Techniques d'enquête, décembre 1998
Vol. 24, n° 2, pp. 133-140
Statistique Canada

Estimation des flux bruts de la population active provenant d'enquêtes donnant lieu à une non-réponse dont il faut tenir compte au niveau du ménage

PAUL S. CLARKE et RAY L. CHAMBERS[1]

RÉSUMÉ

La mesure des flux bruts de la population active est un objectif important des enquêtes continues sur la population active effectuées par un grand nombre d'offices nationaux de la statistique. Cependant, il est bien connu que l'estimation de ces flux peut être compliquée par une non-réponse, des erreurs de mesure, un renouvellement de l'échantillon et des effets complexes du plan de sondage. Le présent article, inspiré par des modèles de non-réponse dans les enquêtes sur les ménages, porte sur l'estimation des flux bruts tout en apportant des ajustements en fonction de la non-réponse dont il faut tenir compte. Les approches antérieures basées sur un modèle en ce qui concerne l'estimation des flux bruts supposaient que la non-réponse était un processus au niveau de la personne. Nous proposons une catégorie de modèles qui permettent une non-réponse dont il faut tenir compte au niveau du ménage. On a recours à une étude en simulation pour démontrer que les estimations des flux bruts de la population active au niveau de la personne provenant des données d'enquêtes sur les ménages peuvent être biaisées et que les estimations en fonction de modèles au niveau du ménage peuvent permettre de réduire ce biais.

MOTS CLÉS: Flux bruts; enquêtes sur les ménages; non-réponse dont il faut tenir compte.

1. INTRODUCTION

On définit de façon générale les flux bruts de la population active comme étant des transitions dans le temps entre les trois grands états de la population active, personnes occupées, personnes sans emploi et personnes économiquement inactives. Les estimations des flux bruts sont un outil important dans l'étude de la dynamique de la population active (par exemple, voir Vanski 1985). Les enquêtes continues à grande échelle telle la Labour Force Survey en Grande-Bretagne et la Current Population Survey aux États-Unis fournissent des données pour l'estimation des flux bruts. Cependant, la non-réponse, l'erreur de mesure, le renouvellement de l'échantillon et les effets complexes du plan de sondage ont une incidence sur l'estimation des flux bruts de ces enquêtes. Hogue (1985) traite de ces facteurs et d'autres qui ont une incidence sur l'estimation des flux bruts. Dans le présent article, nous nous concentrons sur le problème de la non-réponse.

Nous supposons qu'un mécanisme de non-réponse a pour conséquence que les données observées sont incomplètes. Si la probabilité de la non-réponse dépend des données manquantes, alors le mécanisme de non-réponse est un mécanisme dont il faut tenir compte (Rubin 1976). L'approche basée sur des modèles pour analyser des données incomplètes d'enquête est détaillée dans Little (1982). Les approches basées sur des modèles relativement à l'estimation des flux bruts de la population active font intervenir la modélisation des flux de la population active et du mécanisme de non-réponse tout en assurant l'ajustement des deux modèles aux données incomplètes. Des exemples de ces modèles sont données dans Stasny et Fienberg (1985), Stasny (1986) et, dans le cas de la non-réponse dont il faut tenir compte, dans Little (1985). Nous disons de ces modèles qu'ils sont au niveau de la personne parce que les personnes sont modélisées comme répondant ou ne répondant pas, indépendamment des autres personnes échantillonnées.

Tant la Labour Force Survey que la Current Population Survey sont des exemples d'enquêtes sur les ménages, c'est-à-dire des enquêtes se fondant sur un échantillon aléatoire des ménages plutôt que sur des personnes. Les enquêtes sur les ménages peuvent donner lieu à un comportement corrélé de non-réponse au sein des ménages. Par exemple, dans la Current Population Survey, un seul membre du ménage (habituellement le chef du ménage) agit en tant que représentant des membres du ménage; ainsi, si le membre choisi du ménage ne répond pas, les autres membres du ménage ne répondent pas non plus. Il s'ensuit que, en raison du comportement corrélé de non-réponse au sein du ménage, les modèles de non-réponse au niveau de la personne ne conviennent pas à l'estimation des flux bruts de la population active qui utilise des données d'enquêtes sur les ménages.

Dans le présent article, nous proposons une catégorie de modèles pour les flux de la population active au niveau de la personne et une non-réponse au niveau du ménage qui tient compte du comportement corrélé de non-réponse au sein du ménage. On présente également un certain nombre de modèles plausibles de non-réponse qui sont estimables à partir des données observées, tant pour ce qui est de la non-réponse dont on n'a pas à tenir compte que pour la non-

[1] Paul S. Clarke et Ray L. Chambers, Department of Social Statistics, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom.

réponse dont il faut tenir compte. Nous simulons alors des données d'enquêtes sur les ménages à l'aide de ces modèles au niveau du ménage afin de démontrer l'utilité éventuelle de notre approche: premièrement, on démontre que les estimations des flux bruts de la population active au niveau de la personne sont biaisées lorsqu'elles sont ajustées en fonction des données d'enquêtes sur les ménages; deuxièmement, on compare le biais des estimations des flux bruts au niveau de la personne à celui au niveau du ménage afin de démontrer les avantages d'ajuster les modèles au niveau des ménages aux données d'enquêtes sur les ménages. Enfin, nous résumons les conclusions de nos études en simulation et nous présentons des orientations pour des études ultérieures.

2. UN MODÈLE POUR LA NON-RÉPONSE AU NIVEAU DU MÉNAGE

2.1 Les données

Un flux brut est la probabilité ou la fréquence de personnes au sein de la population qui font une transition d'états entre deux points dans le temps, t_1 et $t_2 (t_1 < t_2)$. Les flux bruts de la population active renvoient aux transitions entre les trois grands états de la population active: 1 = «personne occupée», 2 = «personne sans emploi» et 3 = «hors de la population active», la dernière catégorie faisant référence aux personnes économiquement inactives telles les personnes à la retraite et les étudiants. Soit S qui représente un échantillon aléatoire simple de ménages, indexé par h. Au sein du ménage h, il y a n_h personnes admissibles, dont $n_h(ab)$ ont un flux de population active (a, b) entre t_1 et t_2, où $\sum_{a,b} n_h(ab) = n_h$ et $a, b = 1, 2, 3$. Nous disons que $\{n_h(ab)\}$ sont les données complètes, c'est-à-dire les fréquences qui seraient observées en l'absence de non-réponse.

Le tableau 1 illustre les données des flux de la population active complètes pour le ménage h comme étant un tableau de contingence de 3×3. Si h répond les deux fois, les données observées sont les cellules de ce tableau à double entrée. Cependant, si le ménage ne répond pas à t_1 ou t_2, les données observées correspondent aux marges du tableau: $n_h(1+), n_h(2+), n_h(3+)$ sont les données observées si h répond à t_1 mais ne répond pas à t_2; et $n_h(+1)$, $n_h(+2), n_h(+3)$ sont les données observées si h répond à t_2 mais ne répond pas à t_1. (Un indice remplacé par « + » représente la somme de tous les niveaux de cet indice.) En outre, si h ne répond pas à t_1 et t_2, les données observées sont la taille du ménage, n_h, que nous supposons connues et fixes entre t_1 et t_2.

Tableau 1
Données complètes des flux de la population active pour le ménage h

État		t_2 1	2	3	
t_1	1	$n_h(11)$	$n_h(12)$	$n_h(13)$	$n_h(1+)$
	2	$n_h(21)$	$n_h(22)$	$n_h(23)$	$n_h(2+)$
	3	$n_h(31)$	$n_h(32)$	$n_h(33)$	$n_h(3+)$
		$n_h(+1)$	$n_h(+2)$	$n_h(+3)$	n_h

2.2 Spécification du modèle

Il ne convient pas de traiter le comportement de non-réponse de personnes au sein des ménages comme étant indépendant dans les enquêtes sur les ménages. Dans la Labour Force Survey, par exemple, un membre admissible du ménage détermine si le ménage peut être interviewé. Par conséquent, s'il n'y a aucune personne admissible que l'on peut contacter, chaque personne du ménage ne répond pas. Pour construire un modèle de la non-réponse au niveau du ménage, nous prenons les idées à l'origine de la non-réponse au niveau de la personne et nous les étendons au ménage en considérant qu'un ménage est une entité ayant son propre flux de non-réponse entre t_1 et t_2. Pour permettre une non-réponse dont il faut tenir compte, la probabilité d'un flux de non-réponse au sein d'un ménage est modélisée en tant qu'une fonction de ses flux individuels de la population active, que nous devons maintenant décrire.

Soit $\boldsymbol{N}_h = (N_h(11), N_h(21), ..., N_h(33))$ le vecteur aléatoire des fréquences des flux de la population active pour le ménage h, où $N_h(ab)$ est la variable aléatoire dont l'extrant correspond au nombre de personnes du flux de la population active $(a, b), a, b = 1, 2, 3$. En outre, désignons le vecteur aléatoire pour le flux de non-réponse du ménage h par $\boldsymbol{R}_h = (R_{h1}, R_{h2})$, où

$$R_{hj} = \begin{cases} 1, & \text{si ménage répond à } t_j \\ 0, & \text{sinon} \end{cases}$$

est la variable aléatoire de l'état de non-réponse pour h à t_j, $j = 1, 2$. Les réalisations de ces quantités aléatoires sont désignées par $\boldsymbol{n}_h$ et $\boldsymbol{r}_h$. Nous supposons maintenant que $\boldsymbol{n}_h$ et $\boldsymbol{r}_h$ sont connus et formulons la probabilité conjointe de $\boldsymbol{N}_h$ et $\boldsymbol{R}_h$ comme

$$\Pr(\boldsymbol{N}_h = \boldsymbol{n}_h, \boldsymbol{R}_h = \boldsymbol{r}_h) = \Pr(\boldsymbol{N}_h = \boldsymbol{n}_h) \Pr(\boldsymbol{R}_h = \boldsymbol{r}_h \mid \boldsymbol{N}_h = \boldsymbol{n}_h),$$

où $\Pr(\boldsymbol{N}_h = \boldsymbol{n}_h)$ est le modèle des flux de la population active et $\Pr(\boldsymbol{R}_h = \boldsymbol{r}_h \mid \boldsymbol{N}_h = \boldsymbol{n}_h)$ est appelé le modèle des flux de non-réponse.

Le modèle des flux de la population active est multinomial avec une fonction de probabilité

$$\Pr(\boldsymbol{N}_h = \boldsymbol{n}_h; \omega) = n_h! \prod_{a,b} \frac{\omega(ab)^{n_h(ab)}}{n_h(ab)!}, \quad (1)$$

où $\omega(ab) > 0$ est la probabilité pour une personne d'avoir un flux de la population active (a, b) et $\sum_{a,b}\omega(ab) = 1$. Le vecteur des paramètres des flux de la population active est désigné par $\omega = (\omega(11), \omega(21), ..., \omega(33))$, dont 8 sont libres. L'hypothèse de l'échantillonnage multinomial dans (1) sous-entend que le comportement des flux de la population active des personnes est indépendant au sein des ménages et que les ménages sont homogènes pour ce qui est du comportement de leurs flux de la population active. Ces hypothèses ne sont pas réalistes, mais (1) peut facilement être étendu à un modèle plus réaliste des flux de la population active, tel qu'il est indiqué à la section 4.

La probabilité pour le ménage h d'avoir un flux de non-réponse (u, v) est

$$\pi(uv \mid \boldsymbol{n}_h) = \Pr(\boldsymbol{R}_h = (u, v) \mid \boldsymbol{N}_h = \boldsymbol{n}_h; \psi) = \frac{1}{n_h} \sum_{a,b} n_h(ab)\psi(uv \mid ab), \quad (2)$$

lorsque $u, v = 0, 1$, à savoir une moyenne pondérée des paramètres du modèle de non-réponse. En établissant $n_h = 1$, on peut constater que $\psi(uv \mid ab) > 0$ est la probabilité d'un ménage d'une taille un (c'est-à-dire une personne) qui a un flux de non-réponse (u, v) s'il a un flux de population active (a, b). Ainsi, $\sum_{u,v}\psi(uv \mid ab) = 1$ et $\psi = (\psi(11 \mid 11), \psi(01 \mid 11), ..., \psi(00 \mid 33))$ est le vecteur des paramètres de non-réponse, dont 27 sont libres.

Avant de définir la fonction de vraisemblance des données complètes, partageons S en 4 sous-ensembles exhaustifs et mutuellement exclusifs

$$S = S_{11} \cup S_{01} \cup S_{10} \cup S_{00},$$

où $S_{uv} = \{h : \boldsymbol{r}_h = (u, v)\}$ est le sous-ensemble des ménages ayant un flux de non-réponse (u, v). Ainsi, étant donné que S est un échantillon aléatoire simple des ménages, la fonction de vraisemblance des données complètes est

$$L(\omega, \psi; \{\boldsymbol{n}_h, \boldsymbol{r}_h\}) = \prod_{u,v} \prod_{h \in S_{uv}} L_h(\omega, \psi; \boldsymbol{n}_h, (u, v)), \quad (3)$$

où $L_h(\omega, \psi; \boldsymbol{n}_h, (u, v))$ est la contribution du ménage $h \in S_{uv}$ à la vraisemblance, le produit de (1) et (2).

2.3 Ajustement des modèles

2.3.1 Estimation de vraisemblance maximale

Étant donné que les données complètes ne sont pas disponibles, (3) doit être modifié de façon à donner la vraisemblance fondée sur les données observées. Représentons les données observées par $\{\boldsymbol{n}_h^*\}$. Tel qu'on l'a indiqué à la section 2.1, les données observées pour les ménages qui répondent à t_1 et t_2 sont la classification croisée complète du tableau 1, à savoir $\boldsymbol{n}_h^* = \boldsymbol{n}_h$. De même, si $h \in S_{10}$ alors $\boldsymbol{n}_h^* = (n_h(1+), n_h(2+), n_h(3+))$; si $h \in S_{01}$ alors $\boldsymbol{n}_h^* = (n_h(+1), n_h(+2), n_h(+3))$; et si $h \in S_{00}$ alors $\boldsymbol{n}_h^* = n_h$.

On obtient la contribution du ménage $h \in S_{uv}$ à la vraisemblance des données observées en faisant le total de $L_h(\omega, \psi; \boldsymbol{n}_h, (u, v))$ pour toutes les valeurs possibles que la classification croisée complète 3×3 des flux de la population active peut prendre compte tenu de la marge observée. Représentant cet ensemble de tableaux par $\boldsymbol{n}_h : \boldsymbol{n}_h^*$, la vraisemblance des données observées pour S est

$$L(\omega, \psi; \{\boldsymbol{n}_h^*, \boldsymbol{r}_h\}) = \prod_{u,v} \prod_{h \in S_{uv}} \sum_{\boldsymbol{n}_h : \boldsymbol{n}_h^*} L_h(\omega, \psi; \boldsymbol{n}_h, (u, v)). \quad (4)$$

L'ajustement du modèle nécessite le calcul de (4) à chaque étape d'un processus d'optimisation itérative. Sur le plan des calculs, ceci est exigent parce que la fonction de vraisemblance des données complètes doit être additionnée explicitement pour les données manquantes. Par exemple, les données observées pour $h \in S_{10}$ est $\boldsymbol{n}_h^* = (n_h(1+), n_h(2+), n_h(3+))$ et la contribution de vraisemblance de ce ménage à la vraisemblance des données observées est

$$\sum_{\boldsymbol{n}_h : \boldsymbol{n}_h^*} L_h(\omega, \psi; \boldsymbol{n}_h, (1, 0)).$$

Pour calculer explicitement cette contribution, chaque tableau des données complètes 3×3 $\boldsymbol{n}_h$ pour $\boldsymbol{n}_h^*$ fixe est produit et $L_h(\omega, \psi; \boldsymbol{n}_h, (1, 0))$ est évaluée pour chacun. Dans le cas d'un ménage de taille $n_h = 5$, il y a au moins 21 tableaux possibles et au plus 108 tableaux possibles, selon les valeurs de la marge fixe; lorsque $n_h = 15$, un ménage d'une très grande taille, les nombres respectifs sont de 136 et de 9 261. Une procédure semblable est utilisée pour $h \in S_{01}$, sauf qu'ici $\boldsymbol{n}_h^* = (n_h(+1), n_h(+2), n_h(+3))$ est la marge fixe. Si $h \in S_{00}$ alors aucune donnée n'est observée au sujet de la situation vis-à-vis de la population active, seulement la taille du ménage n_h. Donc chaque tableau 3×3 ayant le total n_h doit être produit et la fonction de vraisemblance être calculée pour chacun: lorsque $n_h = 5$ il y a 1 287 tableaux et lorsque $n_h = 15$ il y a 490 314 tableaux. Il est impossible, en ce qui concerne le temps d'exécution machine, de calculer ces sommes directement. Le nombre de calculs explicites peut être réduit si l'on reconnaît que chaque ménage est défini uniquement par ses fréquences des flux de la population active observés et son flux de non-réponse. Ainsi, la somme des données manquantes ne doit être effectuée qu'une fois pour un ménage qui a des fréquences de flux de la population active et un flux de non-réponse donnés; la contribution de ce ménage à la vraisemblance est alors élevée à la puissance du nombre de ménages définis de façon semblable dans S.

2.3.2 Estimabilité des paramètres

Si nous posons $n_h = 1$ pour tous les h, les données complètes n'ont aucune structure de ménage et constituent un tableau à entrée quadruple recoupé par la situation vis-à-vis de la population active et l'état de non-réponse à t_1 et t_2. Le logarithme du rapport de vraisemblance (4) des données observées est maintenant l'équivalent de celui des modèles au niveau de la personne dans Stasny et Fienberg (1985), Little (1985) et Stasny (1986). Pour ces modèles, l'estimabilité nécessite que le nombre de paramètres des modèles ne soit pas supérieur à 15 (un pour chaque cellule de tableau observée moins une pour la contrainte de l'échantillonnage multinomial). En conséquence, (ω, ψ) sont impossibles à estimer parce qu'il y a $8 + 27 = 35$ paramètres libres. Étant donné que l'intérêt est concentré sur les probabilités des flux bruts de la population active, ω, il est nécessaire de limiter ψ afin d'assurer l'estimabilité.

Lorsque $n_h > 1$, déterminer l'estimabilité des paramètres est plus difficile parce que (4) a une expression en forme analytique fermée compliquée. Fitzmaurice, Laird et Zahner (1996) utilisent une méthode numérique pour déterminer l'estimabilité qui implique une démonstration que la matrice d'information est régulière dans le voisinage de l'estimation de vraisemblance maximale. Cependant, ce n'est non seulement peu pratique pour les problèmes d'une dimension élevée, mais évaluer la matrice d'information pour le modèle au niveau des ménages est particulièrement difficile dans ce cas. Au lieu, nous adoptons une approche pragmatique pour déterminer l'estimabilité des paramètres: tout d'abord, nous restreignons l'attention aux modèles qui remplissent la condition nécessaire pour l'estimabilité lorsque $n_h = 1$; et deuxièmement, on utilise différentes valeurs de départ pour chaque ajustement. Si les valeurs de départ différentes révèlent une estimation de vraisemblance maximale non unique, ou si les estimations des paramètres demeurent inchangées par rapport à leurs valeurs de départ, alors on suppose que les paramètres du modèle ne peuvent être estimés.

2.4 Modèles de non-réponse

Pour permettre d'obtenir les estimations des paramètres à partir des données observées, θ et ψ doivent être limités conformément aux hypothèses au sujet du mécanisme de non-réponse. Les paramètres de non-réponse sont interprétés comme des probabilités de non-réponse individuelles, mais en ce qui concerne le cadre de ménages établi jusqu'à maintenant, il n'est pas approprié de parler de personnes qui ne répondent pas. Cependant, en réalité ce sont les personnes au sein des ménages qui déterminent un flux de non-réponse d'un ménage, et non le ménage lui-même. Par conséquent, les contraintes sont placées sur les paramètres de non-réponse au niveau de la personne qui s'appliquent au niveau du ménage par la dépendance fonctionnelle de $\pi(uv \mid \boldsymbol{n}_h)$ sur ψ dans (2). Par exemple, si les paramètres de non-réponse sont limités de sorte que $\psi(uv \mid ab) = \psi(uv)$ pour la totalité de a,b, alors on n'a pas à tenir compte du mécanisme de non-réponse du ménage parce que les flux de non-réponse sont indépendants des flux de la population active.

Nous présentons maintenant quatre modèles pour le mécanisme de non-réponse, deux dont on doit tenir compte et deux dont on n'a pas à tenir compte.

- Modèles dont on n'a pas à tenir compte.
  - Modèle I_A: Probabilité de non-réponse constante,

$$\psi(uv \mid ab) = \lambda^{1-u}(1-\lambda)^u \times \lambda^{1-v}(1-\lambda)^v,$$

    qui a 1 paramètre, λ, la probabilité d'une personne qui ne répond pas;
  - Modèle I_B: Indépendant de la situation vis-à-vis de la population active, mais probabilités de non-réponse différentes à t_1 et t_2,

$$\psi(uv \mid ab) = \lambda^{1-u}(1-\lambda)^u \times \theta^{1-v}(1-\theta)^v,$$

    qui a 2 paramètres, λ, θ, soit les probabilités de non-réponse à t_1 et t_2, respectivement.
- Modèles dont il faut tenir compte.
  - Modèle N_A: Les distributions de non-réponse à t_1 et t_2 sont indépendantes, mais dépendent de la situation vis-à-vis de la population active à t_1 et t_2, respectivement,

$$\psi(uv \mid ab) = \lambda(a)^{1-u}(1-\lambda(a))^u \times \theta(b)^{1-v}(1-\theta(b))^v$$

    qui a 6 paramètres, $\boldsymbol{\lambda} = (\lambda(1), \lambda(2), \lambda(3))$ et $\boldsymbol{\theta} = (\theta(1), \theta(2), \theta(3))$, où $\lambda(a)$ est la probabilité de non-réponse à t_1 si on a la situation vis-à-vis de la population active a à t_1, et $\theta(b)$ à t_2 si on a la situation vis-à-vis de la population active b à t_2;
  - Modèle N_B: Les distributions de non-réponse à t_1 et t_2 dépendent de la situation vis-à-vis de la population active à t_1 et t_2 respectivement, c.-à-d. un processus de Markov d'ordre un. Contrairement à N_A, les distributions de non-réponse à t_1 et t_2 sont dépendantes: si l'état de non-réponse à t_1 est 1, alors la distribution de non-réponse à t_2 est la même qu'à t_1; mais si l'état de non-réponse à t_1 est 0, les distributions de non-réponse sont distinctes,

$$\psi(uv \mid ab) = \lambda(a)^{1-u}(1-\lambda(a))^u \times \begin{cases} \lambda(b)^{1-v}(1-\lambda(b))^v, & \text{if } u = 1, \\ \theta(b)^{1-v}(1-\theta(b))^v, & \text{if } u = 0, \end{cases}$$

lorsque $a, b = 1, 2, 3$ et $u, v = 0, 1$. En vertu du modèle I_A, il y a un total de $8 + 1 = 9$ paramètres libres, remplissant la condition nécessaire pour l'estimabilité d'un modèle au niveau de la personne. Les modèles I_B, N_A et N_B ont 10, 14 et 14 paramètres libres, respectivement, et remplissent donc aussi la condition nécessaire pour l'estimabilité.

3. ÉTUDE EN SIMULATION

3.1 Procédure de simulation

Nous avons eu recours à une étude en simulation pour examiner les conséquences de ne pas tenir compte de la structure du ménage dans le cas des données d'enquêtes sur les ménages pour comparer les estimations des flux bruts de la population active pour les modèles au niveau de la personne et au niveau du ménage. À cette fin, on a produit des données d'enquêtes sur les ménages à l'aide de l'échantillonnage de Monte Carlo. Chaque ensemble de données de l'échantillon se composait de 10 000 personnes réparties dans des ménages de taille $n_h = k$ pour la totalité des h. Au sein de chaque ménage, on a produit les flux de la population active à partir de (1) et on a produit le flux de non-réponse à partir de (2) en vertu d'un des modèles N_A ou N_B. On a rendu les données incomplètes en regroupant chaque tableau des flux de la population active des données complètes pour qu'il soit uniforme avec le flux de non-réponse du ménage. Au total, 1 000 ensembles de données indépendantes ont été produits de cette façon.

Les paramètres de la population utilisés pour produire les flux de la population active sont illustrés dans le tableau suivant:

		b		
$\omega(ab)$		1	2	3
	1	0.43	0.245	0.035
a	2	0.02	0.16	0.01
	3	0.015	0.035	0.05

Il s'agit de toute évidence d'une population en récession étant donné que la probabilité de passer de personne employée à personne sans emploi est très grande ($\omega(12) = 0.245$). En vertu des modèles N_A et N_B les paramètres de la population sont

	i		
	1	2	3
$\lambda(i)$	0.2	0.8	0.5
$\theta(i)$	0.5	0.2	0.8

On doit remarquer que ces valeurs des paramètres ne représentent pas un comportement des flux de non-réponse réaliste, elles ont été choisies dans le but d'illustrer la présente méthodologie. Cependant, cela n'influe pas sur les conclusions générales de l'article, qui sont également pertinentes pour des valeurs réalistes des véritables probabilités de non-réponse.

3.2 Résultats de la simulation

On obtient les estimations pour les modèles au niveau de la personne en corrigeant (4) avec $n_h = 1$ pour chaque ensemble de données incomplètes. La figure 1 résume les distributions d'échantillonnage de l'estimation de la vraisemblance maximale au niveau de la personne de $\omega(12)$, $\hat{\omega}(12)$, pour les modèles de non-réponse I_A, I_B, N_A et N_B (les estimations pour les modèles dont on n'a pas à tenir compte I_A et I_B sont incluses parce que les deux donnent les mêmes estimations des flux de la population active). Les traits verticaux représentent les intervalles entre le percentile 2,5 et le percentile 97,5 de chaque distribution de l'échantillonnage de l'estimation, et les points en caractères gras représentent ses médianes. On obtient trois distributions pour chaque estimation au niveau de la personne: la distribution la plus à gauche survient lorsque la taille du ménage est $k = 1$, les données simulées n'ont aucune structure de ménage; et si on lit de gauche à droite, les deux distributions suivantes sont celles que l'on obtient lorsque la taille du ménage est $k = 2$ et $k = 5$, respectivement. Le trait vertical plein désigne la véritable probabilité de flux, $\omega(12) = 0{,}0245$. Le comportement de la distribution d'échantillonnage de $\hat{\omega}(12)$ dans la présente étude reflète celui des autres estimations des flux bruts de la population active.

La figure 1a) en résume les distributions d'échantillonnage lorsque N_A est le véritable modèle. Si le modèle ajusté au niveau de la personne est I_A, I_B ou N_B, les estimations des flux bruts de la population active ont d'importants biais, quelle que soit la taille du ménage. Comme on devrait s'y attendre, l'estimation de la médiane pour le modèle correct N_A n'est pas biaisée si $k = 1$ et a un petit biais pour $k = 2$ et $k = 5$ (quoique ce biais soit plus faible dans le cas de $k = 5$ que dans celui de $k = 2$). La diminution du biais lorsque la valeur de k augmente est également apparente dans le cas des estimations au niveau de la personne I_A, I_B et N_B. Ce comportement n'est pas prévu étant donné qu'il semble naturel de s'attendre à ce que le biais des estimations au niveau de la personne augmente avec la taille du ménage. Les résultats sont légèrement différents dans la figure 1b) lorsque N_B est vrai. Dans le présent cas, l'estimation pour le modèle N_B au niveau de la personne devient plus biaisée à mesure que la valeur de k augmente, mais le biais diminue pour les modèles mal spécifiés I_A, I_B et N_B au niveau de la personne. En outre, les estimations mal spécifiées pour I_A et I_B comportent un petit biais lorsqu'on les compare à celles du modèle mal spécifié N_A. On traite de ces résultats à la section 3.3.

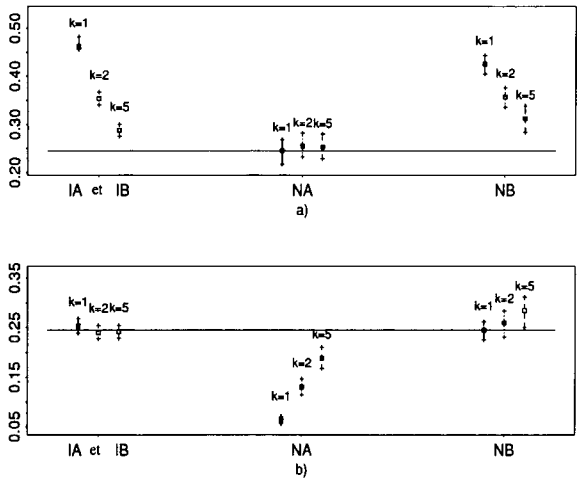

Figure 1. Distribution d'échantillonnage de $\hat{\omega}(12)$ pour les modèles au niveau de la personne, I_A, I_B, N_A et N_B lorsque le véritable modèle de non-réponse est a) N_A et b) N_B et la taille du ménage est $k = 1, 2, 5$.

Une comparaison des estimations de la médiane de $\omega(12)$ pour les modèles corrigés au niveau de la personne et au niveau du ménage lorsque N_B est vrai est illustrée à la figure 2. Il y a quatre distributions d'échantillonnage associées à chaque modèle: les deux premières représentent celles dérivées d'un modèle de non-réponse au niveau de la personne et un modèle de non-réponse au niveau du ménage lorsque la taille du ménage est $k = 2$; et de même, les deux distributions suivantes sont celles que l'on obtient lorsque la taille du ménage est 5.

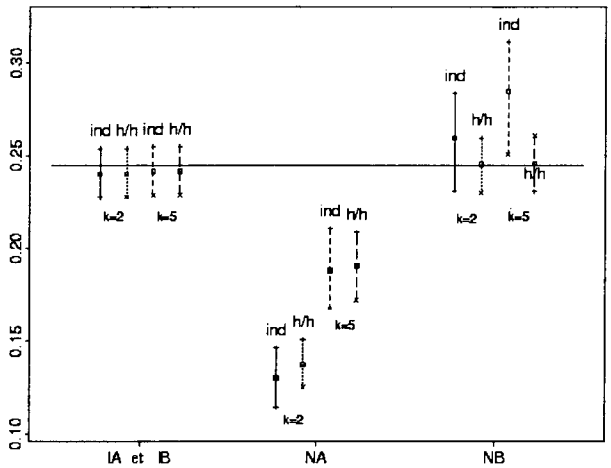

Figure 2. Les distributions d'échantillonnage de $\hat{\omega}$ (12) pour les modèles au niveau de la personne et au niveau du ménage I_A, I_B, N_A et N_B lorsque le véritable modèle de non-réponse est N_B et la taille du ménage est $k = 2, 5$.

Pour une paire donnée de distributions d'échantillonnage au niveau de la personne et au niveau du ménage, on peut constater que l'estimation au niveau du ménage est moins biaisée que celle au niveau de la personne, et l'étendue de chaque distribution d'échantillonnage au niveau du ménage est plus faible. L'exception à cet énoncé est lorsque l'on corrige le modèle I_A, où les distributions au niveau de la personne et au niveau du ménage sont identiques. Cette égalité survient parce que la vraisemblance des données observées pour les modèles au niveau de la personne et au niveau du ménage sont équivalentes lorsque l'on n'a pas à tenir compte du modèle de non-réponse. Une autre caractéristique est que si le modèle de non-réponse est correctement spécifié, les estimations au niveau du ménage ne sont pas biaisées.

3.3 Résumé

Les estimations des flux bruts de la population active en vertu des modèles au niveau de la personne ne sont jamais moins biaisées que celles des modèles au niveau du ménage lorsqu'ils sont corrigés en fonction des données d'enquêtes sur les ménages dans notre étude. On doit remarquer que si l'on n'a pas à tenir compte du véritable modèle de non-réponse, il n'est pas nécessaire d'utiliser un modèle de non-réponse au niveau du ménage parce que les modèles au niveau de la personne et au niveau du ménage sont équivalents. Par exemple, si I_A est vrai, (2) réduit à $\lambda^{u+v}(1-\lambda)^{1-u-v}$ et (4) se factorise en deux composantes dépendantes de ω seulement et λ seulement; on peut démontrer que le facteur dépendant de ω est l'équivalent de celui du modèle au niveau de la personne et que par conséquent les estimations des flux de la population active sont les mêmes.

Il semble, à mesure qu'augmente la taille du ménage, que le biais des estimations des flux de la population active diminue si on doit tenir compte du modèle véritable. En fait, ce résultat survient parce que nous utilisons (1) pour produire les flux de la population active et non parce que les estimations du modèle ne sont pas biaisées pour une valeur importante n_h. Pour en connaître les raisons, prenons en considération le processus de formation du ménage utilisé pour produire chaque échantillon de Monte Carlo: à mesure que n_h augmente, chaque fréquence du ménage tend vers la même valeur, c'est-à-dire, $n_h(ab)$ converge vers $n_h\omega(ab)$; par conséquent,

$$\pi(uv \mid \boldsymbol{n}_h) \to \frac{1}{n_h}\sum_{a,b} n_h\omega(a,b)\psi\ (uv \mid ab)$$

$$= \sum_{a,b} \omega(ab)\psi(uv \mid ab),$$

qui est indépendant de $\boldsymbol{n}_h$, c'est-à-dire que l'on n'a pas à tenir compte du mécanisme de non-réponse du ménage simulé. Par conséquent, les estimations des flux de la population active ne sont pas biaisées du fait que la correction des modèles dont il faut tenir compte en fonction des données simulées donne des estimations de paramètres qui sont conformes à la non-réponse dont on n'a pas à tenir

compte. Pour produire la non-réponse au niveau du ménage dont il faut tenir compte, il est nécessaire de prévenir $n_h(ab) \rightarrow n_h\omega(ab)$ en étendant (1) de façon à permettre des flux de la population active différentiels entre les ménages. Ces extensions du modèle des flux de la population active font l'objet de la section 4.

La figure 1b) illustre deux résultats anormaux qui contredisent l'explication ci-dessus lorsque N_B est le modèle véritable. Tout d'abord, le biais de l'estimation du modèle N_B au niveau de la personne augmente à mesure que n_h augmente. Cependant, d'autres simulations avec un ménage d'une taille $n_h = 10$ ont révélé que le biais de l'estimation au niveau de la personne est zéro. Ainsi, la non-réponse asymptotique dont on n'a pas à tenir compte est également évidente lorsque N_B est vrai, mais n_h doit être importante avant que son incidence devienne apparente pour le modèle N_B. Deuxièmement, le biais des estimations du modèle dont on n'a pas à tenir compte au niveau de la personne est faible, presque zéro, lorsque N_B est vrai. Ce faible biais diminue encore plus à mesure que n_h augmente, conformément à l'ignorabilité asymptotique, mais nous devons encore parvenir à une explication satisfaisante quant à savoir pourquoi les modèles dont on peut ne pas tenir compte se comparent si bien dans cette situation. Des études plus approfondies sont nécessaires pour examiner cette constatation.

4. DISCUSSION

Dans les sections 3 et 4, on a démontré grâce à une étude en simulation que la modélisation de la non-réponse dont il faut tenir compte au niveau du ménage lors de l'estimation des flux bruts de la population active provenant d'enquêtes sur les ménages entraîne une réduction du biais dans les estimations des flux, par rapport aux flux provenant des modèles au niveau de la personne. S'il s'agit d'un modèle de non-réponse dont on n'a pas à tenir compte, il n'est pas nécessaire de recourir à des modèles au niveau du ménage parce que les modèles au niveau du ménage et au niveau de la personne sont équivalents. En outre, on a démontré que le contrôle de la non-réponse au niveau du ménage n'élimine pas nécessairement tout le biais provenant des estimations des flux de la population active. La spécification correcte du modèle de non-réponse est toujours considérée impérative, quoique le fait de tenir compte de la structure des données du ménage peut entraîner un peaufinage des estimations des flux si le modèle de non-réponse est mal spécifié. En particulier, nous démontrons que les estimations au niveau du ménage sont moins biaisées que leurs estimations équivalentes au niveau de la personne.

Notre modèle de non-réponse est une extension de l'idée qui veut que la non-réponse peut dépendre des caractéristiques d'une unité, dans le présent cas les flux de la population active des membres du ménage. La non-réponse dans les enquêtes sur les ménages peut survenir pour plusieurs raisons, par ex., un refus, un non-contact, un déménagement ou un renouvellement de l'échantillon. Le modèle actuel peut facilement être étendu pour modéliser des tendances plus complexes de non-réponse en précisant l'indicateur de non-réponse comme étant une variable polytome et en paramétrant le modèle de non-réponse conformément aux tendances complexes de non-réponse. Il est également à remarquer que nous ne supposons pas que le modèle au niveau du ménage est une représentation exacte du comportement de non-réponse du ménage; plutôt, nous supposons que le modèle au niveau de la personne offre une approximation de la dynamique de non-réponse au sein du ménage.

Un problème important, mis en évidence par les résultats provenant de l'étude en simulation, est notre hypothèse que le comportement des flux de la population active individuels est homogène au sein des ménages. De toute évidence, il s'agit d'une hypothèse qui n'est pas réaliste. Le modèle est facilement étendu en précisant les flux de la population active et les probabilités de flux de non-réponse comme modèles de régression pour tenir compte des renseignements liés à la covariable au niveau de la personne, au niveau du ménage ou à un niveau plus élevé. Par exemple, les probabilités de flux de la population active pourraient être précisées comme étant une régression multinomiale-logistique:

$$\log\left(\frac{\omega_{hi}(ab)}{\omega_{hi}(11)}\right) = \beta_0^{(ab)} + \boldsymbol{\beta}_1^{(ab)}\boldsymbol{x}_{hi}^T,$$

où $\omega_{hi}(ab)$ désigne la probabilité d'une personne i dans un ménage h du flux de la population active (a, b), $\boldsymbol{x}_{hi}$ est un vecteur de covariable (rangée), et $(\beta_0^{(ab)}, \beta_1^{(ab)})$ sont les coefficients de régression pour le mutinomial-logit (a, b). Cependant, l'ajustement de ces modèles nécessite que l'on émette des hypothèses d'indépendance conditionnelles au sujet de la relation entre les distributions des covariables, les flux de la population active et les flux de non-réponse parce que les renseignements sur les covariables peuvent être manquants dans le cas des ménages qui ne répondent pas. Une autre solution est de permettre des covariables hétérogènes entre les flux de la population active du ménage, à l'aide d'effets aléatoires, en adoptant des hypothèses relativement à la distribution des différences entre les ménages. L'ajustement de ces modèles est également compliquée et nécessiterait, par exemple, une méthode de Monte Carlo de chaîne de Markov pour effectuer l'intégration nécessaire. Si S n'est pas un échantillon aléatoire simple, on peut incorporer des variables auxiliaire du plan dans le processus d'ajustement en utilisant le cadre de régression que nous venons de décrire.

REMERCIEMENTS

Le travail de Paul Clarke dans le présent article était financé grâce à une bourse d'études du Conseil de la recherche économique et sociale (prix n° R00429614273); le travail de Ray Chambers était financé grâce au contrat conclu entre l'Office for National Statistics et l'Université de Southampton pour la prestation de services de recherches en méthodologie statistique. Les deux auteurs aimeraient remercier les examinateurs, dont les observations et les conseils pratiques ont aidé à rendre la version définitive du manuscrit considérablement plus compréhensible.

BIBLIOGRAPHIE

FITZMAURICE, G.M., LAIRD, N.M., et ZAHNER, G.E.P. (1996). Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*, 91, 99-107.

HOGUE, C.R. (1985). History of the problems encountered in estimating gross flows. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 1-8.

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

LITTLE, R.J.A. (1985). Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bulletin de l'Institut International de Statistique*, 15, 1-15.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

STASNY, E.A. (1986). Estimating gross flows using panel data with nonresponse: an example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.

STASNY, E.A., et FIENBERG, S.E. (1985). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 25-39.

VANSKI, J.E. (1985). Uses of gross change data in assessing demographic labor market dynamics. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 9-12.