallée [1]

Résumé

Nous cherchons à corriger la surestimation causée par la non-réponse de lien dans l'échantillonnage indirect lorsque l'on utilise la méthode généralisée de partage des poids (MGPP). Nous avons élaboré quelques méthodes de correction pour tenir compte de la non-réponse de lien dans la MGPP applicables lorsque l'on dispose ou non de variables auxiliaires. Nous présentons une étude par simulation de certaines de ces méthodes de correction fondée sur des données d'enquête longitudinale. Les résultats des simulations révèlent que les corrections proposées de la MGPP réduisent bien le biais et la variance d'estimation. L'accroissement de la réduction du biais est significatif.

Mots-clés : Méthode de partage des poids ; non-réponse ; échantillonnage indirect ; enquête longitudinale.

1. Introduction

Par échantillonnage indirect, on entend la sélection d'échantillons dans une population qui n'est pas celle que l'on veut étudier, mais qui y est reliée. Un scénario d'échantillonnage de ce genre est souvent mis en œuvre lorsque l'on ne dispose pas de bases de sondage pour la population cible, mais que l'on en possède pour une autre population qui y est reliée. Nous appelons cette dernière la population d'échantillonnage. Par exemple, dans Lavallée (2007), nous considérons la situation où l'estimation a trait aux jeunes enfants dans les familles, mais où nous disposons seulement d'une liste de noms de parents comme base de sondage. Par conséquent, nous devons d'abord sélectionner un échantillon de parents avant de pouvoir sélectionner l'échantillon d'enfants. Dans cette situation type d'échantillonnage indirect, la population d'échantillonnage est celle des parents, tandis que la population cible est celle des enfants. Il convient de souligner que les enfants d'une famille particulière peuvent être sélectionnés par l'entremise de leur père ou de leur mère. La figure 1 donne une illustration simple de ce scénario d'échantillonnage indirect (figure 1.2, Lavallée 2007).

La littérature concernant les problèmes d'estimation associés à l'échantillonnage indirect est abondante et nous ne nommerons que quelques-uns de ces travaux ici. Ernst (1989) est le premier à discuter des méthodes appliquées pour produire des estimations transversales au moyen de données provenant d'une enquête-ménages longitudinale. Il présente la méthode de partage des poids dans le contexte d'une enquête longitudinale et montre aussi qu'elle fournit un estimateur sans biais du total pour toute caractéristique de la population d'intérêt. Kalton et Brick (1995) concluent qu'une telle méthode fournit aussi la variance minimale du total de population estimé sous certains scénarios d'échantillonnage simples pour une enquête-ménages longitudinale par panel. Lavallée (1995) étend la méthode de partage des poids à un contexte tout à fait général d'échantillonnage indirect qui comprend l'enquête longitudinale comme cas particulier, pour obtenir la méthode généralisée de partage des poids (MGPP). Il montre que ce scénario de pondération produit des estimations sans biais quel que soit le plan d'échantillonnage utilisé pour obtenir un échantillon de la population d'échantillonnage. Comme cela est le cas de tout scénario de pondération, la mise en œuvre de la MGPP requiert une correction pour divers problèmes de non-réponse. Lavallée (2001) fournit une MGPP corrigée tenant compte des problèmes de non-réponse dans l'échantillonnage indirect. Ce dernier comporte un autre type de non-réponse, dénommée non-réponse de lien (*link nonresponse*) et que Lavallée (2001) avait appelée « non-réponse de relation (*relationship nonresponse*) », dû au fait qu'il est impossible de déterminer, ou que l'on n'a pas réussi à déterminer, si une unité de la population d'échantillonnage est apparentée (liée) à une unité dans la population cible. Lavallée (2001) souligne qu'en cas de non-réponse de lien, la MGPP donne lieu à une surestimation, mais laisse ouverte la question de la correction qu'il convient d'apporter pour tenir compte de la non-réponse de lien. L'objectif de la présente étude est d'élaborer des méthodes pour traiter le biais d'estimation causé par cette non-réponse de lien.

La présentation de la suite de l'article est la suivante. À la section 2, nous décrivons la notation et le problème étudié. À la section 3, nous proposons quelques modifications de la MGPP en vue d'intégrer la non-réponse de lien. À la section 4, nous décrivons une étude par simulation portant sur un ensemble de données réelles et à la section 5, nous présentons certaines conclusions. Il convient de souligner que nous nous servons dans le présent article d'une étude par simulation pour illustrer les progrès réalisés grâce aux nouvelles méthodes, et que d'autres contributions théoriques à la résolution du problème peuvent être consultées dans

1. Xiaojian Xu, Département de mathématique, Université Brock, St. Catharines (Ontario) Canada, L2S 3A1. Courriel : xxu@brocku.ca ; Pierre Lavallée, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario), K1A 0T6. Courriel : Pierre.lavallée@statcan.gc.ca.

Lavallée (2002), Deville et Lavallée (2006), et Lavallée (2007).

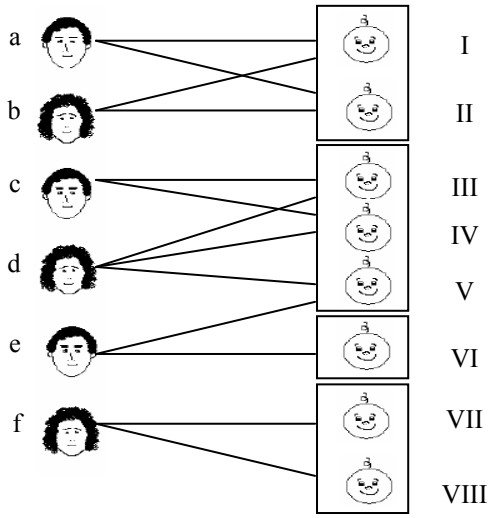

Figure 1 Échantillonnage indirect d'enfants

2. Notation et problème

Soit U^A et U^B la population d'échantillonnage et la population cible, respectivement. Donc, U^A est la population reliée à U^B pour laquelle existe une base de sondage connue. Soit s^A, M^A et m^A, un échantillon tiré de U^A, le nombre d'unités dans U^A et le nombre d'unités dans s^A, respectivement. Soit π_j^A la probabilité de sélection de la j^{e} unité dans U^A, avec $\pi_j^A > 0$ et $\sum_{j=1}^{M^A} \pi_j^A = m^A$. Nous utilisons également la notation suivante : M^B, N, U_i^B et M_i^B, le nombre d'unités dans U^B, le nombre de grappes dans U^B, la i^{e} grappe de U^B avec $\cup_{i=1}^{N} U_i^B = U^B$, et le nombre d'unités dans la i^{e} grappe U_i^B, respectivement.

Soit $l_{j,ik}$, une variable indicatrice de l'existence d'un lien : $l_{j,ik} = 1$ indique qu'il existe un lien entre la j^{e} unité de U^A et la k^{e} unité de U_i^B, tandis que $l_{j,ik} = 0$ indique qu'il n'existe pas de lien. Soit aussi $L_{j,i}^B$, le nombre total de liens existants entre l'unité j de U^A et les unités de U_i^B, c'est-à-dire $L_{j,i}^B = \sum_{k=1}^{M_i^B} l_{j,ik}$. Soit L_i^B, le nombre total de liens existants entre les unités de U^A et les unités de U_i^B, c'est-à-dire $L_i^B = \sum_{j=1}^{M^A} L_{j,i}^B$. Nous désignons par y_{ik} les caractéristiques de la k^{e} unité de la i^{e} grappe dans la population U^B, et par Y^B le total de tous les y_{ik}'. Nous avons alors $Y^B = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} y_{ik}$.

Désignons par Ω^B les grappes dans U^B où il existe au moins une unité ik telle que $l_{j,ik} = 1$ pour une j^{e} unité dans s^A, et disons qu'elle peut être identifiée par les unités j dans s^A, c'est-à-dire qu'une telle unité i satisfait $L_i^B = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} l_{j,ik} > 0$. Le nombre de grappes dans Ω^B est n. Après l'échantillonnage, nous avons réétiqueté les grappes comprises dans Ω^B au moyen de l'indice $i = 1, 2, \ldots, n$. Nous désignons par w_{ik} le poids d'estimation appliqué à la k^{e} unité de la i^{e} grappe, par Ω_i^A l'ensemble d'unités comprises dans U^A qui ont des liens avec certaines unités comprises dans U_i^B avec $i \in \Omega^B$, et par Ω^A l'ensemble d'unités comprises dans U^A qui ont des liens avec certaines unités comprise dans Ω^B, c'est-à-dire $\Omega^A = \{j | \sum_{i \in \Omega^B} L_{j,i}^B \neq 0\}$. Nous utilisons s_i^A pour indiquer l'ensemble d'unités dans s^A qui ont des liens avec certaines unités de U_i^B avec $i \in \Omega^B$. Soit T^A, T_i^A et m_i^A, le nombre d'unités dans Ω^A, le nombre d'unités dans Ω_i^A et le nombre d'unités dans s_i^A, respectivement. Enfin, nous utilisons les trois indicateurs suivants : soit t_j la variable indicatrice de sélection dans $s^A : t_j = 1$ indique que la j^{e} unité de U^A est dans s^A et $t_j = 0$ indique le contraire. Soit t_j^L la variable indicatrice d'inclusion dans s^A pour les unités de $\Omega^A : t_j^L = 1$ indique que la j^{e} unité de Ω^A est dans s^A et $t_j^L = 0$ indique le contraire ; soit, enfin, $t_{j,i}^L$ la variable indicatrice d'inclusion dans s_i^A pour les unités de $\Omega_i^A : t_{j,i}^L = 1$ indique que la j^{e} unité de Ω_i^A est dans s_i^A et $t_{j,i}^L = 1$ indique le contraire.

Notre objectif est d'estimer le total Y^B, qui est notre paramètre d'intérêt, pour la population cible U^B qui est divisée en N grappes. Pour le faire, nous sélectionnons un échantillon s^A dans U^A avec la probabilité de sélection π_j^A. Puis, nous identifions Ω^B en utilisant $l_{j,ik} \neq 0$. Toutes les unités des grappes comprises dans Ω^B font partie de l'étude dans laquelle nous mesurons y_{ik} et l'ensemble des $l_{j,ik}$.

En appliquant la MGPP, nous attribuerons un poids d'estimation w_{ik} à chaque unité k des grappes étudiées i. Ces poids peuvent être choisis de manière appropriée pour que l'estimateur de Y^B :

$$\hat{Y}^B = \sum_{i=1}^{n} \sum_{k=1}^{M_i^B} w_{ik} y_{ik} \tag{1}$$

donne de bons résultats pour l'estimation de Y^B.

Nous souhaitons estimer la quantité Y^B en utilisant $\hat{Y}^B$. En suivant Horvitz et Thompson (1952), posons que w_{ik} est égal à l'inverse de la probabilité de sélection, π_{ik}, de la k^{e} unité de U_i^B dans la population cible. Alors, $\hat{Y}^B$ donne un estimateur sans biais de Y^B. Cependant, le calcul de π_{ik} est difficile, voire même impossible, dans le cas qui nous occupe, à cause de la complexité du scénario d'échantillonnage indirect. Par conséquent, nous introduisons la MGPP pour résoudre ce problème. Pour la commodité du lecteur, nous donnons ici les grandes lignes de la MGPP en calculant les poids pour chaque grappe qui a été observée.

Étape 1 : Fournir les poids initiaux w_{ik}'

$$w_{ik}' = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A}; \tag{2}$$

Étape 2 : Calculer L_i^B

$$L_i^B = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} l_{j,ik} \; ; \tag{3}$$

Étape 3 : Obtenir le poids final w_i

$$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{L_i^B} \; ; \tag{4}$$

Étape 4 : Poser que $w_{ik} = w_i$ pour tout k dans la i^{e} grappe.

Il découle du théorème de la section 3 de Lavallée (2001) que

$$\hat{Y}^B = \sum_{i=1}^{n} \frac{\sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{L_i^B} \sum_{k=1}^{M_i^B} y_{ik} \tag{5}$$

offre un estimateur sans biais de Y^B à condition que tous les liens $l_{j,ik}$ puissent être identifiés correctement. Les poids d'estimation attribués dans (5) sont

$$w_{ik} = \begin{cases} \dfrac{\sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{L_i^B}, & \text{pour toutes les unités } k \text{ dans la grappe } i \text{ quand } i \text{ est dans } \Omega^B \; ; \\ 0, & \text{quand } i \text{ n'est pas dans } \Omega^B . \end{cases} \tag{6}$$

Un exemple simple est illustré à la figure 2. Nous voulons estimer le total Y^B lié à la population cible U^B. Supposons que nous sélectionnons les unités $j = 1$ et 2 dans U^A. En sélectionnant l'unité $j = 1$, nous étudions les unités de la grappe $i = 1$. De même, en sélectionnant l'unité $j = 2$, nous étudions les unités des grappes $i = 1$, et 2. Par conséquent, nous avons $\Omega^B = \{1, 2\}$. Pour chaque unité k des grappes i de Ω^B, nous calculons les poids initiaux w'_{ik} donnés par (2), le nombre total de liens qui existent entre les unités de U^A et les unités de U_i^B, L_i^B, et les poids finaux w_{ik}. Alors, d'après (5), l'estimateur résultant de Y^B prend la forme qui suit (voir Lavallée 2007, pages 17 et 18 pour plus de précisions) :

$$\hat{Y}^B = \frac{1}{2}\left[\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A}\right] y_{11} + \frac{1}{2}\left[\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A}\right] y_{12} + \frac{1}{3\pi_2^A} y_{21} + \frac{1}{3\pi_2^A} y_{22} + \frac{1}{3\pi_2^A} y_{23}. \tag{7}$$

Soulignons que, dans le cas de l'estimateur pour lequel $l_{j,ik}$ est connu, la seule hypothèse pour l'absence de biais est que $L_i^B > 0$ pour toutes les grappes i dans U^B. Autrement dit, chaque grappe de la population cible doit posséder au moins un lien provenant de U^A. Nous savons que si certains liens manquent, l'estimateur (5) sera biaisé. En cas de non-réponse de lien, comme il est mentionné dans Lavallée (2001), on ne peut déterminer L_i^B. Habituellement, l'utilisation du nombre total de liens observés pour remplacer cette quantité inconnue produit une surestimation de Y^B parce que certaines composantes des liens manquent réellement dans la sommation L_i^B. Ce problème est exactement celui que nous nous proposons d'examiner dans la présente étude et nous essayons de corriger les poids d'estimation w_{ik} en estimant L_i^B de façon à obtenir une meilleure estimation de Y^B.

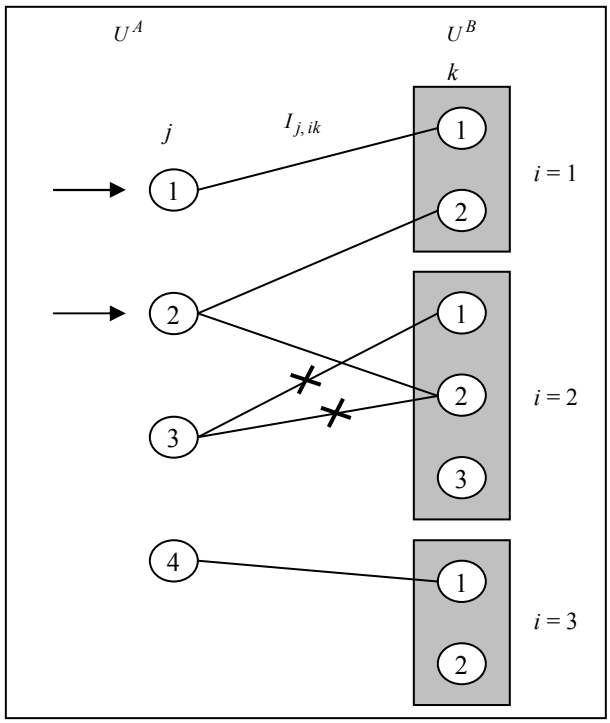

Figure 2 Exemple de liens dans l'échantillonnage indirect

3. Traitements du biais dans les estimations

Comme nous l'avons mentionné à la section 1, le biais observé dans les estimations lorsque l'on utilise la MGPP est causé par des problèmes de non-réponse de lien. Dans cette situation, il est impossible d'identifier et d'observer tous les éléments composant L_i^B. En pratique, bien qu'on puisse normalement déterminer les liens entre les unités de s^A et celles de U^B, il est souvent difficile, voire même impossible, d'identifier les parties des liens en dehors de s^A. Nous disons que de telles unités ont des liens

manquants avec U^B. Soit $\Delta^A = \Omega^A \setminus s^A$ l'ensemble d'unités pour lesquelles des liens pourraient manquer. Alors,

$$L_i^B = \sum_{j \in s^A} \sum_{k=1}^{M_i^B} l_{j,ik} + \sum_{j \in \Delta^A} \sum_{k=1}^{M_i^B} l_{j,ik}. \quad (8)$$

Si nous exécutons la MGPP sans tenir compte de ces liens manquants, nous nous servons du total des $l_{j,ik}$ observés comme valeur de L_i^{B*} pour calculer $\hat{Y}^B$ en utilisant

$$L_i^{B*} = \sum_{j \in s^A} \sum_{k=1}^{M_i^B} l_{j,ik} + \sum_{j \in \Delta_0^A} \sum_{k=1}^{M_i^B} l_{j,ik}, \quad (9)$$

où Δ_0^A est un sous-ensemble de Δ^A qui contient uniquement les unités dont les liens sont observés. Le prix de cette approche est la surestimation de Y^B en utilisant (5), puisque

$$L_i^B \geq L_i^{B*}.$$

Nous proposons quelques méthodes pour appliquer la MGPP en tenant compte de la non-réponse de lien dans l'estimation de L_i^B.

3.1 Estimation de L_i^B en l'absence de variables auxiliaires

3.1.1 Estimation de L_i^B par ajustement proportionnel pour chaque grappe (méthode 1)

Pour aborder le problème de la non-réponse de lien, nous nous concentrons sur l'estimation de L_i^B en utilisant l'information connue sur les liens à l'intérieur de s^A. Pour calculer les poids donnés par (6) en utilisant la MGPP, il nous suffit d'estimer L_i^B pour les $i \in \Omega^B$. Pour tout $i \in \Omega^B$,

$$L_i^B = \sum_{j=1}^{T_i^A} L_{j,i}^B. \quad (10)$$

Un estimateur général de ce total peut s'exprimer sous la forme

$$\hat{L}_i^B = \sum_{j=1}^{T_i^A} w_{j,i}^L L_{j,i}^B, \quad (11)$$

où $w_{j,i}^L$ est un poids aléatoire qui prend la valeur $w_{j,i}^L = 0$ si j n'est pas dans l'échantillon s_i^A. Pour chaque $i \in \Omega^B$, nous utilisons l'information connue sur les liens entre s_i^A et U_i^B pour estimer l'information sur les liens entre Ω_i^A et U_i^B. L'espérance de $\hat{L}_i^B$ est

$$E(\hat{L}_i^B) = \sum_{j=1}^{T_i^A} E(w_{j,i}^L) L_{j,i}^B. \quad (12)$$

En comparant (10) et (12), nous constatons que $\hat{L}_i^B$ est sans biais pour L_i^B sous tout scénario de pondération avec $E(w_{j,i}^L) = 1$ pour tout j.

Pour commencer, nous adoptons l'estimateur d'Horvitz-Thompson (Horvitz et Thompson 1952), également appelé estimateur π (Särndal, Swensson et Wretman 1991). Notons qu'en vertu de la définition de Ω_i^A, $\Omega_i^A \supset s_i^A$ pour tout i. Nous *imitons* une procédure d'estimation du nombre de liens dans Ω_i^A en utilisant celui dans s_i^A. La procédure consiste à sélectionner un « échantillon » s_i^A dans la « population » Ω_i^A. Soit $\pi_{j,i}^L$ la probabilité que j (qui est dans Ω_i^A) soit incluse dans s_i^A. Posons alors que

$$w_{j,i}^L = \begin{cases} 1/\pi_{j,i}^L, & j \text{ est dans } s_i^A, \\ 0, & j \text{ est dans } \Omega_i^A \setminus s_i^A. \end{cases} \quad (13)$$

Selon le corollaire 3.1 dans Cassel, Särndal et Wretman (1977), ce scénario de pondération donne un estimateur sans biais de L_i^B. Nous avons

$$\hat{L}_i^B = \sum_{j=1}^{T_i^A} \frac{L_{j,i}^B t_j^L}{\pi_{j,i}^L}. \quad (14)$$

Cela nous donne un estimateur asymptotiquement sans biais (la preuve suit) de Y^B :

$$\tilde{Y}^B = \sum_{i=1}^{n} \frac{\sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{\sum_{j=1}^{T_i^A} \frac{L_{j,i}^B t_{j,i}^L}{\pi_{j,i}^L}} \sum_{k=1}^{M_i^B} y_{ik}. \quad (15)$$

Afin de démontrer l'absence de biais de cet estimateur, nous employons un développement en série de Taylor. Selon le corollaire 5.1.5 (Fuller 1996), nous obtenons

$$\frac{1}{\hat{L}_i^B} = \frac{1}{L_i^B} - \frac{1}{(L_i^B)^2}(\hat{L}_i^B - L_i^B) + O([\hat{L}_i^B - L_i^B]^2)$$

$$= \frac{1}{(L_i^B)^2}(2L_i^B - \hat{L}_i^B) + O_p(n^{-1}).$$

Il s'ensuit que

$$p\lim\left\{ n^{1/2}\left[\frac{1}{\hat{L}_i^B} - \frac{1}{(L_i^B)^2}(2L_i^B - \hat{L}_i^B) \right] \right\} = 0.$$

Par conséquent, en vertu du théorème 5.2.1 (Fuller 1996), la loi limite de $n^{1/2}[1/\hat{L}_i^B]$ est la loi limite de $n^{1/2}[1/(L_i^B)^2(2L_i^B - \hat{L}_i^B)]$. Nous notons que $\tilde{Y}^B$ est une fonction de la variable aléatoire t_j ainsi que de la variable aléatoire $t_{j,i}^L$; par conséquent, nous désignons l'espérance de $\tilde{Y}^B$ par rapport à t_j par $E_{t_j}(\cdot)$ et celle par rapport à $t_{j,i}^L$ par $E_{t_{j,i}^L}(\cdot)$. D'où nous obtenons asymptotiquement

$$
\begin{aligned}
&E(\tilde{Y}^B) \\
&\approx \sum_{i=1}^{n} E_{t_j}\left[E_{t^L_{j,i}}\left(\frac{1}{(L_i^B)^2}\left(2L_i^B - \sum_{j=1}^{T_i^A} \frac{L_{j,i}^B t_{j,i}^L}{\pi_{j,i}^L} \right) \right.\right. \\
&\qquad \left.\left. \sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A} \right) \middle| \Omega^B \right] \sum_{k=1}^{M_i^B} y_{ik} \\
&= \sum_{i=1}^{n} E_{t_j}\left(\frac{1}{L_i^B} \sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A} \right) \sum_{k=1}^{M_i^B} y_{ik} \qquad (16) \\
&= E_{t_j}\left(\sum_{i=1}^{n} \left(\frac{1}{L_i^B} \sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A} \right) \sum_{k=1}^{M_i^B} y_{ik} \right) \qquad (17) \\
&= E_{t_j}(\hat{Y}^B).
\end{aligned}
$$

Selon Lavallée (1995), $E_{t_j}(\hat{Y}^B) = Y^B$. Par conséquent, $\tilde{Y}^B$ est un estimateur approximativement sans biais de Y^B.

Maintenant, nous devons calculer $\pi_{j,i}^L$. Il s'agit d'une fonction de π_j^A, mais pourtant elle dépend de la façon dont s_i^A influe sur U_i^B, donc sur Ω_i^A. Ce genre d'effet est difficile à dépister et varie d'un cas à l'autre ; cependant, nous pouvons en donner une estimation générale. La première approche que nous proposons ici est d'estimer la probabilité de sélection, $\pi_{j,i}^L$, en utilisant la proportion d'unités de s^A qui prennent dans Ω^A, c'est-à-dire

$$\hat{\pi}_{j,i}^{L(1)} = \frac{m_i^A}{T_i^A}. \qquad (18)$$

Donc,

$$
\begin{aligned}
\hat{L}_i^{B(1)} &= \sum_{j=1}^{T_i^A} \frac{L_{j,i}^B t_j^L}{\hat{\pi}_{j,i}^{L(1)}} \\
&= \frac{T_i^A}{m_i^A} \sum_{j=1}^{m_i^A} L_{j,i}^B. \qquad (19)
\end{aligned}
$$

et

$$\hat{Y}^{B(1)} = \sum_{i=1}^{n} \frac{\sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{\frac{T_i^A}{m_i^A} \sum_{j=1}^{m_i^A} L_{j,i}^B} \sum_{k=1}^{M_i^B} y_{ik} = \sum_{i=1}^{n} w_i^{(1)} \sum_{k=1}^{M_i^B} y_{ik}, \qquad (20)$$

avec

$$w_i^{(1)} = \frac{m_i^A}{T_i^A} \frac{\sum_{j=1}^{m^A} \frac{L_{j,i}^B}{\pi_j^A}}{\sum_{j=1}^{m_i^A} L_{j,i}^B}. \qquad (21)$$

Revenons à l'exemple de la figure 2, en supposant qu'il existe deux non-réponses de lien qui se produisent entre l'unité $j=3$ dans U^A et les unités $k=1,2$ de la grappe $i=2$ dans U^B. Si nous utilisons la MGPP sans correction donnée par (5), l'estimateur résultant de Y^B n'est plus (7). À la place, nous avons

$$
\begin{aligned}
\hat{Y}^B = &\frac{1}{2}\left(\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A} \right) y_{11} + \frac{1}{2}\left(\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A} \right) y_{12} \\
&+ \frac{1}{\pi_2^A} y_{21} + \frac{1}{\pi_2^A} y_{22} + \frac{1}{\pi_2^A} y_{23}, \qquad (22)
\end{aligned}
$$

qui comporte un biais. Afin d'appliquer (20), nous commençons par calculer m_i^A / T_i^A. Les poids résultant de l'application de la méthode (1) donnés par (21) pour cet exemple sont présentés au tableau 1. Donc, cette méthode modifiée donne l'estimateur :

$$
\begin{aligned}
\hat{Y}^B = &\frac{1}{2}\left(\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A} \right) y_{11} + \frac{1}{2}\left(\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A} \right) y_{12} \\
&+ \frac{1}{2\pi_2^A} y_{21} + \frac{1}{2\pi_2^A} y_{22} + \frac{1}{2\pi_2^A} y_{23}, \qquad (23)
\end{aligned}
$$

qui est moins biaisé que (22).

Tableau 1
Poids initiaux, nombre total de liens qui ont répondu et poids finaux selon (21)

i	k	w'_{ik}	L_i^B	m_i^A	T_i^A	m_i^A/T_i^A	$w_i^{(1)}$
1	1	$1/\pi_1^A$	1	2	2	1	$1/2(1/\pi_1^A + 1/\pi_2^A)$
1	2	$1/\pi_2^A$	1	2	2	1	$1/2(1/\pi_1^A + 1/\pi_2^A)$
2	1	0	0 (manquant)	1	2	1/2	$1/2\pi_2^A$
2	2	$1/\pi_2^A$	1 (un lien manque)	1	2	1/2	$1/2\pi_2^A$
2	3	0	0	1	2	1/2	$1/2\pi_2^A$

3.1.2 Estimation de L_i^B par ajustement proportionnel global (méthode 2)

Dans l'approche précédente, l'information concernant m_i^A et T_i^A est nécessaire pour chaque grappe i. En supposant que nous ignorons la variation de Ω_i^A parmi toutes les grappes i, nous proposons simplement d'utiliser

$$L_i^{B*} = \sum_{j=1}^{T^A} \frac{L_{j,i}^B t_j^L}{\pi_j^L} \tag{24}$$

en nous servant de l'information sur les liens dans s^A pour estimer l'information sur les liens dans T^A, où t_j^L est la variable indicatrice de la présence dans s^A en provenance de Ω^A. Nous devons maintenant calculer π_j^L. De nouveau, il s'agit d'une fonction de π_j^A qui dépend pourtant de la complexité des effets de s^A sur Ω^B, donc sur Ω^A. Le calcul est compliqué et varie d'un cas à l'autre sans avoir de forme générale, mais nous pouvons habituellement donner une estimation grossière de la valeur.

La deuxième approche que nous proposons ici consiste à estimer π_j^L en utilisant la proportion d'unités dans s^A qui apparaît dans Ω^A, c'est-à-dire $\pi_j^{L*} = m^A / T^A$. Cela nous informe que

$$\hat{L}_i^{B(2)} = \frac{T^A}{m^A} \sum_{j=1}^{m^A} L_{j,i}^B. \tag{25}$$

Dans le cas de plans de sondage aléatoires simples avec ou sans stratification, $\hat{L}_j^{B(2)}$ donne un estimateur sans biais de L_i^B. Pour les plans plus complexes, il fournit un estimateur fondé sur un modèle sans biais sous l'hypothèse (A) de la façon suivante :

(A) Supposons que pour toute grappe *i*, la moyenne de tous les liens existants associés avec toutes les unités comprises dans l'échantillon s^A est la même que celle des liens existants associés à toutes les unités comprises dans U^A, c'est-à-dire

$$\frac{\sum_{j=1}^{m^A} L_{j,i}^B}{m^A} = \frac{\sum_{j=1}^{M^A} L_{j,i}^B}{T^A}. \tag{26}$$

Donc, les poids d'estimation sont donnés par

$$w_{ik}^{(2)} = w_i^{(2)} = \frac{m^A}{T^A} \frac{\sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{\sum_{j=1}^{M^A} L_{j,i}^B t_j}, \text{ pour toutes les unités } k \tag{27}$$

dans la grappe *i*.

Il s'ensuit que Y^B peut être estimé par

$$\hat{Y}^{B(2)} = \frac{m^A}{T^A} \sum_{i=1}^{n} \frac{\sum_{j=1}^{m^A} \frac{L_{j,i}^B}{\pi_j^A}}{\sum_{j=1}^{m^A} L_{j,i}^B} \sum_{k=1}^{M_i^B} y_{ik} = \sum_{i=1}^{n} w_i^{(2)} \sum_{k=1}^{M_i^B} y_{ik}, \tag{28}$$

Revenons à l'exemple de la figure 2 avec deux non-réponses de lien qui surviennent entre l'unité $j = 3$ dans U^A et les unités $k = 1, 2$ de la grappe $i = 2$ dans U^B. Afin d'appliquer (28), nous calculons d'abord m^A / T^A. Pour notre exemple, nous avons $m^A = 2$ et $T^A = 3$. L'estimateur résultant de Y^B en utilisant la méthode d'ajustement (2) pour cet exemple est alors

$$\hat{Y}^B = \frac{2}{3}\left[\frac{1}{2}\left(\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A}\right) y_{11} + \frac{1}{2}\left(\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A}\right) y_{12} + \frac{1}{\pi_2^A} y_{21} + \frac{1}{\pi_2^A} y_{22} + \frac{1}{\pi_2^A} y_{23}\right]. \tag{29}$$

Par conséquent, la correction faite dans (28) est différente de celle résultant de la méthode (1) dans cet exemple.

Nous savons que $\text{var}(\hat{Y}^{B(1 \text{ ou } 2)}) = \text{var}\{E(\hat{Y}^{B(1 \text{ ou } 2)} \mid s^A)\} + E\{\text{var}(\hat{Y}^{B(1 \text{ ou } 2)} \mid s^A)\}$. L'espérance et la variance intérieures (conditionnellement à s^A) sont calculées sur tous les ensembles possibles de liens $l_{j,ik}$ « répondants » sachant l'échantillon s^A, tandis que l'espérance et la variance extérieures sont calculées sur tous les échantillons s^A possibles. En général, les corrections faites ci-dessus n'éliminent pas le deuxième terme qui dépend du caractère aléatoire de $l_{j,ik}$.

3.2 Estimation de L_i^B en disposant de variables auxiliaires

3.2.1 Estimation de $l_{j,ik}$ en utilisant un modèle logistique

Les méthodes d'estimation de L_i^B proposées à la section 3.1 sont faciles à appliquer et ne nécessitent pas d'information supplémentaire. Toutefois, les hypothèses peuvent être violées, ce qui produit une estimation indésirable. Par exemple, $L_{j,i}^B$ peut dépendre de certaines caractéristiques de l'unité *j* et de la grappe *i*.

Nous supposons que la probabilité qu'il existe un lien entre une unité dans la population d'échantillonnage et une unité dans la population cible dépend de certaines variables auxiliaires en spécifiant un modèle de régression logistique. Nous pouvons estimer cette fonction de probabilité de façon que l'estimation de la quantité d'intérêt dans la population cible soit désirable. Soit $P_{j,ik} = P(l_{j,ik} = 1)$ qui est affectée par un certain vecteur de variables $\mathbf{x}_j^A$ dans U^A et $\mathbf{x}_{ik}^B$ dans U^B.

Nous pouvons ajuster le modèle logistique

$$\log\left(\frac{P_{j,ik}}{1 - P_{j,ik}}\right) = \mathbf{a}'\mathbf{x}_j^A + \mathbf{b}'\mathbf{x}_{ik}^B \tag{30}$$

en utilisant les liens observés et leurs variables caractéristiques correspondantes. Les vecteurs de paramètres inconnus **a** et **b** peuvent être estimés. Nous proposons d'imputer les liens $l'_{j,ik}$ par leur probabilité estimée :

$$\hat{P}_{j,ik} = \frac{e^{\hat{\mathbf{a}}'\mathbf{x}_j^A + \hat{\mathbf{b}}'\mathbf{x}_{ik}^B}}{1 + e^{\hat{\mathbf{a}}'\mathbf{x}_j^A + \hat{\mathbf{b}}'\mathbf{x}_{ik}^B}}, \tag{31}$$

où $(\hat{\mathbf{a}}, \hat{\mathbf{b}})$ est un estimateur de $(\mathbf{a}, \mathbf{b})$; par exemple, nous utilisons l'estimateur du maximum de vraisemblance (pseudo-vraisemblance) pondéré. Nous obtenons alors

$$\hat{L}_i^{B(3)} = \sum_{j \in s^A \cup \Delta_0^A} L_{j,i} + \sum_{j \in \Omega^A \setminus (s^A \cup \Delta_0^A)} \hat{L}_{j,i}$$

$$= \sum_{j \in s^A \cup \Delta_0^A} L_{j,i} + \sum_{j \in \Omega^A \setminus (s^A \cup \Delta_0^A)} \sum_{k=1}^{M_i^B} \frac{e^{\hat{\mathbf{a}}'\mathbf{x}_j^A + \hat{\mathbf{b}}'\mathbf{x}_{ik}^B}}{1 + e^{\hat{\mathbf{a}}'\mathbf{x}_j^A + \hat{\mathbf{b}}'\mathbf{x}_{ik}^B}}. \tag{32}$$

En remplaçant L_i^B par $\hat{L}_i^{B(3)}$, (5) nous donne un estimateur convergent de Y^B quand le modèle spécifié dans (30) est correct et que $(\hat{\mathbf{a}}, \hat{\mathbf{b}})$ est convergent. Notons qu'il existe d'autres options que le modèle logistique, telles que les modèles logit et log-log complémentaire. Pour plus de renseignements, consulter Draper et Smith (1998). Ces auteurs mentionnent aussi que le choix du modèle n'est pas toujours évident en pratique.

3.2.2 Estimation directe de L_i^B en utilisant un modèle log-linéaire

Nous considérons qu'il existe un vecteur de variables $\mathbf{x}_i^B$ qui affecte la valeur de L_i^B. Cela signifie que le nombre total de liens dans une grappe varie seulement en fonction des caractéristiques de la grappe proprement dite. En utilisant le modèle log-linéaire, nous pouvons proposer l'expression suivante :

$$\log(L_i^B) = \theta^T \mathbf{x}_i^B. \tag{33}$$

Si la qualité de l'ajustement est raisonnable, nous pouvons estimer L_i^B directement par

$$\hat{L}_i^{B(4)} = e^{\hat{\theta}^T \mathbf{x}_i^B}, \tag{34}$$

où $\hat{\theta}$ est un estimateur de θ. Si $\hat{\theta}$ est convergent, après remplacement de L_i^B par $\hat{L}_i^{B(4)}$, (5) nous donne un estimateur convergent de Y^B. Nous notons que $\hat{L}_i^{B(4)}$ pourrait avoir une valeur non entière et donc devoir être arrondi à la valeur entière la plus proche.

4. Étude par simulation

Si, sous un plan de sondage longitudinal, on souhaite produire des estimations transversales à un point particulier dans le temps après le point de départ, la situation devient un exemple pratique de problème d'échantillonnage indirect. Puisque la population évolue au cours du temps, la population cible n'est pas la même que la population initiale à partir de laquelle a été tiré l'échantillon longitudinal. À la présente section, nous utiliserons l'Enquête sur la dynamique du travail et du revenu (EDTR) comme exemple pour démontrer la performance d'un des estimateurs que nous avons présentés à la section 3.1.

Le plan d'échantillonnage de l'EDTR est décrit en détail dans Lavallée (1993). Certains termes que nous utilisons dans le présent article, tels que cohabitants, individus initialement présents et individus initialement absents, sont ceux utilisés dans Lavallée (1995). Les individus initialement absents dans la population sont ceux qui ne faisaient pas partie de la population durant l'année où l'échantillon longitudinal a été sélection, mais qui sont considérés dans l'échantillon ultérieur ; font partie de ce groupe les nouveaunés et les immigrants. Après l'année initiale de sélection, la population compte des individus longitudinaux, des individus initialement présents et des individus initialement absents. Si nous nous concentrons sur les ménages contenant au moins un individu longitudinal (c'est-à-dire les ménages longitudinaux), les individus initialement présents ou absents qui se joignent à ces ménages sont appelés cohabitants.

Dans cet exemple particulier, U^A est la population à l'année initiale, disons yr_0, de l'enquête longitudinale et U^B est la population durant n'importe quelle année, disons l'année yr_t, après l'année initiale. L'échantillon s^A est formé de tous les individus longitudinaux. $L_{j,i}$ est une variable binaire ; sa valeur est 1 si l'individu j vit dans le i^{e} ménage à l'année yr_t et elle est 0 autrement. L_i^B est le nombre total de personnes longitudinales et de cohabitants initialement présents à l'année yr_0 qui vivent dans le i^{e} ménage à l'année yr_t.

Pour un individu longitudinal, le lien sera de type un à un (biunivoque). Dans le cas des cohabitants, il est fort probable qu'il sera impossible d'identifier le lien quelques années après l'année initiale de l'enquête, à cause, par exemple, des nouvelles naissances et de l'immigration ; en outre, plus la proportion de cohabitants dans la population cible est élevée, plus cette probabilité devient forte. Par exemple, dans le panel 3 de l'EDTR, les cohabitants représentaient 7,8 % de 47 377 individus en l'an 2000, c'est-à-dire un an après l'année initiale. La proportion était passée à 13,87 % en l'an 2002 (trois ans plus tard) et à 15,22 % en 2003 (quatre ans plus tard). Nous voyons donc que l'on ne peut pas fermer les yeux sur les non-réponses de lien chez une proportion aussi importante de cohabitants. Comme nous disposons d'information observée, nous exécutons l'estimation de L_i^B par les deux types d'ajustements proportionnels que nous avons proposés aux

sections 3.1.1 et 3.1.2. Afin d'évaluer les propriétés des estimations obtenues par ces approches, nous exécutons une étude par simulation en nous servant des données de l'EDTR. Nous nous intéressons aux estimations transversales pour quatre variables de revenu pour l'année 2003. Ces quatre variables sont le revenu total avant impôt, le revenu total après impôt, les gains (c'est-à-dire les traitements et salaires avant retenues et le revenu d'un travail autonome) et les salaires et traitements avant retenues (également appelés revenu d'emploi). Nous nous intéressons au total de population pour ces quatre variables. Nous avons estimé ces quatre grandeurs d'intérêt aux niveaux national et provincial.

Dans une enquête longitudinale, le nombre total de liens dans la grappe *i* n'est généralement pas supérieur au nombre total d'individus dans cette grappe ni inférieur au nombre d'individus longitudinaux dans cette grappe. Puisque T_i^B est inconnu, nous remplaçons T_i^B par M_i^B dans (5) pour notre étude par simulation.

Premièrement, nous supposons que les liens entre toutes les unités sélectionnées durant l'année initiale (1999) et toutes les unités présentes dans l'ensemble de la population en 2003 sont spécifiées correctement. Puis, nous calculons les totaux en utilisant la MGPP. Nous utilisons ces totaux comme estimations cibles, c'est-à-dire les valeurs « réelles. »

Deuxièmement, nous supprimons aléatoirement 50 % des liens associés aux individus initialement présents en convertissant aléatoirement certains cohabitants initialement présents en cohabitants initialement absents. Le nombre de liens enlevés de la sorte représente environ 6,3 % de la population totale à laquelle nous nous intéressons, dont la taille est de 30 224. Sans aucune correction, nous recalculons l'estimation en utilisant la MGPP. Nous utilisons cette estimation comme estimation de référence c'est-à-dire le « placebo ».

Troisièmement, nous estimons les mêmes quantités en utilisant la MGPP avec les méthodes d'ajustement proportionnelles, c'est-à-dire les méthodes (1) et (2) de la section 3.1, pour voir si les estimations sont suffisamment proches de la « valeur réelle » et déterminer l'importance de l'amélioration due à ces ajustements.

Cette étude par simulation en utilisant les données de l'EDTR démontre que la méthode proposée donne de très bons résultats en ce qui concerne la correction de la surestimation due à la non-réponse de lien.

Nous écrivons

$$w_i^{\text{moyen}} = \frac{\sum_{j=1}^{m^A} L_{j,i}^B \frac{1}{\pi_j^A}}{\sum_{j=1}^{m^A} L_{j,i}^B} \tag{35}$$

Puis, en nous servant des méthodes (1) et (2) de la section 3.1, nous estimons Y^B par

$$\hat{Y}_{\text{moyen}}^{B(1)} = \sum_{i=1}^{n} \frac{m_i^A}{T_i^A} w_i^{\text{moyen}} \sum_{k=1}^{M_i^B} y_{ik}, \tag{36}$$

et

$$\hat{Y}_{\text{moyen}}^{B(2)} = \frac{m^A}{T^A} \sum_{i=1}^{n} w_i^{\text{moyen}} \sum_{k=1}^{M_i^B} y_{ik}, \tag{37}$$

respectivement.

Notons que w_i^{moyen} est le poids moyen des personnes longitudinales qui vivent dans le i^{e} ménage durant l'année yr_t. Par conséquent, il est également raisonnable d'utiliser à sa place le poids médian :

$$w_i^{\text{moyen}} = \text{la médiane de} \frac{1}{\pi_j^A}, j = 1, 2, \ldots, m^A. \tag{38}$$

pour accroître la robustesse des estimations. De même, nous estimons Y^B par

$$\hat{Y}_{\text{moyen}}^{B(1)} = \sum_{i=1}^{n} \frac{m_i^A}{T_i^A} w_i^{\text{moyen}} \sum_{k=1}^{M_i^B} y_{ik}, \tag{39}$$

et

$$\hat{Y}_{\text{médian}}^{B(2)} = \frac{m^A}{T^A} \sum_{i=1}^{n} w_i^{\text{médian}} \sum_{k=1}^{M_i^B} y_{ik}. \tag{40}$$

Les comparaisons de ces méthodes proposées avec et sans intégration du problème de non-réponse en utilisant le poids moyen ainsi que le poids médian dans chaque ménage sont présentées aux tableaux 2 à 5.

Les quatre tableaux suivants donnent l'évaluation de la performance de nos estimations fondées sur l'erreur relative définie comme étant :

$$\left| \frac{\text{estimation - “valeur réelle”}}{\text{“valeur réelle”}} \right| \times 100\ \%.$$

Tableau 2
Revenu total avant impôt (en dollars canadiens)

Province	Estim. par MGPP sans liens manquants	Estim. par MGPP avec liens manquants	Estim. par MGPP corrigée avec moyenne	Estim. par MGPP corrigée avec médiane
T.-N.-L.	9 261 958 108	9 788 749 735	9 317 420 236	9 304 530 248
Î.-P.-É.	2 720 448 008	2 858 506 466	2 735 943 043	2 734 922 451
N.-É.	18 277 017 251	19 573 546 299	18 140 076 618	18 067 144 557
N.-B.	15 297 155 323	16 281 178 934	15 291 696 585	15 236 482 035
Qc	1,57839E+11	1,69664E+11	1,56533E+11	1,56405E+11
Ont.	2,895E+11	3,07642E+11	2,85409E+11	2,85599E+11
Man.	23 436 397 548	25 043 168 032	23 632 717 226	23 553 543 216
Sask.	20 185 285 649	21 595 804 296	20 163 683 598	20 095 359 071
Alb.	69 063 402 292	74 576 351 600	68 716 661 193	68 582 541 733
C.-B.	81 749 374 346	86 593 614 506	81 387 640 982	81 248 680 715
Canada	6,8733E+11	7,33617E+11	6,8286E+11	6,82356E+11

Tableau 3
Revenu total avant impôt (en dollars canadiens)

Province	Estim. par MGPP sans liens manquants	Estim. par MGPP avec liens manquants	Estim. par MGPP corrigée avec moyenne	Estim. par MGPP corrigée avec médiane
T.-N.-L.	7 846 587 557	8 287 351 908	7 892 754 014	7 882 437 105
Î.-P.-É.	2 300 092 795	2 416 503 441	2 314 256 124	2 313 544 320
N.-É.	15 154 508 564	16 257 679 161	15 080 155 194	15 020 088 623
N.-B.	12 878 350 198	13 718 260 686	12 894 700 593	12 849 252 205
Qc	1,27632E+11	1,37514E+11	1,27118E+11	1,26999E+11
Ont.	2,3788E+11	2,53073E+11	2,35192E+11	2,3534E+11
Man.	19 541 510 220	20 877 377 918	19 713 628 649	19 649 142 217
Sask.	16 894 929 025	18 073 635 883	16 890 410 993	16 834 787 407
Alb.	57 466 974 767	62 055 315 246	57 183 814 491	57 073 904 623
C.-B.	68 710 569 670	72 770 595 462	68 431 531 373	68 309 055 749
Canada	5,66306E+11	6,05044E+11	5,63958E+11	5,63518E+11

Tableau 4
Gains (en dollars canadiens)

Province	Estim. par MGPP sans liens manquants	Estim. par MGPP avec liens manquants	Estim. par MGPP corrigée avec moyenne	Estim. par MGPP corrigée avec médiane
T.-N.-L.	6 433 112 169	6 837 522 157	6 541 306 193	6 530 174 122
Î.-P.-É.	1 898 192 704	2 019 341 995	1 964 066 449	1 962 669 664
N.-É.	12 772 667 160	13 809 197 160	12 999 111 234	12 939 785 579
N.-B.	11 250 688 811	12 030 378 710	11 411 530 716	11 370 222 533
Qc	1,18878E+11	1,28949E+11	1,19797E+11	1,19717E+11
Ont.	2,27577E+11	2.43404E+11	2,26812E+11	2,27092E+11
Man.	17 560 695 670	18 995 682 322	18 066 353 153	18 001 882 362
Sask.	15 159 319 031	16 340 668 148	15 381 733 004	15 319 210 228
Alb.	56 152 023 359	61 059 244 608	56 540 145 524	56 418 889 147
C.-B.	60 532 655 979	64 499 398 960	61 192 920 832	61 085 986 951
Canada	5,28214E+11	5,67945E+11	5,3199E+11	5,31722E+11

Tableau 5
Salaires et traitements avant retenues (en dollars canadiens)

Province	Estim. par MGPP sans liens manquants	Estim. par MGPP avec liens manquants	Estim. par MGPP corrigée avec moyenne	Estim. par MGPP corrigée avec médiane
T.-N.-L.	6 180 713 343	6 572 345 010	6 283 079 555	6 272 429 515
Î.-P.-É.	1 636 344 440	1 747 755 878	1 713 809 312	1 713 157 676
N.-É.	12 327 220 137	13 341 912 666	12 579 519 733	12 521 159 025
N.-B.	10 742 381 379	11 508 445 078	10 961 105 589	10 921 102 477
Qc	1,08636E+11	1,18092E+11	1,10024E+11	1,09898E+11
Ont.	2,07331E+11	2,22043E+11	2,07265E+11	2,07495E+11
Man.	16 146 993 217	17 504 024 442	16 701 823 718	16 641 840 086
Sask.	13 982 423 360	15 129 217 320	14 311 467 435	14 255 519 224
Alb.	52 594 490 290	57 359 188 114	53 195 227 508	53 077 388 907
C.-B.	56 206 787 033	59 886 429 369	56 875 663 895	56 764 297 512
Canada	4,85784E+11	5,23184E+11	4,91116E+11	4,90763E+11

Tableau 6
Comparaison des erreurs relatives dans l'estimation du revenu avant impôt (%)

Province	MGPP avec liens manquants	Méthode (1) avec moyenne	Méthode (1) avec médiane	Méthode (2) avec moyenne	Méthode (2) avec médiane
T.-N.-L.	5,688	0,599	0,460	1,059	2,397
Î.-P.-É.	5,075	0,570	0,532	2,859	4,063
N.-É.	7,094	0,749	1,148	3,549	2,459
N.-B.	6,433	0,037	0,397	2,693	2,987
Qc	7,492	0,828	0,909	4,372	2,896
Ont.	6,267	1,413	1,348	4,691	1,771
Man.	6,856	0,838	0,500	1,644	3,654
Sask.	6,988	0,107	0,446	2,480	2,598
Alb.	7,982	0,502	0,696	3,185	2,407
C.-B.	5,926	0,442	0,612	3,995	3,343
Canada	6,734	0,650	0,724	3,868	2,662

Tableau 7
Comparaison des erreurs relatives dans l'estimation du revenu après impôt (%)

Province	MGPP avec liens manquants	Méthode (1) avec moyenne	Méthode (1) avec médiane	Méthode (2) avec moyenne	Méthode (2) avec médiane
T.-N.-L.	5,617	0,588	0,457	1,101	2,409
Î.-P.-É.	5,061	0,616	0,585	2,832	4,121
N.-É.	7,279	0,491	0,887	3,338	2,765
N.-B.	6,522	0,127	0,226	2,539	3,150
Qc	7,742	0,403	0,496	3,991	3,375
Ont.	6,387	1,130	1,068	4,432	2,081
Man.	6,836	0,881	0,551	1,645	3,733
Sask.	6,977	0,027	0,356	2,406	2,675
Alb.	7,984	0,493	0,684	3,180	2,415
C.-B.	5,909	0,406	0,584	3,989	3,419
Canada	6,841	0,415	0,492	3,657	2,927

Les tableaux montrent que nos estimations au moyen de la méthode (1) ainsi que de la méthode (2) donnent de très bons résultats en ce qui concerne la réduction du biais. Dans l'ensemble, la méthode (1) est meilleure que la méthode (2), mais l'amélioration constatée est nettement moindre que celle observée en passant de l'absence de correction à la méthode (2). Puisque la méthode (2) produit des résultats de qualité et requiert beaucoup moins d'information que la méthode (1), nous recommandons de la choisir.

Maintenant, nous examinons la méthode (2) avec utilisation du poids moyen, qui donne l'estimation $\hat{Y}^{B(2)}_{\text{moyen}}$, afin d'analyser les propriétés de sa variance pour l'estimation de Y^B. Nous utilisons la méthode bootstrap pour estimer la variance de $\hat{Y}^{B(2)}_{\text{moyen}}$ aux niveaux national et provincial. La méthode utilisée pour la simulation est celle du bootstrap classique avec remise, où le bootstrapage est effectué à la première étape de l'échantillonnage. Les poids bootstrap utilisés sont ceux fournis avec les données de l'EDTR et comprennent toutes les corrections nécessaires. Voir Lévesque (2001) et LaRoche (2003) pour une explication détaillée de l'utilisation du bootstrap pour l'EDTR. Le gain en ce qui concerne la réduction de la variance n'est pas aussi important que dans le cas du biais ; toutefois, l'étude par simulation montre que la méthode proposée produit une plus petite variance que l'application de la MGPP sans correction pour les liens manquants. Les résultats sont présentés au tableau 10.

Les résultats des simulations présentés ici sont fondés sur un seul échantillon de l'EDTR et une seule suppression aléatoire des liens des individus initialement présents. Pour évaluer complètement les propriétés des estimateurs susmentionnés, nous aurions dû utiliser un processus Monte Carlo. Des simulations de ce type ont été effectuées par Hurand (2006) en se basant sur des données agricoles. Dans ces simulations, 1 000 échantillons ont été sélectionnés et, pour chaque échantillon sélectionné, le pire scénario a été utilisé, c'est-à-dire l'élimination de tous les liens provenant des unités non échantillonnées. Les résultats de ces simulations ont montré que l'ajustement proportionnel et l'ajustement proportionnel global sont les deux méthodes qui s'approchent, en moyenne, le plus du total réel et celles dont les biais sont négligeables.

Tableau 8
Comparaison des erreurs relatives dans les estimations des gains (%)

Province	MGPP avec liens manquants	Méthode (1) avec moyenne	Méthode (1) avec médiane	Méthode (2) avec moyenne	Méthode (2) avec médiane
T.-N.-L.	6,286	1,682	1,509	0,041	3,585
Î.-P.-É.	6,382	3,470	3,397	0,0739	7,115
N.-É.	8,115	1,773	1,308	1,265	5,281
N.-B.	6,930	1,430	1,062	1,279	4,512
Qc	8,472	0,773	0,706	2,827	4,560
Ont.	6,955	0,336	0,213	3,760	2,920
Man.	8,172	2,879	2,512	0,291	5,835
Sask.	7,793	1,467	1,055	0,979	4,324
Alb.	8,739	0,691	0,475	2,140	3,777
C.-B.	6,553	1,091	0,914	2,643	5,081
Canada	7,522	0,715	0,664	2,628	4,131

Tableau 9
Comparaison des erreurs relatives dans les estimations des traitements et salaires avant retenues (%)

Province	MGPP avec liens manquants	Méthode (1) avec moyenne	Méthode (1) avec médiane	Méthode (2) avec moyenne	Méthode (2) avec médiane
T.-N.-L.	6,336	1,656	1,484	0,1012	3,593
Î.-P.-É.	6,809	4,734	4,694	1,056	8,424
N.-É.	8,231	2,047	1,573	0,939	5,509
N.-B.	7,131	2,036	1,664	0,685	5,133
Qc	8,704	1,278	1,162	2,294	5,070
Ont.	7,096	0,0317	0,0791	3,473	3,265
Man.	8,404	3,436	3,065	0,787	6,469
Sask.	8,202	2,353	1,953	0,107	5,213
Alb.	9,059	1,142	0,918	1,713	4,247
C.-B.	6,547	1,190	0,992	2,565	5,234
Canada	7,699	1,098	1,025	2,251	4,541

Tableau 10
Comparaison des estimations de l'écart-type

	Variables	Revenu total avant impôt	Revenu total après impôt	Gains	Salaires et traitements avant retenues
Niveau national	MGPP sans liens manquants	9 677 258 789	7 343 792 762	8 850 202 075	8 468 718 449
	Méthode (2) avec moyenne	9 471 103 083	7 238 715 323	8 593 015 854	8 232 428 642
Ontario	MGPP sans liens manquants	7 888 106 377	6 101 001 739	7 245 688 373	7 149 203 530
	Méthode (2) avec moyenne	7 601 169 501	5 939 509 894	6 952 217 872	6 831 300 511
Québec	MGPP sans liens manquants	4 341 215 711	3 113 247 130	3 772 369 180	3 162 277 660
	Méthode (2) avec moyenne	4 160 251 472	2 974 248 451	3 668 996 929	3 100 868 366

5. Conclusion

Nous avons élaboré quatre méthodes d'estimation pour traiter le problème de la non-réponse de lien dans l'échantillonnage indirect. Les résultats des simulations exposés dans l'article montrent que les méthodes de correction que nous avons présentées pour illustrer l'utilisation de la MGPP avec intégration de la non-réponse de lien donnent de bons résultats pour ce qui est de réduire le biais de l'estimation et produisent une amélioration globale de la variance. Le progrès en ce qui concerne la réduction du biais semble important. L'application de la méthode proposée à la section 3.2 à des ensembles de données réelles sera étudiée prochainement.

Les constatations importantes qui se dégagent de la présente étude sont les suivantes :

1. Les méthodes d'ajustement sont faciles à appliquer.
2. Dans une situation plus générale, telle que $L_{j,i} > 1$ pour certaines unités j, (35) représente la moyenne pondérée par $L^B_{j,i}$. Conséquemment, l'approche de la médiane donnée par (39) et (40) peut être modifiée en utilisant une version généralisée de la médiane, c'est-à-dire la médiane « pondérée ». Autrement dit, nous remplaçons (38) par

$$w_i^{\text{médian}} = \text{la médiane de} \frac{1}{\pi_j^A}$$

où $j = 1, 2, \ldots, L^B_{1,i}\,; 1, 2, \ldots, L^B_{2,i}\,; \ldots; 1, 2, \ldots, L^B_{m^A,i}$.

3. Certaines réponses de lien valides en dehors de s^A ne peuvent pas être utilisées pour estimer L_i^B par les méthodes proposées à la section 3.1. Toutefois, cette information valide pourrait être avantageuse dans les approches de prédiction de $l_{j,ik}$ en utilisant des variables auxiliaires, comme le montre la section 3.2.1.

Remerciements

Les auteurs remercient le rédacteur associé et deux examinateurs de leurs suggestions et commentaires constructifs au sujet des versions précédentes du présent article. Ces travaux de recherche ont été financés par le Conseil de recherches en sciences naturelles et en génie du Canada et par Mathématiques des technologies de l'information et des systèmes complexes.

Bibliographie

Cassel, C.-M., Särndal, C.-E. et Wretman, J. (1977). *Foundations of Inference in Survey Sampling*. New York : John Wiley & Sons, Inc.

Deville, J.-C., et Lavallée, P. (2006). Sondage indirect : les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*, 32, 185-196.

Draper, N.R., et Smith, H. (1998). *Applied Regression Analysis*, 3ième Éd. New York : John Wiley & Sons, Inc.

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. Dans *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York : John Wiley & Sons, Inc., 135-159.

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

Hurand, C. (2006). La méthode généralisée du partage des poids et le problème d'identification des liens. Rapport interne, Divison des méthodes d'enquêtes sociales, Statistique Canada, juillet 2006.

Kalton, G., et Brick, J.M. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, 21, 37-49.

LaRoche, S. (2003). Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics. *Income Research Paper Series*, Catalogue no. 75F0002MIE - No. 007, Statistique Canada.

Lavallée, P. (1993). Sample representativity for the Survey of Labour and Income Dynamics. *Statistics Canada*, *Research Paper of the Survey of Labour and Income Dynamics*, Catalogue No. 93-19, décembre 1993.

Lavallée, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquêtes*, 21, 27-35.

Lavallée, P. (2001). Correcting for non-response in indirect sampling. *Proceedings of Statistics Canada's Symposium 2001*.

Lavallée, P. (2002). *Le sondage indirect, ou la méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles et Éditions Ellipse.

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

Lévesque, I. (2001). Enquête sur la dynamique du travail et du revenu - Estimation de la variance. Rapport interne de Statistique Canada, 2 juillet 2001.

Särndal, C.-E., Swensson, B. et Wretman, J. (1991). *Model Assisted Survey Sampling*. New York : Springer-Verlag.