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

Article

L'estimation des flux bruts dans les enquêtes à base de sondage double

par Yan Lu et Sharon Lohr

Juin 2010

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2010
Vol. 36, N° 1, pp. 13-24
Statistique Canada, N° 12-001-X au catalogue

L'estimation des flux bruts dans les enquêtes à base de sondage double

Yan Lu et Sharon Lohr [1]

Résumé

Les flux bruts sont souvent utilisés pour étudier les transitions concernant la situation d'emploi ou d'autres variables catégoriques chez les individus formant une population. Dans les enquêtes longitudinales à base de sondage double, pour lesquelles des échantillons indépendants sont tirés de deux bases de sondage afin de réduire les coûts d'enquête ou d'améliorer la couverture, l'estimation efficace et cohérente des flux bruts peut poser des défis, à cause des plans de sondage complexes et des données manquantes dans l'un ou l'autre échantillon, ou les deux. Nous proposons des estimateurs des flux bruts dans les enquêtes à base de sondage double et examinons leurs propriétés asymptotiques. Puis, nous estimons les transitions entre les situations d'emploi en utilisant des données provenant de la Current Population Survey et de la Survey of Income and Program Participation.

Mots clés : Enquêtes complexes ; enquêtes à base de sondage double ; jackknife ; estimation longitudinale ; données manquantes.

1. Introduction

À l'heure actuelle, de nombreuses enquêtes sont conçues en vue de suivre les mêmes personnes à intervalle de temps régulier afin que des quantités longitudinales, telles que les transitions entre les situations d'emploi ou les situations de pauvreté puissent être étudiées. Par exemple, aux États-Unis, la Current Population Survey (CPS ; United States Census Bureau 2006) s'appuie sur un plan de sondage à panel rotatif en vertu duquel les personnes habitant l'unité de logement sélectionnée pour l'enquête sont interviewées pendant quatre mois d'affilée, cessent de l'être pendant huit mois, puis sont de nouveau interviewées pendant quatre mois consécutifs. Ce plan permet d'estimer des quantités liées aux changements qui surviennent chez les individus au cours du temps. Puisque de nombreuses réponses aux enquêtes sont catégoriques, les flux bruts, qui sont des transitions entre états d'une variable catégorique au cours du temps, sont particulièrement importants.

Le tableau 1 donne les nombres d'occurrences d'une variable catégorique mesurée à deux périodes dans une population de N unités. À la période 1, la variable peut se trouver dans l'un de r états, et à la période 2, elle peut se trouver dans l'un de c états. L'exemple qui suit illustre le tableau 1. Dans l'étude des changements de situation d'emploi, nous pourrions avoir $r = 2$ et $c = 2$, l'état 0 représentant le chômage et l'état 1, l'emploi. Alors, X_{00} donne le nombre de membres de la population qui sont en chômage aux deux périodes, X_{10} est le nombre qui sont occupés à la période 1, mais en chômage à la période 2, X_{0+} est le nombre total en chômage à la période 1, et ainsi de suite. Nous voulons calculer les estimations et les erreurs-types des flux bruts X_{kl}, $k = 0, ..., r-1$, $l = 0, ..., c-1$, en utilisant des données d'enquête. En pratique, cet exercice est parfois compliqué à cause des données manquantes et d'autres problèmes.

Même si des estimations transversales successives permettent d'évaluer une variation des taux de chômage au cours du temps, seule une enquête longitudinale permet d'étudier des questions telles que la persistance du chômage chez les individus. L'estimation des flux bruts en utilisant des données d'enquête a été étudiée par de nombreux auteurs, dont Chambers, Woyzbun et Pillig (1988), Hocking et Oxspring (1971), Blumenthal (1968), Chen et Fienberg (1974), Stasny (1984, 1987), ainsi que Stasny et Fienberg (1986). La plupart de ces travaux portaient sur des méthodes en vue d'obtenir des estimateurs du maximum de vraisemblance (EMV) pour les fréquences de cellules attendues dans les tables de contingence en se servant de données partiellement croisées. Pfeffermann, Skinner et Humphreys (1998) ont proposé des estimateurs qui tiennent compte des erreurs de classification dans les données d'enquête. Tous ces travaux s'appuient sur l'hypothèse qu'un échantillon probabiliste, habituellement un échantillon aléatoire simple, a été tiré d'une base de sondage unique.

Tableau 1
Table des flux bruts pour la population

		Période 2					
		0	**1**	**2**	**⋯**	**$c-1$**	
Période 1	**0**	X_{00}	X_{01}	X_{02}	⋯	$X_{0,\,c-1}$	X_{0+}
	1	X_{10}	X_{11}	X_{12}	⋯	$X_{1,\,c-1}$	X_{1+}
	2	X_{20}	X_{21}	X_{22}	⋯	$X_{2,\,c-1}$	X_{2+}
	⋮	⋮	⋮	⋮	⋱	⋮	⋮
	$r-1$	$X_{r-1,\,0}$	$X_{r-1,\,1}$	$X_{r-1,\,2}$	⋯	$X_{r-1,\,c-1}$	$X_{r-1,\,+}$
		X_{+0}	X_{+1}	X_{+2}	⋯	$X_{+,\,c-1}$	N

1. Yan Lu, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, 87131-0001. Courriel : yljenlu@gmail.com ; Sharon Lohr, School of Mathematical and Statistical Sciences, Arizona State University, Tempe AZ 85287-1804. Courriel : sharon.lohr@asu.edu.

Un certain nombre de programmes d'enquêtes longitudinales, telles l'Enquête longitudinale nationale auprès des enfants et des jeunes du Canada et l'Enquête par panel auprès des ménages canadiens, ont maintenant commencé à mettre en œuvre un plan à base de sondage double ou à base de sondage multiple, ou envisagent de le faire. Dans le cas d'une enquête à base de sondage multiple, les échantillons probabilistes sont tirés indépendamment de deux ou plusieurs bases de sondage. L'utilisation de plus d'une base de sondage donne souvent une meilleure couverture de la population et permet de réaliser des économies considérables dans le cas de certaines populations. Ainsi, l'Assets and Health Dynamics Survey (Heeringa 1995), dont l'objectif était d'estimer les caractéristiques de la population de 65 ans et plus, s'appuyait sur un plan à base de sondage double dans lequel la base de sondage A était celle d'une enquête nationale auprès de la population générale et la base de sondage B était une liste de personnes inscrites au régime Medicare. La structure de cette enquête est illustrée à la figure 1. La base de sondage A couvrait l'entièreté de la population, mais nécessitait une présélection de grande portée afin d'identifier les individus faisant partie de la population cible, ce qui rendait l'échantillonnage coûteux ; l'échantillonnage à partir de la base de sondage B était moins cher, mais cette base ne contenait pas l'entièreté de la population. Kalton et Anderson (1986) décrivent l'utilisation de plans à base de sondage double pour échantillonner les populations rares ; Blair et Blair (2006) soutiennent que les enquêtes à base de sondage double permettent de tirer parti des modes d'échantillonnage moins coûteux, tels que les méthodes en ligne pour échantillonner les populations rares.

Dans d'autres situations, les deux bases de sondage peuvent être incomplètes, comme l'illustre la figure 2. Hartley (1962, 1974) a été le premier à proposer des estimateurs pour le plan de sondage à base double de la figure 2, quand des échantillons indépendants sont tirés de chaque base de sondage. Des progrès subséquents sont décrits dans Bankier (1986), Fuller et Burmeister (1972), Skinner et Rao (1996), et Lohr et Rao (2000). Lohr et Rao (2006) résument les méthodes d'estimation de quantités de population dans les enquêtes transversales à base de sondage multiple.

Dans le présent article, nous proposons des estimateurs des flux bruts qui peuvent être appliqués aux enquêtes à base de sondage double dans lesquelles l'information longitudinale est recueillie auprès de l'un des échantillons ou des deux. Les unités échantillonnées dans l'une des enquêtes ou dans les deux sont suivies au cours du temps ; dans certains cas, des unités supplémentaires sont échantillonnées à des périodes ultérieures afin d'intégrer de nouvelles unités de population ou de compenser l'érosion de l'échantillon. Une enquête longitudinale à base de sondage double ajoute de nouveaux défis à ceux que posent les enquêtes longitudinales à base de sondage unique ou les enquêtes transversales à base de sondage double. Des données peuvent manquer dans l'échantillon de chacune des bases de sondage et des unités peuvent voir leur appartenance passer d'une base de sondage à l'autre entre les interviews de l'enquête. En outre, les deux plans d'échantillonnage peuvent être complexes, présentant une stratification et une mise en grappes. Dans le cas d'une enquête à deux bases de sondage chevauchantes, comme celles illustrée à la figure 2, l'objectif est d'utiliser aussi efficacement que possible l'information contenue dans la zone de chevauchement. Dans le présent article, nous étudions le problème consistant à utiliser toute l'information tirée de la base de sondage A et de la base de sondage B pour estimer les probabilités de transition au sein de la population.

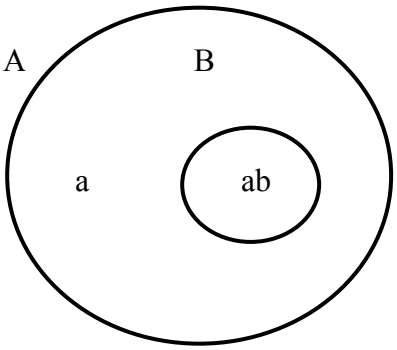

Figure 1 La base de sondage B est un sous-ensemble de la base de sondage A

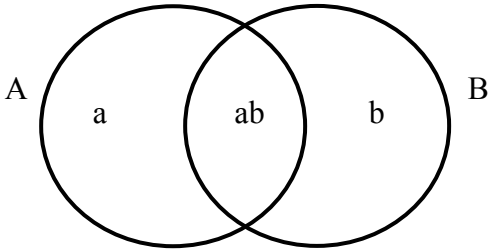

Figure 2 Les bases de sondage A et B sont toutes deux incomplètes, mais chevauchantes

La présentation de l'article est la suivante. À la section 2, nous exposons le problème de recherche. À la section 3, nous dérivons les estimateurs des flux bruts dans les enquêtes à base de sondage double pour des échantillons complexes pouvant présenter des données manquantes. À la section 4, nous établissons les propriétés asymptotiques et discutons de l'estimation de la variance. À la section 5, nous décrivons une application de notre recherche à la Current

Population Survey et à la Survey of Income and Program Participation. Enfin, à la section 6, nous présentons nos conclusions.

2. Notation et quantités dans les échantillons

Supposons qu'il existe deux bases de sondage, A et B, qui ensemble couvrent la population d'intérêt $A \cup B$ comme l'illustre la figure 2. En utilisant la notation de Hartley (1962), il existe trois domaines non chevauchants : $a = A \cap B^c$, $b = A^c \cap B$ et $ab = A \cap B$, où c désigne le complément d'un ensemble. Les tailles de population des bases de sondage A et B sont N_A et N_B, et les tailles de population des domaines sont N_a, N_b et N_{ab}. Nous supposons que N_A et N_B sont connues, mais que la taille de population $N = N_A + N_B - N_{ab}$ peut être inconnue. Dans le présent article, nous supposons que la population, ainsi que les bases de sondage ne varient pas au cours du temps. Il s'agit d'hypothèses fortes, mais dans de nombreuses enquêtes longitudinales, la population d'intérêt et les bases de sondage peuvent être définies pour la période 1.

Supposons, à la présente section, que l'appartenance à un domaine est constante au cours du temps. Pour simplifier la notation, nous supposons ici que $r = 2$ et $c = 2$, de sorte qu'il existe deux catégories possibles à chaque période ; le cas général est similaire. Puisque les trois domaines sont non chevauchants, chaque dénombrement de population X_{kl}, $k = 0, 1$, $l = 0, 1$, peut s'écrire $X_{kl} = X_{kla} + X_{klab} + X_{klb}$, où X_{kld} est le nombre d'unités de la population du domaine d qui se trouve dans l'état k à la période 1 et dans l'état l à la période 2. Les probabilités de population et de domaine correspondantes sont $p_{kl} = X_{kl}/N$ et $p_{kld} = X_{kld}/N_d$ pour $d \in \{a, ab, b\}$.

Nous tirons des échantillons probabilistes indépendants, S_A et S_B, de tailles n_A et n_B, des bases de sondage A et B. Soit w_i^A le poids de l'unité échantillonnée i pour l'échantillon tiré de la base de sondage A et soit w_j^B, le poids de l'unité échantillonnée j pour l'échantillon tiré de la base de sondage B. Nous pouvons donner à w_i^A la forme d'un poids d'échantillonnage $[P(i \in S_A)]^{-1}$ ou d'un poids de type Hájek $[P(i \in S_A)]^{-1} N_A/$ (somme des poids d'échantillonnage dans S_A). D'autres scénarios de pondération pour les données longitudinales, discutés dans Verma, Betti et Ghellini (2007) et dans Lavallée (2007), pourraient également être utilisés. Soit $\mathbf{y}_i = (y_{i1}, y_{i2})$ la réponse de l'unité i dans S_A, avec $y_{i1}, y_{i2} \in \{0, 1, M\}$, où M indique que la valeur est manquante. Alors, $\hat{X}^A_{kla} = \sum_{i \in S_A} w_i^A \, I(y_{i1} = k) \, I(y_{i2} = l) \, I(i \in a)$ et $\hat{X}^A_{klab} = \sum_{i \in S_A} w_i^A \, I(y_{i1} = k) \, I(y_{i2} = l) \, I(i \in ab)$ estiment les dénombrements de population pour la cellule (k, l) dans les domaines a et ab provenant de S_A, pour $k, l \in \{0, 1, M\}$. Soit $\mathbf{y}_j = (y_{j1}, y_{j2})$ la réponse de l'unité j dans S_B, et soit $\hat{X}^B_{klb} = \sum_{j \in S_B} w_j^B \, I(y_{j1} = k) \, I(y_{j2} = l) \, I(j \in b)$ et $\hat{X}^B_{klab} = \sum_{j \in S_B} w_j^B \, I(y_{j1} = k) \, I(y_{j2} = l) \, I(j \in ab)$ les estimateurs correspondants provenant de S_B.

Dans le présent article, nous supposons que l'appartenance à un domaine peut être déterminée pour chaque unité de l'échantillon et que les réponses $\mathbf{y}_i$ ne contiennent pas d'erreur de classification. Donc, nous supposons que nous savons si chaque unité de l'échantillon tiré de la base de sondage A ou de la base de sondage B appartient à l'autre base de sondage ou non. Nous supposons aussi que $\mathbf{y}_i$ et $\mathbf{y}_j$ sont mesurés sans erreur – dans l'exemple de l'emploi, cela signifie que chaque répondant donne la bonne réponse concernant sa situation d'emploi. Donc, les méthodes proposées ici sont sensibles à l'erreur de classification des observations dans les domaines et dans les cellules. Si les moyennes de domaine diffèrent ou si les observations sont classées incorrectement, les estimateurs des flux bruts pourraient présenter un biais ; Pfeffermann et coll. (1998) discutent des méthodes permettant de tenir compte des erreurs de classification dans les enquêtes à base de sondage unique.

Les estimateurs dérivés de S_A sont présentés au tableau 2. Un tableau similaire peut être conçu pour les estimateurs dérivés de S_B. Nous supposons que chaque unité est échantillonnée durant l'une des périodes ou les deux. En l'absence de données manquantes, tous les dénombrements estimés pour les cellules (k, M) et (M, l) sont nuls. En utilisant l'absence exacte ou approximative de biais des estimateurs, selon que l'on se sert des poids d'échantillonnage ou des poids de Hájek, en l'absence de données manquantes, $E[\hat{X}^A_{kla}] \approx X_{kla}$, $E[\hat{X}^A_{klab}] \approx E[\hat{X}^B_{klab}] \approx X_{klab}$ et $E[\hat{X}^B_{klb}] \approx X_{klb}$.

Tableau 2
Estimateurs dérivés de l'échantillon tiré de la base de sondage *A*

			Période 2			
			0	1	Val. Manquante	
Période 1	Domaine *a*	0	$\hat{X}^A_{00a}$	$\hat{X}^A_{01a}$	$\hat{X}^A_{0Ma}$	$\hat{X}^A_{0+a}$
		1	$\hat{X}^A_{10a}$	$\hat{X}^A_{11a}$	$\hat{X}^A_{1Ma}$	$\hat{X}^A_{1+a}$
		Val. Manquante	$\hat{X}^A_{M0a}$	$\hat{X}^A_{M1a}$		$\hat{X}^A_{M+a}$
	Domaine *ab*	0	$\hat{X}^A_{00ab}$	$\hat{X}^A_{01ab}$	$\hat{X}^A_{0Mab}$	$\hat{X}^A_{0+ab}$
		1	$\hat{X}^A_{10ab}$	$\hat{X}^A_{11ab}$	$\hat{X}^A_{1Mab}$	$\hat{X}^A_{1+ab}$
		Val. Manquante	$\hat{X}^A_{M0ab}$	$\hat{X}^A_{M1ab}$		$\hat{X}^A_{M+ab}$
			$\hat{X}^A_{+0}$	$\hat{X}^A_{+1}$	$\hat{X}^A_{+M}$	$\hat{N}_A$

3. Estimateurs des flux bruts dans les enquêtes à base de sondage double

À la présente section, nous dérivons des estimateurs des flux bruts pour des échantillons complexes dans des

enquêtes à base de sondage double. Nous suivons une approche de pseudo-vraisemblance à base de sondage double pour tenir compte des plans d'échantillonnage et des mécanismes de génération des données manquantes. L'approche à base de sondage double permet d'améliorer la précision des estimateurs et offre plus de souplesse pour modéliser les mécanismes susmentionnés. Les méthodes utilisées à l'heure actuelle pour traiter les données manquantes sont fondées sur des méthodes statistiques classiques et entrent dans quatre catégories générales (Little et Rubin 2002) : l'analyse des cas complets, les méthodes de pondération, les méthodes d'imputation et les méthodes fondées sur un modèle. Ici, nous adoptons une approche fondée sur un modèle pour traiter les données manquantes. À la présente section, nous considérons des conditions simples, c'est-à-dire des échantillons aléatoires simples tirés d'une population sans données manquantes. Puis, nous ajoutons un modèle pour le mécanisme de génération de données manquantes. Enfin, nous discutons d'estimateurs applicables à des plans de sondage plus complexes.

3.1 Échantillons aléatoires simples avec données complètes

Pour justifier l'estimateur utilisé dans le cas général, nous commençons par étudier l'estimation des flux bruts quand il n'existe pas de données manquantes et quand l'échantillon tiré de chaque base de sondage est un échantillon aléatoire simple. Alors, $x^A_{kld} = n_A \hat{X}^A_{kld}/N_A$, pour $d = a, ab$ est le nombre observé d'unités échantillonnées dans la cellule kl et le domaine d provenant de S_A ; $x^B_{kld} = n_B \hat{X}^B_{kld}/N_B$ pour $d = b, ab$ est le nombre observé correspondant d'unités échantillonnées provenant de S_B.

Si les fractions d'échantillonnage sont faibles, nous pouvons utiliser une approximation multinomiale de la vraisemblance. Dans le cas de l'échantillon tiré de la base de sondage A, il existe huit cellules dont les probabilités associées sont $P^A_{kld} = p_{kld} N_d/N_A$, pour $k, l \in \{0, 1\}$ et $d \in \{a, ab\}$. Les probabilités correspondantes pour l'échantillon tiré de la base de sondage B sont $P^B_{kld} = p_{kld} N_d/N_B$ pour $k, l \in \{0, 1\}$ et $d \in \{b, ab\}$. En utilisant la loi multinomiale et en supposant que les échantillons provenant des deux bases de sondage sont tirés indépendamment, la fonction de vraisemblance est donnée par

$$L(\mathbf{p}, N_{ab}) \propto \prod_{k,l,d} (P^A_{kld})^{x^A_{kld}} \times \prod_{k,l,d} (P^B_{kld})^{x^B_{kld}}.$$

Pour simplifier, nous écrivons la vraisemblance en fonction de P^A_{kld} et P^B_{kld}, mais les paramètres d'intérêt sous-jacents sont $\mathbf{p} = (p_{00a}, p_{01a}, ..., p_{11b})$ et N_{ab}.

En posant que les dérivées partielles de la log-vraisemblance par rapport aux paramètres sont nulles, les estimateurs du maximum de vraisemblance sont données par $\hat{p}_{kla} = x_{kla}/n^A_a$, $\hat{p}_{klb} = x_{klb}/n^B_b$ et $\hat{p}_{klab} = (x^A_{klab} + x^B_{klab})/(n^A_{ab} + n^B_{ab})$, où $n^A_{ab} = \sum_{i \in S_A} I(i \in ab)$, $n^B_{ab} = \sum_{j \in S_B} I(j \in ab)$, $n^A_a = n_A - n^A_{ab}$ et $n^B_b = n_B - n^B_{ab}$. L'EMV pour N_{ab}, $\hat{N}_{ab}$, est la plus petite racine de l'équation quadratique

$$[n_A + n_B]\, \hat{N}^2_{ab} - [n_A N_B + n_B N_A + n^A_{ab} N_A + n^B_{ab} N_B]\, \hat{N}_{ab} + [n^A_{ab} + n^B_{ab}]\, N_A N_B = 0. \quad (1)$$

Enfin, en utilisant les résultats susmentionnés, nous construisons les EMV pour X_{kl} et p_{kl} :

$$\hat{X}_{kl} = (N_A - \hat{N}_{ab})\, \hat{p}_{kla} + \hat{N}_{ab} \hat{p}_{klab} + (N_B - \hat{N}_{ab}) \hat{p}_{klb}.$$

$$\hat{p}_{kl} = \frac{(N_A - \hat{N}_{ab})\, \hat{p}_{kla} + \hat{N}_{ab} \hat{p}_{klab} + (N_B - \hat{N}_{ab})\, \hat{p}_{klb}}{N_A + N_B - \hat{N}_{ab}}.$$

Ces estimateurs sont les mêmes que ceux obtenus par Skinner (1991). Cependant, ce dernier a utilisé la loi normale approximative de la moyenne des réponses $\bar{y}$ dans chaque domaine pour obtenir les EMV, tandis que nos estimateurs proviennent d'un modèle multinomial. Ce modèle nous permet d'inclure des données partiellement classifiées provenant des unités observées durant une seule période, comme nous le montrons à la section suivante.

3.2 Échantillons aléatoires simples avec données manquantes

En pratique, certaines personnes peuvent n'apparaître dans l'échantillon qu'à l'une des deux périodes seulement. Cette situation peut être due à l'érosion de l'échantillon (quand des membres de l'échantillon cessent de participer à l'étude) ou à d'autres causes. Dans une enquête à panel rotatif, telle que la CPS, les personnes qui sortent de l'échantillon à la période 1 ne seront pas interrogées à la période 2 et, par conséquent, leur situation d'emploi durant cette période sera inconnue. Dans d'autres situations, l'un des échantillons peut être transversal, auquel cas toutes les observations sont mesurées exactement à une seule période.

3.2.1 Modèle pour les données manquantes

Blumenthal (1968), Chen et Fienberg (1974), Stasny (1984, 1987), et Stasny et Fienberg (1986) ont utilisé une procédure à deux phases pour modéliser les données manquantes dans un échantillon unique. Un modèle est d'abord proposé pour les données complètes, puis le mécanisme de génération des données manquantes est modélisé. Nous étendons cette procédure à nos structures à base de sondage double. L'un des avantages d'une enquête à base de sondage double est qu'elle offre plus de souplesse pour la modélisation des données manquantes.

Premièrement, nous supposons que si toutes les unités étaient mesurées aux deux périodes, le modèle de la section 3.1 pourrait être utilisé. Pour le mécanisme de non-réponse, nous supposons que chaque observation dans la cellule (k, l) et le domaine d provenant de S_A a la probabilité

ϕ^A_{kld} de manquer à la période 1 et la probabilité ψ^A_{kld} de manquer à la période 2. Nous supposons aussi que l'unité ne peut pas manquer aux deux périodes à la fois.

Cette formulation repose sur l'hypothèse que la probabilité qu'une observation manque dans une cellule, un domaine et une base de sondage particuliers est constante. Si des données pouvaient manquer pour d'autres raisons, des paramètres supplémentaires pourraient être utilisés pour faire la distinction entre les observations dont la classification est partielle, à cause, disons, du plan à panel rotatif et celles dont la classification est partielle à cause de la non-réponse. À la section 5, nous discutons d'une autre approche qui pourrait être utilisée avec des mécanismes multiples de génération des données manquantes.

Pour $k, l \in \{0, 1\}$, la probabilité qu'une unité provenant de S_A soit observée dans la cellule (k, l) et le domaine d est

$$Q^A_{kld} = P^A_{kld}(1 - \phi^A_{kld} - \psi^A_{kld}).$$

La probabilité qu'une unité provenant de S_A soit observée dans la cellule (k, M) et le domaine d est

$$Q^A_{kMd} = \sum_{l=0}^{1} P^A_{kld}\psi^A_{kld}.$$

De même, la probabilité qu'une unité provenant de S_A soit observée dans la cellule (M, l) et le domaine d est

$$Q^A_{Mld} = \sum_{k=0}^{1} P^A_{kld}\phi^A_{kld}.$$

Nous définissons de la même manière les probabilités pour la base de sondage B, soit $Q^B_{kld} = P^B_{kld}(1 - \phi^B_{kld} - \psi^B_{kld})$, $Q^B_{kMd} = \sum_{l=0}^{1} P^B_{kld}\psi^B_{kld}$ et $Q^B_{Mld} = \sum_{k=0}^{1} P^B_{kld}\phi^B_{kld}$.

Sous ce modèle à deux phases et en émettant l'hypothèse d'indépendance des échantillons, la fonction de vraisemblance pour les deux échantillons est donnée par :

$$\begin{aligned} L(\mathbf{p}, \boldsymbol{\psi}, \boldsymbol{\phi}, N_{ab}) \propto & \prod_{k\in\{0,1\}} \prod_{l\in\{0,1\}} \prod_{d\in\{a,\,ab\}} (Q^A_{kld})^{x^A_{kld}} \\ & \times \prod_{k\in\{0,1\}} \prod_{l\in\{0,1\}} \prod_{d\in\{b,\,ab\}} (Q^B_{kld})^{x^B_{kld}} \\ & \times \prod_{k\in\{0,1\}} \prod_{d\in\{a,\,ab\}} (Q^A_{kMd})^{x^A_{kMd}} \\ & \times \prod_{l\in\{0,1\}} \prod_{d\in\{a,\,ab\}} (Q^A_{Mld})^{x^A_{Mld}} \\ & \times \prod_{k\in\{0,1\}} \prod_{d\in\{b,\,ab\}} (Q^B_{kMd})^{x^B_{kMd}} \\ & \times \prod_{l\in\{0,1\}} \prod_{d\in\{b,\,ab\}} (Q^B_{Mld})^{x^B_{Mld}}, \end{aligned} \qquad (2)$$

où $\boldsymbol{\psi}$ est le vecteur des ψ^A_{kld} et des ψ^B_{kld}, et $\boldsymbol{\phi}$ est le vecteur des ϕ^A_{kld} et des ϕ^B_{kld}.

L'expression (2) correspond au modèle le plus général, dans lequel les deux enquêtes sont longitudinales et présentent toutes deux des données manquantes à chaque période. Si la base de sondage A est utilisée dans une enquête à panel rotatif, par exemple, alors toutes les probabilités Q^A_{kld} sont non nulles : les unités comprises dans les panels mesurés aux deux périodes seront incluses dans les estimateurs x^A_{kld} pour $k, l \in \{0, 1\}$, les unités dans les panels sortant de l'enquête après la période 1 seront incluses dans les estimateurs x^A_{kMd}, et les unités figurant dans les panels entrants seront incluses dans les estimateurs x^A_{Mld}. Selon la structure des enquêtes, certains facteurs de l'expression (2) peuvent être omis. Par exemple, si l'enquête s'appuyant sur la base de sondage B est une enquête transversale répétée dont la fraction d'échantillonnage est faible, les probabilités Q^B_{kld} pour $k, l \in \{0, 1\}$ seront presque nulles et nous omettrons ces facteurs dans l'expression de la vraisemblance.

La vraisemblance donnée par (2) peut s'écrire sous la forme du produit d'un facteur contenant N_{ab} et d'un facteur contenant les autres paramètres. Par conséquent, l'EMV pour N_{ab} est de nouveau la plus petite racine de l'équation (1). Nous discutons des estimateurs des paramètres restants à la section suivante.

3.2.2 Identificabilité du modèle et modèles réduits

L'une des difficultés que pose la maximisation de la vraisemblance donnée par (2) est que, sous le modèle général, il existe un total de 42 paramètres, tandis que les deux échantillons ne comprennent que 32 dénombrements de cellule observés. Donc, nous ne pouvons pas estimer tous les paramètres sous le modèle le plus général. Toutefois, nous pouvons envisager des modèles dont le nombre de paramètres est réduit, comme l'ont fait Chen et Fienberg (1974) pour les enquêtes à base de sondage unique. En fait, le cas de la base de sondage double donne nettement plus de souplesse pour la modélisation des données manquantes, grâce à l'information indépendante provenant des deux échantillons au sujet du domaine ab.

Nous commençons par énoncer les conditions pour qu'un modèle réduit soit localement identifiable. Soit $\boldsymbol{\theta}$ le vecteur de dimension s des paramètres d'intérêt ; dans notre cas, $\boldsymbol{\theta}$ comprend les composantes linéairement indépendantes de $\mathbf{p}$, N_{ab} / N et les paramètres pour le mécanisme de génération des données manquantes. Dans l'expression de la vraisemblance (2), les probabilités provenant des échantillons multinomiaux indépendants sont Q^A_{kld} et Q^B_{kld}. Ces probabilités pourraient s'écrire sous la forme de fonctions de $\boldsymbol{\theta}$, avec $\mathbf{Q}^A(\boldsymbol{\theta}) = (Q^A_{00a}, ..., Q^A_{lMab})$, un vecteur de dimension g des Q^A_{kld} non nulles et $\mathbf{Q}^B(\boldsymbol{\theta}) = (Q^B_{00b}, ..., Q^B_{lMab})$, un vecteur de dimension q des Q^B_{kld} non nulles. Quand toutes les cellules du tableau 2 et celles du tableau analogue

pour la base de sondage B ont des probabilités non nulles, $g = q = 16$. Soit $\mathbf{D} = (\mathbf{D}'_A, \mathbf{D}'_B)'$ la matrice des dérivées de la transformation, avec $\mathbf{D}_{A(\alpha\beta)} = \partial \mathbf{Q}^A_\alpha / \partial \theta_\beta$ et $\mathbf{D}_{B(\delta\beta)} = \partial \mathbf{Q}^B_\delta / \partial \theta_\beta$ pour $\alpha = 1, ..., g-1$, $\delta = 1, ..., q-1$, et $\beta = 1, ..., s$. Alors, en utilisant les théorèmes 3, 4 et 5 de Catchpole et Morgan (1997), le modèle est localement identifiable si la matrice $\mathbf{D}$ est de plein rang. La preuve pour le cas d'une base de sondage double est donnée dans Lu (2007).

Dans le cas d'une enquête à base de sondage double, nous considérons deux types de modèles pour les données manquantes. Dans un modèle de type (1), la probabilité que l'information manque à la période 1 ou à la période 2 pour la cellule (k, l) est la même pour chaque domaine dans une base de sondage, c'est-à-dire $\phi^A_{kla} = \phi^A_{klab} = \phi_{klA}$, $\psi^A_{kla} = \psi^A_{klab} = \psi_{klA}$, $\phi^B_{klb} = \phi^B_{klab} = \phi_{klB}$ et $\psi^B_{klb} = \psi^B_{klab} = \psi_{klB}$. Dans ce type de modèle, nous estimons les ϕ et les ψ séparément pour chaque échantillon. Ce modèle pourrait être pris en considération quand les données sur les échantillons provenant des deux bases de sondage sont recueillies en utilisant des modes de collecte différents. Par exemple, si l'échantillon tiré de la base de sondage A est celui d'une enquête par la poste et l'échantillon tiré de la base de sondage B, celui d'une enquête par téléphone mobile, on pourrait s'attendre à des probabilités différentes d'abandon pour les deux échantillons.

Dans un modèle de type (2), les probabilités qu'il existe des données manquantes sont les mêmes dans chaque domaine, c'est-à-dire $\phi^A_{klab} = \phi^B_{klab} = \phi_{klab}$. Ce type de modèle pourrait être envisagé si l'on s'attend à ce que la non-réponse soit reliée à l'appartenance à une cellule et que l'on pense que l'appartenance à une base de sondage a peu d'effet sur la non-réponse. Par exemple, si les deux enquêtes s'appuient sur des plans de sondage et des procédures administratives de même type, le choix d'un modèle de type (2) pourrait être approprié.

Pour chaque type de modèle, nous pourrions devoir imposer des contraintes supplémentaires sur les paramètres afin de résoudre les équations de vraisemblance. En nous inspirant de Stasny et Fienberg (1986), les contraintes qui suivent sont possibles :

$$\text{Modèle 1} : \phi_{kl} = \lambda_{t-1(l)},\ \psi_{kl} = \lambda_{t(k)} \qquad (3)$$

$$\text{Modèle 2} : \phi_{kl} = \lambda_{t-1},\ \psi_{kl} = \lambda_t$$

$$\text{Modèle 3} : \phi_{kl} = \lambda_l,\ \psi_{kl} = \lambda_k$$

$$\text{Modèle 4} : \phi_{kl} = \lambda_{t-1(l)},\ \psi_{kl} = \lambda_t$$

$$\text{Modèle 5} : \phi_{kl} = \lambda_{t-1},\ \psi_{kl} = \lambda_{t(k)}.$$

Sous le modèle 1, la probabilité qu'une personne ne réponde pas à une période donnée dépend de cette période et de la classification de la personne durant la période observée. Sous le modèle 2, la probabilité qu'une personne ne réponde pas à une période donnée dépend uniquement de la période en question. Sous le modèle 3, la probabilité qu'une personne ne réponde à pas à une période donnée dépend uniquement de la classification de la personne durant la période observée. Sous le modèle 4, la probabilité qu'une personne ne réponde pas à la période 1 dépend de cette période et de la classification de la personne durant le mois observé, et la probabilité qu'une personne ne réponde pas à la période 2 dépend uniquement de cette période 2. Sous le modèle 5, la probabilité qu'une personne ne réponde pas à la période 1 dépend uniquement de la période, et la probabilité qu'une personne ne réponde pas à la période 2 dépend de la période et de la classification de la personne durant le mois observé. Pour chaque type, de nombreux autres modèles que les cinq susmentionnés sont possibles. En utilisant les matrices des dérivées, il est facile de montrer que les modèles 1 à 5 sont tous identifiables.

En général, il n'existe pas de solution analytique pour les estimations des paramètres, qui doivent donc être obtenues par une méthode itérative. Nous utilisons la fonction « nlm » de R (www.r-project.org) pour calculer les estimations des paramètres ; le code peut être obtenu auprès des auteurs.

3.3 Estimateurs pour les échantillons complexes

Quand les données sur l'un des échantillons ou les deux sont recueillies selon un plan de sondage complexe, l'utilisation des dénombrements de cellule directement dans l'expression de la vraisemblance (2) donne des estimateurs qui ne sont pas convergents sous le plan. Skinner et Rao (1996) ont utilisé une méthode du pseudo-maximum de vraisemblance (PMV) pour obtenir de estimateurs convergents sous le plan dans le cas des enquêtes transversales à base de sondage double. Ils ont montré que, contrairement aux estimateurs de Hartley (1962) et de Fuller et Burmeister (1972), le même ensemble de poids modifiés était utilisé dans les estimateurs du PMV (EPMV) de diverses variables de réponse et que ces estimateurs étaient donc intérieurement convergents.

Nous proposons d'étudier des estimateurs inspirés de la méthode PMV pour l'estimation des flux bruts dans les enquêtes longitudinales complexes à base de sondage double qui permettent que des données manquent à l'une ou à l'autre période dans l'un ou l'autre échantillon. L'idée fondamentale consiste à émettre l'hypothèse de travail d'une loi multinomiale issue d'une population finie pour donner la forme des estimateurs et à utiliser un effet de plan pour corriger les dénombrements de cellule afin qu'ils reflètent le plan de sondage complexe.

Dans le cas de l'échantillonnage aléatoire simple, x^A_{kld}/n_A est un estimateur convergent sous le plan de Q^A_{kld}.

Pour obtenir une pseudo-vraisemblance pour des plans de sondage généraux, nous remplaçons x^A_{kld}/n_A par $\hat{X}^A_{kld}/N_A$, un estimateur de Q^A_{kld} convergent sous le plan de sondage complexe, dans l'expression (2) de la vraisemblance. Définissons $\overline{x}^A_{kld} = \overline{n}_A \hat{X}^A_{kld}/N_A$ et $\overline{x}^B_{kld} = \overline{n}_A \hat{X}^B_{kld}/N_B$. En suivant Skinner et Rao (1996), nous permettons que $\overline{n}_A$ et $\overline{n}_B$ soient des constantes arbitraires. Notons que, si N_A ou N_B est inconnu, il doit être estimé par $\hat{N}_A$ ou $\hat{N}_B$.

La pseudo-vraisemblance a la même forme que (2), avec x^A_{kld}, x^B_{kld}, n_A et n_B remplacés par $\overline{x}^A_{kld}$, $\overline{x}^B_{kld}$, $\overline{n}_A$ et $\overline{n}_B$, respectivement. Nous utilisons ensuite des procédures itératives pour trouver les EPMV des quantités d'intérêt p_{kld}, $\boldsymbol{\phi}$, $\boldsymbol{\psi}$ et N_{ab}. Par suite des facteurs de pseudo-vraisemblance, nous trouvons que $\hat{N}_{ab}$ est égal à la plus petite racine de

$$[\overline{n}_A + \overline{n}_B]\,\hat{N}^2_{ab,\,PMV} - [\overline{n}_A N_B + \overline{n}_B N_A + \overline{n}_A \hat{N}^A_{ab} + \overline{n}_B \hat{N}^B_{ab}]\,\hat{N}_{ab,\,PMV} + [\overline{n}_A \hat{N}^A_{ab} N_B + \overline{n}_B \hat{N}^B_{ab} N_A] = 0. \quad (4)$$

Dans une enquête complexe, particulièrement en cas de mise en grappes, les tailles d'échantillon réelles n_A et n_B ne reflètent pas nécessairement les quantités relatives d'information provenant des échantillons. Nous suggérons donc de donner pour valeur à $\overline{n}_A$ et $\overline{n}_B$ la taille d'échantillon effective pour chaque échantillon, avec $\overline{n}_A = n_A/$ (effet de plan de S_A) et $\overline{n}_B = n_B/$ (effet de plan de S_B). L'effet de plan d'un estimateur $\hat{\mu}$ est le ratio

$$\frac{[V(\hat{\mu}) \text{ provenant du plan de sondage complexe}]}{[V(\hat{\mu}) \text{ provenant de l'EAS de même taille}]}.$$

L'effet de plan varie habituellement selon la variable. Toutefois, pour estimer les flux bruts, les seuls estimateurs provenant des enquêtes utilisées sont ceux des dénombrements de cellule, et nous pourrions nous attendre à ce que, pour de nombreuses enquêtes, les effets pour les estimateurs $\hat{X}^A_{kld}$ soient tous les mêmes et qu'ils soient aussi semblables à l'effet de plan de l'estimateur $\hat{N}^A_{ab}$. Donc, à l'instar de Skinner et Rao (1996), nous suggérons d'utiliser l'effet de plan pour l'estimateur $\hat{N}^A_{ab}$ pour déterminer $\overline{n}_A$, et l'effet de plan pour l'estimateur $\hat{N}^B_{ab}$ pour déterminer $\overline{n}_B$. Si les effets de plan des autres variables sont effectivement identiques, alors les EPMV résultants minimiseront les variances des quantités estimées ; s'ils diffèrent, les EPMV ne seront pas optimaux, mais ils seront convergents et, dans la plupart des cas, proches des valeurs optimales (Lohr et Rao 2006). Si l'effet de plan pour $\hat{N}^A_{ab}$ n'est pas connu, comme cela se produirait, par exemple, si l'enquête était poststratifiée en se basant sur N^A_{ab}, alors nous suggérons d'utiliser un effet de plan généralisé, calculé en prenant une moyenne ou une moyenne pondérée des effets de plan d'autres variables de l'enquête.

4. Propriétés des estimateurs

À la présente section, nous examinons les propriétés des estimateurs. Nous calculons les variances asymptotiques, discutons des estimateurs de variance jackknife et exécutons une petite étude en simulations pour explorer les propriétés.

4.1 Propriétés

Nous considérons le cas général dans lequel des échantillons stratifiés à plusieurs degrés sont tirés de chaque base de sondage. Les estimateurs des totaux de population sont les estimateurs classiques de Horvitz-Thompson ou de Hájek pour les enquêtes complexes. À partir de la base de sondage A, nous estimons le vecteur de paramètres $\boldsymbol{\eta}_A = [(\mathbf{Q}^A)', N_{ab}/N_A]'$ au moyen de $\hat{\boldsymbol{\eta}}_A = [(\hat{\mathbf{Q}}^A)', \hat{N}^A_{ab}/N_A]'$, où $\hat{Q}^A_{kld} = \hat{X}^A_{kld}/N_A$; de même, nous estimons $\boldsymbol{\eta}_B = [(\mathbf{Q}^B)', N_{ab}/N_B]'$ au moyen de $\hat{\boldsymbol{\eta}}_B = [(\hat{\mathbf{Q}}^B)', \hat{N}^B_{ab}/N_B]'$ avec $\hat{Q}^B_{kld} = \hat{X}^B_{kld}/N_B$.

Théorème 1 : Soit $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\eta}}'_A, \hat{\boldsymbol{\eta}}'_B)'$ et $\boldsymbol{\eta} = (\boldsymbol{\eta}'_A, \boldsymbol{\eta}'_B)'$. Supposons que les conditions de régularité imposées aux probabilités d'inclusion dans Isaki et Fuller (1982) sont vérifiées pour chaque échantillon. Soit $\tilde{n}_A$ et $\tilde{n}_B$ le nombre d'unités primaires d'échantillonnage dans les bases de sondage A et B, respectivement, et soit $\tilde{n} = \tilde{n}_A + \tilde{n}_B$. Supposons que $\tilde{n}_A$ et $\tilde{n}_B$ augmentent tous deux de telle façon que $\tilde{n}_A/\tilde{n}_B \to \gamma$ pour une certaine valeur $0 < \gamma < 1$. Alors, $\hat{\boldsymbol{\eta}}$ converge vers $\boldsymbol{\eta}$, et

$$\tilde{n}^{1/2}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}), \quad (5)$$

où $\boldsymbol{\Sigma}$ est une matrice diagonale par blocs dont les blocs sont $\boldsymbol{\Sigma}_A$ et $\boldsymbol{\Sigma}_B$, $\boldsymbol{\Sigma}_A$ est la matrice de covariance asymptotique de $\tilde{n}^{1/2}\hat{\boldsymbol{\eta}}_A$ et $\boldsymbol{\Sigma}_B$ est la matrice de covariance asymptotique de $\tilde{n}^{1/2}\hat{\boldsymbol{\eta}}_B$. Si, en outre, nous supposons que $N_{ab}/N \to \kappa$ pour une certaine valeur $0 < \kappa < 1$ et que le modèle est identifiable, alors $\hat{\boldsymbol{\theta}}$ converge vers $\boldsymbol{\theta}$, où $\boldsymbol{\theta}$, le paramètre d'intérêt, est constitué des composantes de $\mathbf{p}$, N_{ab}/N, $\boldsymbol{\phi}$ et $\boldsymbol{\psi}$, et $\hat{\boldsymbol{\theta}}$ est l'estimateur du pseudo-maximum de vraisemblance de $\boldsymbol{\theta}$. De surcroît, $\tilde{n}^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ est asymptotiquement normal de moyenne 0 et de variance asymptotique $\mathbf{H}_A\boldsymbol{\Sigma}_A\mathbf{H}'_A + \mathbf{H}_B\boldsymbol{\Sigma}_B\mathbf{H}'_B$, où $\mathbf{H}_F$ est la matrice des dérivées de la fonction $\boldsymbol{\theta}$ par rapport aux paramètres $\boldsymbol{\eta}_F$ pour les bases de sondage $F \in \{A, B\}$.

Démonstration. Dans le cas des flux bruts, les valeurs observées de toutes les variables sont égales à 0 ou 1. Donc, les conditions de bornage figurant dans les lemmes 1 et 2 de Isaki et Fuller (1982) sont satisfaites, et les estimateurs pour la base de sondage A sont convergents et asymptotiquement normaux avec

$$\tilde{n}^{1/2}_A(\hat{\boldsymbol{\eta}}_A - \boldsymbol{\eta}_A) \xrightarrow{d} N[0, (\gamma/(1+\gamma))\boldsymbol{\Sigma}_A].$$

Le même argument s'applique pour prouver la convergence et la normalité asymptotique du vecteur d'estimateurs provenant de la base de sondage B, avec

$$\tilde{n}_B^{1/2}(\hat{\boldsymbol{\eta}}_B - \boldsymbol{\eta}_B) \xrightarrow{d} N[0, (1 - (\gamma/(1+\gamma)))\, \boldsymbol{\Sigma}_B].$$

En combinant ces deux résultats asymptotiques et en utilisant l'indépendance des plans de sondage en même temps que le théorème de Slutsky, nous obtenons (5). La loi limite de $\tilde{n}^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ s'ensuit par la méthode delta, puisque les paramètres dans $\boldsymbol{\theta}$ sont tous des fonctions deux fois continuellement dérivables de ceux compris dans $\boldsymbol{\eta}$. Puisque les estimateurs des paramètres ne peuvent pas toujours être définis explicitement sous forme d'une fonction d'autres statistiques provenant d'échantillons, nous pouvons dériver les matrices $\mathbf{H}_A$ et $\mathbf{H}_B$ en linéarisant les équations de score (Binder 1983). L'hypothèse que $N_{ab}/N \to \kappa \in (0, 1)$ garantit que la linéarisation est bien définie.

Le théorème 1 montre que la linéarisation peut être utilisée pour estimer les variances des paramètres d'intérêt. Cependant, dans de nombreuses situations, les matrices $\mathbf{H}_A$ et $\mathbf{H}_B$ ont une haute dimensionnalité et les estimateurs de variance linéarisés ont une forme complexe. Un moyen pratique d'estimer les variances des estimateurs consiste à utiliser l'estimateur jackknife proposé par Lohr et Rao (2000). Sous les conditions de régularité exposées dans leur théorème 4, les estimateurs jackknife et par linéarisation de la variance sont asymptotiquement équivalents. La forme de l'estimateur de variance jackknife est $v_{JK}(\hat{\boldsymbol{\theta}}) = v_A(\hat{\boldsymbol{\theta}}) + v_B(\hat{\boldsymbol{\theta}})$, où v_A est un estimateur jackknife obtenu en supprimant une unité primaire d'échantillonnage à la fois de la base de sondage A tout en utilisant l'ensemble de données complet provenant de la base de sondage B, et v_B est un estimateur jackknife obtenu en supprimant une unité primaire d'échantillonnage à la fois de la base de sondage B tout en utilisant l'ensemble de données complet provenant de la base de sondage A.

4.2 Étude en simulations

Le théorème 1 montre que les estimateurs pour base de sondage double sont convergents pour les quantités de population correspondantes sous le mécanisme de génération de données manquantes modélisé. Nous avons exécuté une petite étude en simulations pour examiner les propriétés pour des tailles d'échantillon modérées avec bases de sondage chevauchantes. Nous avons généré les données comme dans l'étude en simulations de Skinner et Rao (1996), avec $\gamma_a = N_a/N$ et $\gamma_b = N_b/N$. Nous avons créé un échantillon en grappes tiré de la base de sondage A contenant $\tilde{n}_A$ UPE et m observations dans chaque UPE, et un échantillon aléatoire simple de n_B observations pour la base de sondage B. Nous avons généré les réponses binaires en grappes pour l'échantillon provenant de la base de sondage A en créant des vecteurs aléatoires normaux multivariés corrélés, puis en utilisant la fonction probit pour convertir les réponses continues en réponses binaires.

Après avoir créé l'échantillon, nous avons calculé les estimateurs des probabilités de l'union de la base de sondage A et de la base de sondage B, ainsi que les moyennes des valeurs absolues du biais et des erreurs quadratiques moyennes empiriques (EQME) sous diverses conditions. Nous calculons l'EQME d'un estimateur donné, $\hat{Y}$, en nous servant de la formule :

$$\text{EQME} = \frac{1}{R}\sum_{r=1}^{R}(\hat{Y}_r - Y)^2, \qquad (6)$$

où $\hat{Y}_r$ est la valeur de $\hat{Y}$ pour la r^{e} exécution de la simulation. Dans notre étude en simulations, nous avons utilisé $R = 100$.

Nous avons effectué l'étude en simulations en prenant les facteurs suivants : 1) γ_a : 0,2 ou 0,4, 2) γ_b : 0,2 ou 0,4, 3) paramètre de groupement ρ : 0,3, 4) mécanisme de génération des données manquantes : la probabilité qu'une personne soit un non-répondant durant un mois donné dépend de la période et de la classification de la personne durant la période observée, ou données manquant entièrement au hasard, 5) quantité de données manquantes : près de 10 % ou près de 20 %, 6) tailles de l'échantillon : $\tilde{n}_A$: 10, 100 ou 500 ; m : 5, n_B : 100, 1 000 ou 5 000. Dans toutes les exécutions, les paramètres de probabilité étaient $\mathbf{p}_a$: (0,3 ; 0,1 ; 0,2 ; 0,4), $\mathbf{p}_{ab}$: (0,3 ; 0,1 ; 0,1 ; 0,5) et $\mathbf{p}_b$: (0,4 ; 0,1 ; 0,1 ; 0,4). Le tableau 3 donne les résultats de l'étude en simulations pour des données manquantes générées sous le modèle 1, quand elles sont prédites au moyen du modèle 1 et au moyen du modèle utilisant les enregistrements complets uniquement.

Quand les données manquent au hasard, tous les modèles donnent des estimateurs des proportions de flux bruts p_{kl} approximativement sans biais, si bien que nous ne présentons pas les résultats ici. L'examen du tableau 3 montre que le modèle correct ainsi que l'analyse des enregistrements complets seulement produisent des estimateurs biaisés des p_{kl}. Cependant, quand les tailles d'échantillon sont plus grandes, le biais persiste dans l'analyse portant sur les enregistrements complets uniquement, tandis qu'il diminue quand le modèle 1 est ajusté. Dans l'exemple présenté ici, les probabilités d'avoir des données manquantes sont relativement faibles. Lorsque la quantité de données manquantes est plus importante, le contraste entre les estimateurs est plus prononcé.

Tableau 3
Résultats de l'étude en simulations pour les données manquantes générées sous le modèle (1). Le cas (1) correspond à l'ajustement du modèle correct : modèle (1) ; le cas (2) correspond à l'utilisation des enregistrements complets uniquement. Le biais est égal au biais absolu moyen pour les proportions de flux bruts dans la population p_{kl} ; l'EQME est égale à l'erreur quadratique moyenne empirique moyenne pour les p_{kl} ; les proportions utilisées pour générer les données manquantes sont $\lambda_{(t-1)0} = 0{,}141$, $\lambda_{(t-1)1} = 0{,}070$, $\lambda_{(t)0} = 0{,}137$ et $\lambda_{(t)1} = 0{,}068$. Ici, $\tilde{n}_A$ est le nombre d'UPE dans l'échantillon A dont la taille est égale à 5 et n_B est le nombre d'éléments dans l'échantillon B

$\tilde{n}_A$	n_B		p_{00}	p_{01}	p_{10}	p_{11}
10	100	Estimateur	0,311	0,120	0,149	0,420
Cas 1		Biais	0,040	0,029	0,029	0,040
		EQME	0,002	0,001	0,001	0,002
			$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{(t)}$
		Estimateur	0,159	0,095	0,146	0,094
		EQME	0,001	0,001	0,002	0,001
10	100	Estimateur	0,286	0,120	0,146	0,448
Cas 2		Biais	0,048	0,029	0,029	0,041
		EQME	0,004	0,001	0,001	0,002
100	1 000	Estimateur	0,321	0,092	0,138	0,449
Cas 1		Biais	0,015	0,011	0,009	0,015
		EQME	3,337e-04	1,798e-04	1,418e-04	3,256e-04
			$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{(t)}$
		Estimateur	0,145	0,074	0,123	0,068
		EQME	2,642e-04	9,389e-05	3,917e-04	8,206e-05
100	1 000	Estimateur	0,293	0,092	0,135	0,480
Cas 2		Biais	0,0280	0,011	0,010	0,040
		EQME	0,001	1,839e-04	1,711e-04	0,002
500	5 000	Estimateur	0,321	0,093	0,135	0,452
Cas 1		Biais	0,006	0,008	0,007	0,012
		EQME	4,960e-05	7,162e-05	6,381e-05	1,857e-04
			$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{(t)}$
		Estimateur	0,140	0,071	0,123	0,064
		EQME	4,466e-05	1,818e-05	2,288e-04	3,545e-05
500	5 000	Estimateur	0,292	0,092	0,132	0,483
Cas 2		Biais	0,028	0,008	0,008	0,043
		EQME	8,265e-04	7,642e-05	9,571e-05	1,906e-03

5. Application

À la présente section, nous appliquons nos résultats à des données provenant de la Survey of Income and Program Participation (SIPP) et de la Current Population Survey (CPS) pour l'Arizona. La CPS et la SIPP sont des enquêtes par panel longitudinales stratifiées à plusieurs degrés. Nous traitons la SIPP et la CPS comme une enquête à base de sondage double ayant la même population cible, à savoir la population de l'Arizona de 18 à 64 ans. En utilisant l'information provenant des deux enquêtes, nous voulons modéliser la variation des probabilités de transition entre les situations d'emploi de janvier 2001 à janvier 2002 chez les personnes de 18 à 64 ans. Notons que, strictement parlant, ces deux enquêtes ne sont pas conçues comme une enquête à base de sondage double. Les questions relatives aux variables de population active ne sont pas les mêmes. Bien que nous ayons recodé les variables conformément aux définitions de la population active appliquées dans la CPS, il se peut que les différences d'énoncé et d'ordre des questions produisent un biais lorsque l'information est combinée. Nous utilisons ces données comme exemple, parce que des données longitudinales issues d'une base de sondage double réelles ne sont pas disponibles. Néanmoins, l'exemple montre les gains d'efficacité qui peuvent être réalisés en combinant l'information provenant de deux enquêtes pour estimer les flux bruts.

Les deux enquêtes ont pour population cible la population civile à domicile des États-Unis. Nous considérons un sous-ensemble de la population, à savoir la population sur le marché du travail âgée de 18 à 64 ans. Donc, $N_A = N_B = N_{ab}$ et le problème d'estimation est un cas particulier de la théorie exposée à la section 3. Le fichier longitudinal pour la SIPP de 2001 et de 2002 (Westat 2001) s'appuie sur un seul panel. Nous avons fusionné la vague 1 (contenant les enregistrements de janvier 2001), la vague 4 (contenant les enregistrements de janvier 2002) et le fichier de poids longitudinaux, dans lequel les poids sont corrigés pour que leur somme concorde avec le chiffre de population. Puisque les poids du panel longitudinal ont été corrigés de la non réponse, nous considérons qu'il s'agit d'un cas sans données manquantes. Le tableau 4 donne la table de contingence résultante des flux bruts pondérés selon la SIPP.

Pour la CPS, le plan à panel rotatif introduit des données partiellement classifiées. Les mois de janvier 2001 et janvier 2002 ont en commun 50 % de l'échantillon. Nous utilisons ces 50 % de données, ainsi que les données partiellement classifiées pour exécuter l'analyse. La variable de pondération que nous utilisons est un poids transversal avec corrections transversales de la non-réponse et calage transversal (United States Census Bureau 2006). Pour les personnes ayant participé à l'enquête l'une des deux années seulement, nous utilisons le poids calculé pour l'année en question. Pour les personnes ayant participé en janvier 2001 ainsi qu'en janvier 2002, nous utilisons la moyenne des deux poids, afin de minimiser la variance de l'estimateur composite. Le groupe de population étudié est celui des 18 à 64 ans, et nous avons exclu les personnes qui n'appartenaient pas à cette catégorie les deux années. Le tableau 5 donne la table de contingence des flux bruts pondérés selon la CPS.

Tableau 4
Table des flux bruts pour la SIPP, en Arizona

		Janvier 2002		
		Occupé(e)	**En chômage**	
Janvier 2001	Occupé(e)	2 491 029	73 204	
	En chômage	30 698	30 160	
				2 625 091

Tableau 5
Table des flux bruts pour la CPS, en Arizona

		Janvier 2002			
		Occupé(e)	En chômage	Données manquantes	
Janvier 2001	Occupé(e)	1 129 656	38 848	689 497	
	En chômage	41 586	8 211	36 041	
	Données manquantes	606 549	57 549		
					2 607 937

Puisque nous considérons que la SIPP est un cas sans données manquantes, nous supposons que $\phi_{kl} = \psi_{kl} = 0$ et utilisons un modèle de type 1 dans l'analyse des données. Dans les données de la CPS, nous ajustons chaque poids en appliquant le facteur 2 625 091/2 607 937 pour atteindre un total de population unique pour les deux périodes et un total de population unique pour les deux enquêtes. Le nombre d'observations dans la SIPP (base de sondage A) après avoir combiné janvier 2001 et janvier 2002 est de 551, et l'effet de plan pour le chômage est de 1,76 environ, de sorte que $\overline{n}_A = 551/1{,}76 = 313$. L'effet de plan pour le chômage dans la CPS (base de sondage B) est de 1,229 environ, de sorte que $\overline{n}_B = 1\,020/1{,}229 = 830$. En raison des facteurs de vraisemblance, les paramètres estimés des probabilités produits par les cinq modèles données en (3) sont tous les mêmes. Le tableau 6 donne les probabilités estimées et les erreurs-types pour la SIPP, la CPS et les données résultant de la combinaison de ces deux enquêtes.

Tableau 6
Probabilités de transition estimées en utilisant la SIPP, la CPS et la méthode à base de sondage double avec la SIPP et la CPS. Les erreurs-types sont entre parenthèses

	p_{00}	p_{01}	p_{10}	p_{11}
SIPP	0,9489	0,0279	0,0117	0,0115
	(0,0124)	(0,0093)	(0,0061)	(0,0060)
CPS	0,9088	0,0454	0,0353	0,0106
	(0,0100)	(0,0072)	(0,0064)	(0,0035)
SIPP et CPS	0,9230	0,0381	0,0262	0,0127
	(0,0080)	(0,0058)	(0,0050)	(0,0030)

Pour des raisons de confidentialité, aucune information sur la mise en grappes n'est disponible dans les ensembles de données à grande diffusion de la CPS. Nous avons utilisé le produit de l'effet de plan publié et de la variance sous échantillonnage multinomial pour estimer les variances pour les données de la SIPP, ainsi que de la CPS. Nous avons appliqué le résultat du théorème 1 pour estimer les variances de $\hat{p}_{kl}$ pour $k, l = 0, 1$. Dans cette situation particulière, l'estimation de la variance résultant de la combinaison des deux ensembles de données se réduit à $(\overline{n}_A/(\overline{n}_A + \overline{n}_B))^2 V_A + (\overline{n}_B/(\overline{n}_A + \overline{n}_B))^2 V_B$, où V_A désigne l'estimation de la variance provenant des données de la SIPP et V_B, l'estimation de la variance provenant des données de la CPS. Le tableau 6 montre que les erreurs-types sont réduites si l'on utilise la méthode à base de sondage double.

Nous avons également effectué sur les cinq modèles donnés en (3) les tests d'adéquation élaborés dans Lu (2007). Les estimations des paramètres produites par les cinq modèles et les résultats des tests d'adéquation des modèles sont présentées au tableau 7. Les cinq modèles étant tous bien ajustés aux données, nous recommandons d'adopter le plus simple, c'est-à-dire le modèle 3, pour les données.

Tableau 7
Paramètres estimés et résultats des tests d'adéquation

	Paramètres estimés				ddl	G^2 corrigé	valeur p
Modèle 1	$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{t(1)}$	3	3,03	0,39
	0,246	0,395	0,277	0,302			
Modèle 2	λ_{t-1}	λ_t			5	8,58	0,12
	0,255	0,278					
Modèle 3	λ_0	λ_1			5	6,61	0,25
	0,262	0,353					
Modèle 4	$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	λ_t		4	4,10	0,39
	0,246	0,397	0,278				
Modèle 5	λ_{t-1}	$\lambda_{t(0)}$	$\lambda_{t(1)}$		4	6,74	0,15
	0,255	0,277	0,313				

Étant donné l'information limitée disponible dans les ensembles de données à grande diffusion, nous avons utilisé de simples corrections des poids pour faire concorder les chiffres estimés de population avec les totaux connus. Les poids inclus dans les ensembles de données de la SIPP et de la CPS ont déjà été calés et corrigés de la non-réponse, de sorte que les modèles utilisés pour les données manquantes reflètent principalement le plan à panel rotatif plutôt qu'une érosion due au déménagement ou à d'autres activités qui pourraient être reliées à la situation d'emploi.

D'autres travaux de recherche sur ces modèles pourraient inclure l'utilisation de diverses corrections de la pondération pour les enquêtes longitudinales. En outre, des paramètres différents pourraient être utilisés pour faire la distinction entre les observations dont la classification est partielle à cause du plan à panel rotatif et celles dont la classification est partielle à cause de la non-réponse. Pour cela, nous pourrions introduire un modèle de chaîne de Markov semblable à celui proposé par Stasny (1987). Dans le modèle avec données complètes, les individus sont répartis dans la table suivant une loi multinomiale unique. À la deuxième étape du processus, qui est également inobservée, chaque individu peut être choisi pour soit sortir de l'échantillon après l'interview du mois $t-1$, soit entrer dans l'échantillon avant l'interview du mois t, conformément au plan d'échantillonnage. Enfin, à la troisième étape du processus, chaque individu restant peut soit perdre sa classification de ligne, soit sa classification de colonne pour d'autres raisons.

En utilisant ce modèle, nous pouvons modéliser la non-réponse aux deux périodes (c'est-à-dire perdre à la fois les classifications de ligne et de colonne).

6. Conclusion

Dans le présent article, nous avons élaboré des méthodes statistiques pour estimer les flux bruts en nous appuyant sur des enquêtes à base de sondage double. Ces méthodes sont nécessaires pour estimer les changements de situation de pauvreté ou de situation d'emploi au cours du temps. Nous avons élaboré des estimateurs du pseudo-maximum de vraisemblance s'appuyant sur la structure à base de sondage double et les propriétés des deux plans de sondage. Nos modèles tiennent également compte des effets des données manquantes dues au fait qu'une personne cesse de participer à l'enquête ou qu'un plan à panel rotatif est utilisé, de sorte qu'ils permettent d'utiliser pleinement l'information partielle qui peut être fournie par certains ménages. Nous utilisons une méthode jackknife pour estimer la variance des estimateurs et examinons les propriétés de ces derniers. Nous avons appliqué les résultats à des ensembles de données réels.

Dans le présent article, les catégories des tables de contingence des flux bruts sont définies indépendamment des résultats de l'échantillon. Il est également possible de définir les catégories en se basant sur des valeurs qui dépendent de l'échantillon. Par exemple, dans les enquêtes sociales, le seuil de pauvreté pourrait être défini en utilisant un centile fondé sur l'échantillon et les catégories pourraient être définies comme étant « sous le seuil de pauvreté » et « au-dessus du seuil de pauvreté ». Les méthodes exposées dans le présent article peuvent être utilisées pour estimer les flux bruts si les définitions des catégories dépendent de l'échantillon, mais les estimateurs de variance doivent tenir compte de l'effet de l'estimation des bornes des catégories.

Bien que les résultats présentés ici aient trait à des enquêtes à base de sondage double, les méthodes sont générales et pourraient être étendues à plus de deux enquêtes en utilisant les estimateurs du pseudo-maximum de vraisemblance (EPMV) élaborés par Lohr et Rao (2006). Toutefois, la complexité des mécanismes éventuels de génération des données manquantes augmente parallèlement au nombre de bases de sondage. Les erreurs de classification pourraient également être plus fréquentes quand le nombre de bases de sondage est plus élevé.

Notre étude est effectuée dans le contexte des sondages, mais elle s'applique aussi à d'autres conditions dans lesquelles des données provenant de deux sources indépendantes pourraient être combinées. Comme il devient de plus en plus difficile de couvrir l'entièreté d'une population d'intérêt au moyen d'une seule enquête, nous pensons que ces méthodes d'estimation des flux bruts permettent d'obtenir une meilleure couverture de la population à un coût moindre. Elles permettent aussi de compléter une enquête auprès de la population générale par des enquêtes auprès de sous-populations particulières.

Remerciements

La présente étude a été financée en partie par la National Science Foundation aux termes des subventions SES-0604373 et DLS-0909630. Les auteurs remercient le rédacteur associé et les examinateurs de leurs commentaires avisés et constructifs.

Bibliographie

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

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Blair, E., et Blair, J. (2006). Dual frame web-telephone sampling for rare groups. *Journal of Official Statistics*, 22, 211-220.

Blumenthal, S. (1968). Multinomial sampling with partially categorized data. *Journal of the American Statistical Association*, 63, 542-551.

Catchpole, E.A., et Morgan, B.J.T. (1997). Detecting parameter redundancy. *Biometrika*, 84, 187-196.

Chambers, R.L., Woyzbun, L. et Pillig, R. (1988). Maximum likelihood estimation of gross flows. *Australian Journal of Statistics*, 30, 149-162.

Chen, T., et Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.

Fuller, W.A., et Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. Dans *Proceedings of the Social Statistics Section*, American Statistical Association, 245-249.

Hartley, H.O. (1962). Multiple frame surveys. Dans *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, Séries C, 36, 99-118.

Heeringa, S.G. (1995). Technical description of the assets and health dynamics (ahead) survey sample design. Technical Paper, Institute for Social Research, University of Michigan, hrsonline.isr.umich.edu/docs/userg/AHDSAMP.pdf.

Hocking, R.R., et Oxspring, H.H. (1971). Maximum likelihood estimation with incomplete multinomial data. *Journal of the American Statistical Association*, 66, 65-70.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under the regression super-population model. *Journal of the American Statistical Association*, 77, 89-96.

Kalton, G., et Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society*, Séries A, 149, 65-82.

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

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

Lohr, S.L., et Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lohr, S.L., et Rao, J.N.K. (2006). Estimation in Multiple-frame Surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

Lu, Y. (2007). Longitudinal estimation in dual frame surveys. Thèse de doctorat, *Arizona State University*.

Pfeffermann, D., Skinner, C. et Humphreys, K. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society*, Séries A, 161, 13-32.

Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.

Skinner, C.J., et Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Stasny, E.A. (1984). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, 25-40.

Stasny, E.A. (1987). Some Markov-Chain models for nonresponse in estimating gross labor force flows. *Journal of Official Statistics*, 4, 359-73.

Stasny, E.A., et Fienberg, S.E. (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.

United States Census Bureau (2006). Current Population Survey: Design and Methodology. Technical Paper 66, U.S. Census Bureau, Washington, DC.

Verma, V., Betti, G. et Ghellini, G. (2007). Cross-sectional and longitudinal weighting in a rotational household panel: application to EU-SILC. *Statistics in Transition*, 8, 5-50.

Westat (2001). Survey of Income and Program Participation Users' Guide (Supplement to the Technical Documentation). Rapport technique, Washington, DC.