Méthodes de sondage : applications aux enquêtes longitudinales, à la santé, aux enquêtes électorales et aux enquêtes dans les pays en développement. Paris : Dunod, 2008

LE TRAITEMENT COMBINE DES EFFETS DE NON-REPONSE NON-IGNORABLE ET DE SONDAGE INFORMATIF DANS L'ANALYSE DES DONNEES ISSUES DES ENQUETES LONGITUDINALES

Gad NATHAN[1] et Abdulhakeem EIDEH[2]

1. Introduction

Les données issues des enquêtes par sondage, et surtout par les enquêtes longitudinales, sont employées fréquemment pour inférer sur des modèles supposés. Souvent on ne tient pas compte des traits du plan du sondage (stratification, sondage par grappes ou à probabilités inégales) et les données venant de l'enquête par sondage sont analysées en employant des méthodes classiques, basées sur le plan de sondage aléatoire simple. Cette approche peut mener à des inférences erronées à cause du biais de sélection, impliqué par un plan de sondage informatif. Pour traiter les effets de tirage par probabilités inégales sur l'analyse de données issues des enquêtes longitudinales, Feder, Nathan et Pfeffermann (2000) ont appliqué des modèles hiérarchiques en combinaison avec des modèles de séries chronologiques. Pfeffermann, Krieger et Rinott (1998) ont proposé l'emploi de la distribution dans l'échantillon induite par un modèle supposé pour la population, sous un plan de sondage informatif, pour une enquête en temps unique, et ont développé des expressions pour son calcul. Une approche similaire est employée par Nathan et Eideh (2004) et par Eideh et Nathan (2006), en proposant des modèles de séries chronologiques pour l'analyse des données issues des enquêtes longitudinales sous un plan de sondage informatif général.

En plus de l'effet du plan de sondage complexe, un des problèmes principaux pour l'analyse des données issues des enquêtes longitudinales est celui des données manquantes. Pour l'analyse longitudinale on cherche à mesurer une série d'observations pour chaque unité dans l'échantillon. Des données manquantes peuvent apparaître quand des observations sont indisponibles pour un ou plusieurs des temps de la série, ou par intermittence, ou pour une période continue jusqu'à la fin de la série.

Dans le contexte d'enquêtes par sondage, le traitement des données manquantes dans les enquêtes longitudinales est considéré, sur la base du plan de sondage, par Kalton (1986) et

[1] Département de Statistique, Université Hébraïque de Jérusalem, gad@huji.ac.il
[2] Département de Mathématique, Université Alquds, Palestine, msabdul@ppu.edu

Lepkowski (1989). Pfeffermann et Nathan (2001) développent des méthodes de redressement des données manquantes dans les enquêtes longitudinales, par un modèle multiniveau intégré dans un modèle autorégressif. Skinner et Holmes (2003) proposent un modèle hiérarchique avec un effet aléatoire permanent au niveau de l'unité et des effets aléatoires temporaires, qui sont autocorrélés, pour les différentes périodes de l'enquête.

Dans cette communication nous étudions le traitement combiné de non-réponse non-ignorable et de sondage informatif pour l'analyse des données issues des enquêtes longitudinales, par la spécification de la distribution jointe des observations quand le plan de sondage est informatif. Cette distribution décrit simultanément l'effet du plan de sondage informatif et celui de la réponse informative.

2. La distribution dans la population

Soit y_{it} la valeur observée pour l'unité i (=1,…,N) en période t(=1,…,T). Avec chaque valeur, y_{it}, sont associées les valeurs (connues), x_{itk} (k=1,…,p), de p variables explicatives. On suppose que les valeurs y_{it} suivent le modèle de régression: $y_{it} = \beta_1 x_{it1} + ... + \beta_p x_{itp} + \varepsilon_{it}$, où les valeurs de ε_{it}, pour t=1,…, T, sont une série aléatoire de longueur T, associée à chacun des N unités. La structure longitudinale des données suggère que les valeurs de ε_{it} sont corrélées à l'intérieur des unités.

Soit $\mathbf{y}_i = (y_{i1},...,y_{iT})'$, $\mathbf{x}_{it} = (x_{it1},...,x_{itp})'$, et soit $\boldsymbol{\beta} = (\beta_1,...,\beta_p)'$ le vecteur des coefficients de régression inconnus. Le modèle linéaire général multivarié pour les données longitudinales considère les vecteurs aléatoires y_i, i=1,…,N, comme des variables normales multivariées, qui sont distribuées $\mathbf{y}_i \mid \mathbf{x}_i \sim MVN(\mathbf{x}_i\boldsymbol{\beta}, \mathbf{V})$, ou $\mathbf{x}_i$ est la matrice de taille T x p de variables explicatives pour l'unité i, et $\mathbf{V}$ a pour son élément (jk): $v_{jk} = \mathrm{cov}_p(y_{ij}, y_{ik}); \quad j,k = 1,...,T$, (Diggle, Liang et Zeger, 1994).

3. La distribution dans l'échantillon

Pour beaucoup d'exemples d'études longitudinales on emploie un sondage de panel, où les unités sélectionnées pour la première période restent dans l'échantillon jusqu'à la fin de l'étude – voir, par exemple, Nathan (1999). Nous supposons, donc, un plan de sondage informatif à un degré pour un échantillon de panel sélectionné à temps t=1 et que toutes les unités restent dans l'échantillon jusqu'au temps t=T. Il est raisonnable, alors, de supposer que les probabilités d'inclusion du premier ordre, π_i, dépendent des valeurs de la variable de réponse à la première occasion seulement, y_{i1}, et des valeurs des variables explicatives pour la première période, $\mathbf{x}_{i1} = (x_{i11},\ldots,x_{i1p})$. Si $\mathbf{y}_i \sim f_p(\mathbf{y}_i \mid \mathbf{x}_i, \boldsymbol{\theta})$ est la distribution conditionnelle dans la population, la distribution marginale dans l'échantillon de $\mathbf{y}_i$, étant donné $\mathbf{x}_i$, est donnée par:

$$f_s(\mathbf{y}_i \mid \mathbf{x}_i, \boldsymbol{\theta}) = \frac{E_p(\pi_i \mid y_{i1}, \mathbf{x}_{i1}, \boldsymbol{\theta})}{E_p(\pi_i \mid \mathbf{x}_{i1}, \boldsymbol{\theta})} f_p(y_{i1} \mid \mathbf{x}_{i1}, \boldsymbol{\theta}) f_p(y_{i2}, y_{i3}, \ldots, y_{iT} \mid y_{i1}, \mathbf{x}_i; \boldsymbol{\theta}) \qquad (3.1)$$

où $E_p(\pi_i \mid \mathbf{x}_{i1}, \boldsymbol{\theta}) = \int E_p(\pi_i \mid y_{i1}, \mathbf{x}_{i1}, \boldsymbol{\gamma}) f_p(y_{i1} \mid \mathbf{x}_{i1}, \boldsymbol{\theta}) dy_{i1}$. La démonstration de ce résultat est donnée par Eideh et Nathan (2006).

En supposant l'indépendance des observations dans la population, Pfeffermann, Krieger, et Rinott (1998) démontrent l'indépendance asymptotique des valeurs des unités sélectionnées sous la distribution dans l'échantillon, pour les plans de sondage avec des probabilités inégales, souvent employés. En conséquence, l'emploi de la distribution dans l'échantillon permet l'utilisation des procédures efficaces d'inférence standardisée, comme l'inférence basée sur le maximum de vraisemblance.

Notons qu'étant donnée la distribution dans la population, $f_p\left(y_{i1}\left|\mathbf{x}_{i1},\boldsymbol{\theta}\right.\right)$, la distribution dans l'échantillon, $f_s\left(y_{i1}\left|\mathbf{x}_{i1},\boldsymbol{\theta}\right.\right)$, est entièrement déterminée par les valeurs des espérances des probabilités d'inclusion, $E_p\left(\pi_i|y_{i1},\mathbf{x}_{i1}\right)$. Nous considérons les modèles approximatifs suivants pour ces espérances des probabilités d'inclusion, proposés par Pfeffermann, Krieger, et Rinott (1998) et par Skinner (1994):

(a) Modèle exponentiel:

$$E_p\left(\pi_i|y_{i1},\mathbf{x}_{i1}\right)=\exp\left(a_0^*+a_0y_{i1}+a_1x_{i11}+a_2x_{i12}+\ldots+a_px_{i1p}\right) \quad (3.2)$$

(b) Modèle linéaire:

$$E_p\left(\pi_i|y_{i1},\mathbf{x}_{i1}\right)=b_0^*+b_0y_{i1}+b_1x_{i11}+b_2x_{i12}+\ldots+b_px_{i1p} \quad (3.3)$$

Eideh et Nathan (2004) considèrent, en plus, les modèles logit et probit pour les espérances des probabilités d'inclusion

Dans le cas du modèle exponentiel pour les espérances des probabilités d'inclusion, équation (3.2), on peut démontrer que la distribution dans l'échantillon de $\boldsymbol{y}_i$, étant donnée la valeur de $\boldsymbol{x}_i$, est:

$$f_s\left(\mathbf{y}_i\left|\mathbf{x}_i,\boldsymbol{\theta}^*\right.\right)=\frac{\exp\left(a_0y_{i1}\right)f_p\left(y_{i1}\left|\mathbf{x}_{i1},\boldsymbol{\theta}\right.\right)}{M_p\left(a_0,\boldsymbol{\theta}\right)}f_p\left(y_{i2},y_{i3},\ldots,y_{iT}|y_{i1},\mathbf{x}_i;\boldsymbol{\theta}\right) \quad (3.4)$$

où $\boldsymbol{\theta}^*=\left(a_0,\boldsymbol{\theta}\right)$ sont des paramètres informatifs, que l'on doit estimer à partir de l'échantillon, et $M_p\left(a_0,\boldsymbol{\theta}\right)=E_p\left[\exp\left(a_0y_{i1}\right)\right]$ est la fonction génératrice des moments de la distribution dans la population de $\mathbf{y}_i$, étant donnée la valeur de $\mathbf{x}_i$.

Dans le cas du modèle linéaire pour les espérances des probabilités d'inclusion, équation (3.3), on peut démontrer que la distribution dans l'échantillon de $\mathbf{y}_i$, étant donnée la valeur de $\mathbf{x}_i$, est:

$$f_s\left(\mathbf{y}_i\left|\mathbf{x}_i,\boldsymbol{\theta}^*\right.\right)=\frac{\left(b_0^*+b_0y_{i1}+b_1x_{i11}+\cdots+b_px_{i1p}\right)f_p\left(y_{i1}|\mathbf{x}_{i1};\boldsymbol{\theta}\right)}{b_0^*+b_0E\left(y_{i1}\right)+b_1x_{i11}+\cdots+b_px_{i1p}}f_p\left(y_{i2},y_{i3},\ldots,y_{iT}|y_{i1},\mathbf{x}_i;\boldsymbol{\theta}\right) \quad (3.5)$$

où $\boldsymbol{\theta}^*=\left(b_0^*,b_{0,},b_1,\ldots,b_p,\boldsymbol{\theta}\right)$ sont des paramètres informatives, que l'on doit estimer à partir de l'échantillon.

Si on suppose que la distribution de $\mathbf{y}_i$, étant donnée la valeur de $\mathbf{x}_i$, dans la population est la distribution multivariée normale, on peut démontrer, en supposant le modèle exponentiel pour

les espérances des probabilités d'inclusion, que la distribution dans l'échantillon de $\mathbf{y}_i$, étant donnée la valeur de $\mathbf{x}_i$, est multivariée normale:

$$\mathbf{y}_i \mid \mathbf{x}_i \sim MVN\left(\boldsymbol{\mu}_i^*, \mathbf{V}^*\right) \tag{3.6}$$

où $\boldsymbol{\mu}_i^* = \left[\mathbf{x}_{i1}'\boldsymbol{\beta} + a_0 v_{11}, \mathbf{x}_{i2}'\boldsymbol{\beta}, ..., \mathbf{x}_{iT}'\boldsymbol{\beta}\right]'$ et la matrice des covariances, $\mathbf{V}^*$, est définie par:

$$v_{11}^* = v_{11}; \quad v_{1t}^* = v_{t1}^* = 0, \quad t > 1; \quad \text{et } v_{tt'}^* = v_{t,t'} - \frac{v_{t,1} v_{1,t'}}{v_{11}}; \quad t, t' = 2, ..., T. \tag{3.7}$$

Notons que si a_0=0, c'est-à-dire que le plan de sondage n'est pas informatif, la moyenne de la distribution dans l'échantillon, égale à celle de la distribution dans la population. Des résultats similaires sont obtenus pour les modèles linéaires, logit et probit.

4. Le traitement des effets de la non-réponse informative

Des observations manquent dans les enquêtes longitudinales de plusieurs manières et pour des raisons différentes, par exemple, des entreprises qui naissent et meurent, les individus qui cessent de répondre etc. Nous considérons ici seulement le cas des observations manquantes qui ne sont suivies que par des observations manquantes, c'est-à-dire que la non-réponse est finale. Si $y_{i1}, ..., y_{iT}$ est la série d'observations souhaitées pour l'unité i (=1,…,n), on suppose que si y_{it} manque, alors toutes les observations $y_{it'}$ pour $t'>t$, manquent aussi. Une question importante est si le manque d'observations est lié aux valeurs observées. En employant la terminologie introduite par Rubin (1976), nous traitons ici le cas du manque informatif (NMAR), c'est-à-dire que le processus de non-réponse dépend des valeurs observées et des valeurs non observées que l'on aurait obtenues, pour une unité qui n'a pas répondu, si elle avait répondu.

Supposons qu'une série complète des observations pourrait être obtenue pour l'unité i, qu'on l'indique par $\mathbf{y}_i^* = \left(y_{i1}^*, y_{i2}^*, ..., y_{iT}^*\right)'$, et on suppose que sa distribution est la distribution normale multivariée. Soit $\mathbf{y}_{i,d_i-1} = \left(y_{i1}, y_{i2}, ..., y_{i,d_i-1}\right)$ le vecteur des valeurs observées, où d_i indique la première période de non-réponse (avec d_i=T+1 si l'unité i répond pour toutes les périodes). On suppose que $y_{it} = y_{it}^*$ si t<d_i, et indiquons par D_i ($2 \le D_i \le T+1$) la variable aléatoire qui prend la valeur de la première période de non-réponse pour l'unité i (=1,…,n). Soit $H_{it} = \left\{y_{i1}, y_{i2}, ..., y_{i,t-1}, \mathbf{x}_i\right\}$ l'ensemble des observations obtenues pour l'unité i avant la période t (pour $t \le d_i$).

Le modèle général pour le processus de non-réponse informative suppose que la probabilité de non-réponse à la période t dépend de $\mathbf{y}_{i,d_i-1}$ et de $\mathbf{y}_i^*$. Diggle et Kenward (1994) proposent le modèle logistique pour le processus de la non-réponse de la forme:

$$\Pr\left(D_i = d_i \middle| H_{id_i}\right) = P_{d_i}\left(H_{id_i}, y_{id_i}^*; \boldsymbol{\varphi}\right) = \frac{\exp\left(\phi_1 y_{i1} + ... + \phi_{d_i-1} y_{i,d_i-1} + \phi_{d_i} y_{id_i}^*\right)}{1 + \exp\left(\phi_1 y_{i1} + ... + \phi_{d_i-1} y_{i,d_i-1} + \phi_{d_i} y_{id_i}^*\right)} \tag{4.1}$$

où $\boldsymbol{\varphi} = \left(\varphi_1, \ldots, \varphi_{d_i}\right)'$ est une vecteur de paramètres inconnus, ou un modèle logit semblable.

Une fois que le modèle du processus de non-réponse est déterminé, on peut obtenir la distribution conjointe du vecteur des observations $\mathbf{y}_{i,d_i-1}$, à partir de la distribution normale multivariée de $\mathbf{y}_i^*$ et de la série des distributions conditionnelles dans l'échantillon de $\mathbf{y}_{it}$, étant donné H_{it}. En employant les équations (3.1), on obtient pour la densité conjointe de la probabilité dans l'échantillon de la série incomplète de d_i-1 observations, $\mathbf{y}_{i,d_i-1}$, et de l'évènement que la non-réponse commence à temps d_i, l'expression suivante:

$$f_s\left(\mathbf{y}_i|\mathbf{x}_i\right)=f^*_{s,d_i-1}\left\{\mathbf{y}_{i,d_i-1}|\mathbf{x}_i\right\}\left\{\prod_{t=2}^{d_i-1}\left[1-P_t\left(H_{it},y_{it};\boldsymbol{\varphi}\right)\right]\right\}P\left(D_i=d_i|H_{id_i}\right) \qquad (4.2)$$

où $f^*_{s,d_i-1}\left\{\mathbf{y}_{i,d_i-1}|\mathbf{x}_i\right\}=f^*_s\left(y_{i1}|\mathbf{x}_{i1}\right)f^*_p\left(y_{i2},...,y_{i,d_i-1}|y_{i1},\mathbf{x}_i\right)$ est la densité de probabilité marginale dans l'échantillon de la série $\boldsymbol{y}_i$. et $\Pr\left(D_i=d_i\middle|H_{id_i}\right)=P_{d_i}\left(H_{id_i},y^*_{id_i};\boldsymbol{\varphi}\right)$ est défini par l'équation (4.1). Notons que ces résultats traitent en même temps les effets de la non-réponse informative et ceux du plan de sondage informatif

5. L'estimation des paramètres

Nous décrivons dans cette partie les méthodes d'estimation des paramètres des modèles, proposés pour l'analyse des données issues des enquêtes longitudinales, pour tenir compte les effets du plan de sondage informatif et de ceux de la non-réponse non-ignorable. On propose deux méthodes d'estimation, toutes les deux basées sur la distribution dans l'échantillon précédemment décrite, en supposant le modèle exponentiel de l'équation (3.2).

5.1 Estimation en deux étapes

Les deux ensembles de paramètres dans l'équation (4.2) qu'on doit estimer sont ceux de la distribution dans la population, $\boldsymbol{\theta}=\left(\boldsymbol{\beta},\mathbf{V}\right)$, et les paramètres de la distribution dans l'échantillon des valeurs observées: a_0, le paramètre du plan de sondage informatif, défini par l'équation (3.2); et $\boldsymbol{\varphi}$, le paramètre du modèle pour le processus de la non-réponse, selon l'équation (4.1). Indiquons par $\boldsymbol{\theta}^*=\left(\boldsymbol{\theta},a_0,\boldsymbol{\varphi}\right)$ l'ensemble des paramètres à estimer. D'abords les paramètres de la distribution dans la population sont estimés par une méthode en deux étapes sur la base de la distribution dans l'échantillon des valeurs observées. A la première étape les paramètres de l'espérance des probabilités d'inclusion sont estimés. A la deuxième étape les autres paramètres sont estimés par la minimisation de la log-vraisemblance, avec les estimateurs issus de la première étape remplaçants les vraies valeurs des paramètres. Selon Pfeffermann, Krieger et Rinott (1998), l'emploi de ce processus d'estimation en deux étapes est nécessaire quand l'espérance conditionnelle des probabilités d'inclusion est exponentielle, car en ce cas il y a un problème d'identifiabilité.

En réalité les espérances conditionnelles des probabilités d'inclusion, $E_p\left(\pi_i|y_{i1}\right)$, ne sont pas connues et les seules données disponibles pour l'analyste pour la première période sont les valeurs de $\left\{y_{i1},w_i;\ \ i\in s\right\}$, où w_i=1/π_i sont les poids de l'échantillon. L'estimation des valeurs de $E_p\left(\pi_i|y_{i1}\right)$, en employant seulement les données $\left\{y_{i1},w_i;\ \ i\in s\right\}$, peut être basée sur la relation suivante (Pfeffermann et Sverchkov, 1999):

$$E_s\left(w_i|y_{i1}\right)=1/E_p\left(\pi_i|y_{i1}\right) \tag{5.1}$$

Alors pour le modèle exponentiel des espérances des probabilité d'inclusion, équation (3.2), l'estimation en deux étapes se déroule comme suit :

- 1e étape - l'estimation de a_0: Selon l'équation (5.1) on peut estimer a_0 par la régression de $-\log(w_i)$ sur les valeurs de $y_{i1},x_{i11},\ldots,x_{i1p}$, pour toutes les unités dans l'échantillon.
- 2e étape - on substitue les estimateurs, $\tilde{a}_0$ de a_0, obtenus par la méthode des moindres carrés dans la distribution dans l'échantillon des valeurs observées. La contribution à la fonction de la log-vraisemblance pour les valeurs observées de l'unité *i* dans l'échantillon est obtenue par l'équation (4.2).

5.2 Estimation par la pseudo-vraisemblance

Cette approche est basée sur la solution des équations estimées du maximum de vraisemblance dans la population. Cet estimateur de maximum de la pseudo-vraisemblance des paramètres $\boldsymbol{\theta}=\left(\boldsymbol{\beta},\mathbf{V}\right)$ est défini comme la solution des équations du maximum de vraisemblance dans la population, qu'on peut exprimer, dans ce cas, comme:

$$U\left(\boldsymbol{\theta}\right)=\frac{\partial}{\partial\boldsymbol{\theta}}\log L_p\left(\boldsymbol{\theta}\right)=\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}}\log L_{ip}\left(\boldsymbol{\beta},\mathbf{V}\right)=0\,. \tag{5.2}$$

Alors l'estimateur de maximum de la pseudo-vraisemblance est défini comme la solution des équations $\hat{\mathrm{U}}\left(\boldsymbol{\theta}\right)=0$, où $\hat{\mathrm{U}}\left(\boldsymbol{\theta}\right)$ est un estimateur, basé sur l'échantillon, de $\mathrm{U}\left(\boldsymbol{\theta}\right)$, défini par l'équation (5.2) – voir Binder (1983).

6. Application empirique

Nous avons appliqué les résultats précédents aux données de l'enquête sur le travail Britannique (LFS), pour l'estimation des flux bruts entre les statuts d'activité pendant deux trimestres consécutifs, en supposant la non-réponse informative, comme étudié déjà par Chambers et Clarke (1998) et par Clarke et Tate (2002). On considère trois statuts de participation à la force du travail – employé (E), non employé (U), et n'appartient pas à la population active (N) – parmi lesquels on veut estimer les flux bruts. Pour l'application des méthodes proposées, on doit modifier les résultats obtenus plus haut pour traiter une variable catégorique. En effet, on se base sur la distribution multinomiale des fréquences pour les neuf catégories de flux entre les paires de statut. On suppose que la non-réponse concerne les ménages entiers, c'est-à-dire que tous les individus du ménage répondent si le ménage répond. Ceci correspond approximativement à la situation réelle pour la plupart des enquêtes sur le travail chez les ménages.

Les résultats pour les estimateurs des probabilités des flux bruts, basés sur les données du LFS pour le 3e trimestre de 1995, sont présentés dans le tableau 1, pour quatre méthodes d'estimation.

TABLEAU 1: ESTIMATEURS DES PROBABILITES DES FLUX BRUTS (%)

Flux	Simple	Pondérée	MVE - Exponentiel	MVE - Heckman
EE	70.62	69.32	69.78	68.23
EU	1.08	1.25	1.17	1.36
EN	1.53	1.71	1.61	1.74
UE	1.61	1.62	1.61	1.61
UU	3.78	4.41	4.16	5.06
UN	1.00	1.12	1.07	1.20
NE	1.40	1.29	1.35	1.28
NU	1.06	1.14	1.12	1.24
NN	17.92	18.12	18.13	18.28

Les méthodes d'estimation considérées sont les suivantes

- Simple (col. 1): L'estimation de probabilités est faite a la base des données non-pondérées, par le maximum de vraisemblance simple, sans tenir en compte ni des effets du plan de sondage et ni de ceux de la non-réponse.
- Ponderée (col. 2): Les probabilités sont estimées par la pondération des données avec les poids individuels calibrés utilisés généralement pour le LFS (Clarke et Tate, 2002), qui sont basés sur le plan de sondage et sur le redressement simple pour la non-réponse.
- Maximum de vraisemblance dans l'échantillon (MVE) – modèle exponentiel (col. 3): On maximise la vraisemblance dans l'échantillon sur la base du modèle exponentiel – équation (3.2), selon l'équation (3.6), c'est à dire en traitant les effets du plan de sondage, mais sans prendre en considération explicitement les effets de la non-réponse. L'estimation est faite par la méthode d'estimation en deux étapes, décrite dans la section 5.1, en employant les poids individuels calibrés qui sont une fonction de l'âge, du sexe, du niveau d'éducation, de la classe sociale et de la forme d'habitation.
- Maximum de vraisemblance dans l'échantillon (MVE) – modèle Heckman (col. 4) : L'analyse par la méthode des probits est appliquée au modèle de Heckman (1976, 1979), proposé pour traiter le biais de spécification, pour estimer les probabilités de réponse. Les variables explicatives employés sont le nombre d'enfants dans le ménage qui ont moins que deux ans et la moyenne des poids individuels calibrés du LFS.

Les résultats montrent des différences importantes entre les estimateurs simples, d'un côté, et les estimateurs pondérés et ceux qui sont basés sur la vraisemblance dans l'échantillon, de l'autre. Au contraire, les différences entre les différentes méthodes pour traiter les effets du plan du sondage complexe et ceux de la non-réponse sont assez petites. Les résultats sont aussi semblables à ceux obtenus par Clarke et Tate (2002). Une explication possible est le fait que toutes les méthodes appliquées emploient, de différentes façons, les mêmes poids individuels calibrés du LFS.

BIBLIOGRAPHIE

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, vol. 51, 279-292.

Chambers, R.L et Clarke, P.S (1998). 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**.** *Techniques d'Enquête,* vol. 24, 133-140.

Clarke, P.S et Tate, P.F. (2002). An application of non–ignorable non–response models for gross flows estimation in the British labour force survey. *Australian and New Zealand Journal of Statistics*, vol. 4, 413-425.

Diggle, P. et Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, vol. 43, 49-73.

Diggle, P. J., Liang, K. Y, et Zeger, S. L. (1994). *Analysis of longitudinal data*. Oxford: Science Publication.

Eideh, A.H. et Nathan, G. (2006). Fitting time series models for longitudinal survey data under informative sampling. *Journal of Statistical Planning and Inference*, vol. 136, 3052-3069.

Feder, M.F, Nathan, G., et Pfeffermann, D. (2000). Modélisation multiniveaux des données longitudinales d'enquêtes complexes à effets aléatoires variables en fonction du temps**.** *Techniques d'Enquête*, vol. 26, 53-65.

Heckman, J.J., (1976). The common structure of statistical models of truncation, sample selection and limited variables, and a simple estimation af such models, *Annals of Econometric and Social Measurement*, vol. 54, 475-492.

Heckman, J.J. (1979). Sample selection bias as a specification error, *Econometrica*, vol. 47, 153-161.

Kalton, G. (1986). Handling wave nonresponse in panel surveys, *Journal of Official Statistics*, vol. 2, 303-314.

Lepkowski, J.M. (1989). Treatment of wave nonresponse in panel surveys. Dans *Panel Surveys,* D . Kasprzyk et coll. (eds.), New York: Wiley, 348-374.

Nathan, G. (1999). *A review of sample attrition and representativeness in three longitudinal surveys*, GSS Methodology Series no. 13, London: Office of National Statistics.

Nathan, G. et Eideh, A.H. (2004). L'analyse des données issues des enquêtes longitudinales sous un plan de sondage informatif. Dans *Échantillonage et Méthodes d'Enquêtes,* P. Ardilly (ed.),. Paris: Dunod, 227-240.

Pfeffermann, D., Krieger, A. M, et Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, vol. 8, 1087-1114.

Pfeffermann, D. et Nathan, G. (2001). Imputation for wave nonresponse – existing methods and a time series approach. Dans *Survey Nonresponse,* R. Groves et coll. (eds.), Chap. 28., 417-429.

Pfeffermann, D. et Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya*, vol. 61, Ser. B, 66-186.

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

Skinner, C. J. (1994). Sample models and weights. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 133-142.

Skinner, C.J., et Holmes, D. (2003). Random effects models for longitudinal data. Dans *Analysis of Survey Data*, R. Chambers et C. Skinner (eds.), New York: Wiley, 175-195.