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

Article

Pondération dans les échantillons rotatifs : le cas de l'enquête SILC en France

par Pascal Ardilly et Pierre Lavallée

Décembre 2007

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2007
Vol. 33, n° 2, pp. 149-156
Statistique Canada, n° 12-001-X au catalogue

Pondération dans les échantillons rotatifs : le cas de l'enquête SILC en France

Pascal Ardilly et Pierre Lavallée [1]

Résumé

L'Enquête européenne sur le revenu et les conditions de vie (*Statistics on Income and Living Conditions*, SILC) a remplacé le Panel européen à partir de 2004. Elle permet de produire des statistiques annuelles sur la répartition des revenus, ainsi que sur la pauvreté et l'exclusion sociale. Cette enquête longitudinale, dont la collecte a eu lieu pour la première fois en France en mai 2004, touche tous les individus de plus de 15 ans occupant les 16 000 logements tirés dans l'échantillon-maître et la base de sondage des logements neufs. Tous ces individus sont suivis au cours du temps, même lorsqu'ils changent de logement. L'enquête doit aussi fournir des estimations transversales de qualité.

Afin de limiter la charge des enquêtés, le plan de sondage préconisé pour SILC par Eurostat est un schéma rotatif basé sur quatre panels d'une durée de quatre ans chacun avec remplacement d'un panel tous les ans. La France a néanmoins choisi de porter la durée de ses panels à neuf années. Le plan de sondage rotatif permet de répondre aux besoins longitudinaux et transversaux de l'enquête. Cependant, il pose des défis en matière de pondération.

Après un rappel du contexte de l'inférence lorsqu'on pratique une enquête longitudinale, l'article traite des pondérations longitudinales et transversales, qui sont conçues de manière à produire des estimateurs approximativement sans biais.

Mots clés : Enquête longitudinale; panel; méthode du partage des poids; pondération longitudinale; pondération transversale.

1. Introduction

L'enquête SILC (*Statistics on Income and Living Conditions*) est une enquête européenne portant sur la mesure du revenu et sur l'évaluation des conditions de vie des personnes vivant en ménage ordinaire (on exclut donc les personnes vivant en communauté). Elle a remplacé, à partir de l'année 2004, le panel communautaire. Bien que l'enquête soit européenne, et donc sous la tutelle d'Eurostat, elle est menée de façon indépendante à l'intérieur de chaque état membre de l'Union européenne. Les états membres - comme ici la France - peuvent ainsi adapter le plan de sondage proposé par Eurostat afin de répondre à leurs besoins nationaux. Le traitement des données est aussi effectué à l'intérieur de chaque état membre, comme c'est le cas habituellement pour les enquêtes d'Eurostat au sein de l'Union européenne. Le présent article se restreint au cas de l'enquête SILC en France, mais il pourrait aussi concerner les autres états membres de l'Union européenne.

L'enquête SILC est une *enquête longitudinale*, menée en mai de chaque année, qui s'intéresse aux individus physiques beaucoup plus qu'aux ménages, et qui s'effectue en face-à-face auprès de toutes les personnes résidant dans les logements échantillonnés. Cette enquête peut se voir comme la version européenne de l'Enquête sur la dynamique du travail et du revenu (EDTR) menée par Statistique Canada (voir Lavallée 1995, ainsi que Lévesque et Franklin 2000).

L'échantillon de l'enquête SILC est rotatif : chaque année, à partir de 2004 , il est constitué par la réunion de neuf sous-échantillons panels, tirés dans des conditions identiques en régime stationnaire, pour partie dans l'échantillon-maître, pour partie dans la base de sondage des logements neufs (BSLN). L'échantillon-maître et la BSLN sont deux bases de sondage de logements construites respectivement à partir du Recensement de la population (RP) de la France et des données du Système d'information et de traitement automatisé des données élémentaires sur les logements et les locaux (SITADEL) (voir Ardilly 2006).

Chaque panel entrant dans SILC réunit tous les individus résidant dans l'ensemble des logements tirés. L'interrogation de l'ensemble des individus des ménages appartenant aux logements sélectionnés permet de produire des estimations à la fois au niveau des individus et des ménages de la population. Elle permet, de plus, d'optimiser les coûts de collecte en maximisant le nombre d'individus atteints pour chaque contact. Certaines estimations concernent néanmoins un champ réduit, défini par les personnes ayant 16 ans ou plus au 31 décembre de l'année d'enquête.

Chaque année, un sous-échantillon sort et un sous-échantillon entre pour le remplacer. La première année où l'enquête a eu lieu, soit 2004, chaque sous-échantillon comprenait 1 780 logements (à quelques unités près, du fait des procédures d'arrondi). À partir de la seconde année, donc dès 2005, la taille du sous-échantillon entrant de l'année a été fixée à 3 000 logements. Notons que

1. Pascal Ardilly, Divison « Échantillonnage et traitement statistique des données », INSEE, Paris, France. Courriel : pascal.ardilly@atih.sante.fr; Pierre Lavallée, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario), K1A 0T6 Canada. Couriel : pierre.lavallee@statcan.ca.

l'initialisation, en 2004, a conduit à un échantillon de 16 000 logements, divisé en neuf parties égales. L'une de ces parties a été interrogée une seule fois (en 2004), une autre deux fois (2004 et 2005), une autre trois fois (2004, 2005 et 2006), *etc*. En régime de croisière, un panel donné sera interrogé neuf années de suite. Durant la phase d'initialisation, qui s'achèvera en réalité en 2012 avec la sortie du neuvième et dernier sous-échantillon issu du tirage de 2004, les sous-échantillons seront évidemment sollicités à moins de neuf reprises.

La procédure d'échantillonnage proprement dite est la procédure standard utilisée lorsqu'on tire dans l'échantillon-maître et dans la BSLN (voir Ardilly 2006). Dans le cas présent, il n'y a aucune surreprésentation de catégories d'individus. C'est une enquête à taux uniforme - aux arrondis près - à l'exception des logements vacants ruraux et des résidences secondaires au RP de 1999 qui sont devenus principaux à la date de l'enquête, qui sont comme de tradition sous-représentés.

Le protocole de collecte permet de considérer chaque sous-échantillon comme un véritable panel d'individus : en effet, on suit physiquement les personnes qui quittent leur logement, les différentes directions régionales de l'INSEE se transmettant les dossiers des individus du panel qui déménagent. Pour plus de détails sur le plan de sondage de SILC, on peut consulter le *Journal Officiel de l'Union Européenne* du 17 novembre 2003 et les documents internes de l'INSEE décrivant l'échantillonnage pratiqué en France.

Comme SILC est une enquête longitudinale où il y a chevauchement de panels dans le temps, la pondération de l'échantillon amène une problématique particulière. Cet article présente en détail les deux types de pondération utilisés pour SILC. On discutera en premier lieu de principes généraux reliés au plan de sondage de SILC. En deuxième lieu, on présentera la pondération longitudinale et finalement, on s'attardera sur la pondération transversale.

Il est à noter que l'on n'abordera pas les questions de correction de la non-réponse ni du redressement des estimations. Le traitement de ces questions renvoie à ce que l'on retrouve en général dans les enquêtes longitudinales comme, par exemple, l'EDTR (voir Lavallée 1995, ainsi que Lévesque et Franklin 2000).

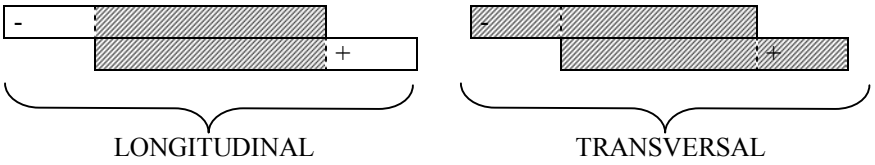
2. Principes généraux

2.1 Deux approches : la vision longitudinale et la vision transversale

Chaque année, on dispose d'un échantillon d'individus tous panélisés dont les huit neuvièmes ont déjà été interrogés au moins une fois les années passées (en l'absence de non-réponse).

On peut s'intéresser en particulier à deux types de paramètres : des totaux annuels Y_t (ou leurs satellites), ou des évolutions de totaux $\Delta_{t,t+1}$ entre deux années données, consécutives ou non. Pour simplifier, on considérera désormais qu'il s'agit de différences de totaux entre deux années consécutives. Quand on parle d'évolutions, il faut préciser les populations d'inférence qui entrent en jeu. Il y a alors deux façons de voir les choses : soit on raisonne sur des populations évolutives avec le temps et l'approche est dite transversale, soit on raisonne à population fixe et l'approche est dite longitudinale. Si on note Ω_t la population complète du champ de l'enquête l'année t, le total annuel à l'année t est donné par $Y_t = \sum_{i\in\Omega_t} Y_i^t$, où Y_i^t est une variable d'intérêt mesurée pour l'individu i. Lorsqu'on parle d'évolution, l'objectif peut être d'estimer la différence $\Delta^*_{t,t+1}$ entre le total Y_{t+1} à $t+1$ sur Ω_{t+1} et le total Y_t à t sur Ω_t, c'est-à-dire $\Delta^*_{t,t+1} = Y_{t+1} - Y_t$. Ceci correspond à une vision transversale. Sinon, l'objectif peut être d'estimer la différence $\Delta_{t,t+1}$ entre les totaux définis sur les unités communes aux populations Ω_{t+1} et Ω_t, les différences d'effectifs entre les deux populations s'expliquant par les unités entrant (naissances) et sortant (morts) de ces populations. Ceci correspond à une vision longitudinale. Soit $\Omega_{t,t+1} = \Omega_t \cap \Omega_{t+1}$, la population commune entre t et $t+1$. On définit alors $\Delta_{t,t+1}$ selon $\Delta_{t,t+1} = \sum_{i\in\Omega_{t,t+1}} (Y_i^{t+1} - Y_i^t)$.

Les schémas suivants synthétisent les deux approches. Le rectangle du haut symbolise la population complète à t et celui du bas la population complète à $t+1$. La partie « moins » représente les morts au sens large (décès, émigration, passage de l'individu en communauté,…) et la partie « plus » représente les naissances au sens large (nouveau-nés, immigration, entrée dans le champ par le franchissement d'un seuil d'âge,…). La partie grisée représente, à chaque date, la population d'inférence.



2.2 Enquêtes répétées dans le temps et stratégies envisageables

L'objectif est évidemment de pouvoir produire à la fois des estimations longitudinales et des estimations transversales. On peut envisager essentiellement trois stratégies :

1. Un échantillonnage « indépendant » chaque année. En fait, compte tenu de l'existence d'un échantillon-maître et d'une base de sondage de logements neufs (BSLN), les tirages s'effectuent tous les ans dans les mêmes communes, et par conséquent il n'y a pas de véritable indépendance entre les différents sous-échantillons. Cette solution est largement perfectible en terme de précision des évolutions.
2. Un échantillonnage intégralement panélisé, c'est-à-dire un tirage initial d'échantillon interrogé chaque année. Ce scénario pose en particulier un problème de charge, car l'opération SILC est engagée pour une durée indéterminée. De ce fait, il est irréaliste.
3. Un échantillon rotatif. C'est ce scénario qui a été choisi, compte tenu des avantages qu'il présente pour satisfaire les attentes en matière à la fois longitudinale et transversale.

Le tableau qui suit qualifie les trois plans de sondage envisageables en fonction des deux approches souhaitées.

TYPE d'échantillon	Approche TRANSVERSALE	Approche LONGITUDINALE
« Indépendant » chaque année	NATUREL	POSSIBLE mais moins efficace
Panel	IMPOSSIBLE sans tirage complémentaire	NATUREL
Rotatif	POSSIBLE	POSSIBLE

La stratégie rotative présente quatre grands atouts :

i. Elle réduit l'erreur d'échantillonnage associée à la mesure des évolutions (sur le principe, comme pour les panels, même si elle est moins efficace en théorie que le panel « pur »).

ii. Elle limite la charge des enquêtés par rapport au panel « pur ». En la circonstance, s'agissant pour la France d'un panel de neuf années, cet argument doit être utilisé avec modestie. Il a cependant plus de force dans le scénario préconisé par Eurostat, qui donne lieu à une enquête annuelle durant quatre années consécutives.

iii. Elle permet de prendre en compte d'une manière très « naturelle » l'évolution de la population avec le temps. Ce point sera plus compréhensible lorsqu'on abordera la question de la couverture des populations nouvelles.

iv. Elle permet de réduire les erreurs d'observation (comme les panels).

En revanche, on peut lui trouver au moins trois défauts :

i. Elle nécessite un suivi des individus dans le temps, ce qui occasionne des coûts de dépistage et des non-réponses du fait des déménagements.

ii. Par nature, la longueur des séries individuelles se limite à neuf années, ce qui est déjà fort appréciable, mais évidemment moins riche qu'un pur panel.

iii. La technique de pondération longitudinale/transversale n'est pas simple…

3. La pondération longitudinale

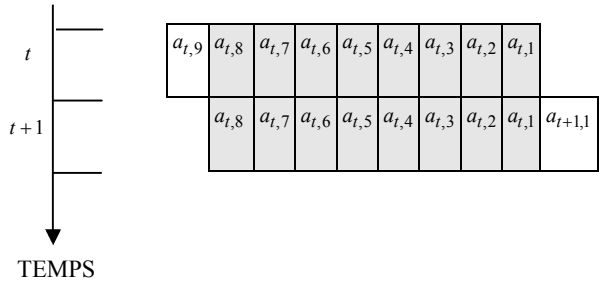
C'est une pondération a priori un peu plus simple à concevoir que la pondération transversale parce qu'il n'y a pas à tenir compte de l'évolution de la population dans le temps (en dehors des « morts », qui par convention disparaissent du champ avec le temps, mais ce point ne pose pas vraiment de problème technique). On rappelle que le principe de l'estimation longitudinale consiste à pratiquer une inférence sur la base d'une unique population considérée à une date initiale.

C'est clairement le caractère rotatif du plan de sondage qui complique la pondération, puisque entre deux années consécutives t et $t+1$, on va mobiliser huit panels distincts, tirés dans des populations différentes (il s'agit de populations d'individus physiques qui sont, bien entendu, différentes d'une année sur l'autre). Si on ne manipulait qu'un seul panel, il suffirait de s'en tenir à l'utilisation des poids de sondage associés aux individus du panel encore dans le champ à la date t, ni plus ni moins, puisque ces poids sont calculés une fois pour toutes au moment du tirage et permettent chaque année, sur toute la durée de vie du panel, une inférence sur la population initiale.

La difficulté essentielle consiste à représenter la population Ω_t à la date t à partir de huit sous-échantillons panels tirés à des dates différentes, donc dans des populations différentes. On peut comprendre intuitivement qu'un individu physique donné ait *in fine* une probabilité de sélection à la date t qui dépend du nombre de sous-échantillons panels dans lesquels il est susceptible d'être tiré. On suppose dans cette partie qu'il n'y a pas de non-réponse. La situation peut être formalisée de la manière suivante, en notant :

$a_{t,k}$ = sous-échantillon panel à enquêter l'année t en $k^{\text{ième}}$ interrogation, et $s_{t,t+1} = \bigcup_{k=1}^{8} a_{t,k}$.

On notera qu'on peut écrire $a_{t+1,k+1} = a_{t,k}$ $(\forall t, \forall k \neq 9)$ puisque par principe on reprend intégralement chaque sous-échantillon panel (non sortant) d'une année sur l'autre. Schématiquement, on a :



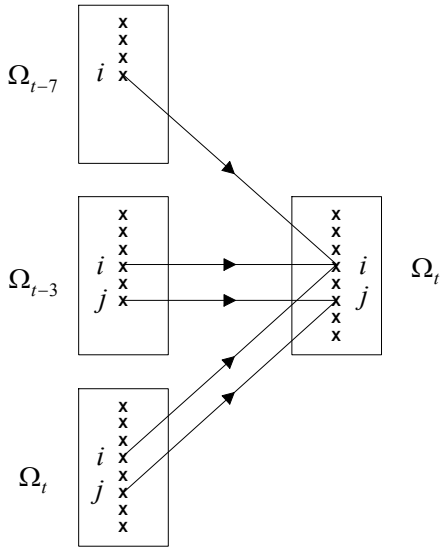
La partie grisée représente $s_{t,t+1}$ qui est l'échantillon exploité dans cette approche longitudinale. C'est en effet sur les individus de $s_{t,t+1}$ que l'on peut obtenir à la fois les informations Y_i^t et Y_i^{t+1} sur l'individu i définies respectivement aux dates t et $t+1$.

Soit un individu i quelconque de Ω_t, dans le champ de l'enquête à t. On note L_i le nombre d'années parmi $\{t-7, t-6, ..., t-1, t\}$ durant lesquelles l'individu i se trouvait dans le champ de l'enquête, donc était susceptible d'être tiré dans un panel « entrant ». Notons que l'on suppose ici que chaque année la base de sondage couvre exactement le champ de l'enquête. On a $L_i \in \{1, 2, 3, \ldots, 8\}$. Par ailleurs, on note K_i l'ensemble des indices k parmi 1, 2, 3, …, 8 pour lesquels on a $i \in a_{t,k}$. Il s'agit donc, à la date t, des numéros des panels dans lesquels on retrouve l'individu i. Pour tout i de $s_{t,t+1}$, K_i sera par construction un ensemble contenant au moins un élément. La plupart du temps, K_i ne comprendra en fait qu'un seul indice, mais parfois il pourra en comprendre deux, voire davantage : en effet, ce cas surviendra si i est tiré dans un panel, qu'il déménage et que son nouveau logement est échantillonné dans un autre panel, une année ultérieure. Il est à noter que notre contexte exclut qu'un logement donné soit tiré deux fois, parce qu'il y a un principe de non-réinterrogation des logements de l'échantillon-maître et de la BSLN. Mais ce n'est qu'une convention de nature pratique, la théorie s'accommodant fort bien d'un système dans lequel on pourrait retirer les logements.

Si $i \in a_{t,k}$, appelons $W_i(t,k)$ son poids de sondage « brut » : il s'agit du poids de sondage du logement dans lequel se trouve i à la date de son tirage en tant qu'individu-panel, donc lors du tirage annuel dans Ω_{t-k+1}. Ce système de poids permet une inférence directe du sous-échantillon $a_{t,k}$ vers la population complète Ω_{t-k+1}. En particulier, $\sum_{i \in a_{t,k}} W_i(t,k)$ estime sans biais le nombre total d'individus appartenant au champ de l'enquête et à la population Ω_{t-k+1}. Pour SILC en France, il s'agit d'un ordre de grandeur de 60 millions. Le poids longitudinal à affecter à tout individu i de $s_{t,t+1}$ sera *in fine* :

$$W_i^{t,t+1} = \frac{1}{L_i} \sum_{k \in K_i} W_i(t,k). \qquad (1)$$

Cette expression découle de l'application de la méthode de partage des poids (voir Lavallée 1995, ainsi que Lavallée 2002) où on définit la population initiale (celle des unités d'échantillonnage) comme réunion des populations $\Omega_{t-7}, ..., \Omega_{t-1}, \Omega_t$ et la population finale (celle des unités d'observation) comme Ω_t. Le schéma ci-dessous illustre le contexte, où, pour plus de clarté, nous n'avons pas reproduit les huit sous-populations initiales, mais seulement trois d'entres-elles. Le nombre de liens apparaît alors clairement égal à L_i (ici, par exemple, i a exactement huit liens, mais j en a strictement moins de huit parce qu'il n'apparaît pas dans les bases de sondage les plus anciennes). Pratiquement, il est réaliste de faire comme si on avait $\Omega_{t-7} \subset \Omega_{t-6} \subset \ldots \subset \Omega_{t-1} \subset \Omega_t$. On peut raisonner sur des populations emboîtées parce que, sauf exception, les individus qui sortent du champ au cours du temps avant t ne seront pas présents dans $s_{t,t+1}$.



La formule (1) fournit l'expression la plus générale possible du poids longitudinal « brut ». On peut ensuite la simplifier dans différents contextes. Si par exemple on néglige les cas où un individu-panel peut être tiré deux fois ou plus, on a

$$W_i^{t,t+1} = \frac{W_i}{L_i} \qquad (2)$$

où W_i est le poids de i relatif à l'unique sous-échantillon panel dans lequel il figure à la date t. Dans le cas de la France, compte tenu des tailles d'échantillon en jeu, adopter

in fine cette expression paraît tout à fait opportun. Si on se place dans un cadre idéal - qui paraît néanmoins trop simplifié dans notre contexte - où la population n'évolue pas dans le temps, on aura $L_i = 8$ pour tout *i*. Notons cependant que la population évolue beaucoup en neuf années, mais avec des durées de panélisation plus courtes, ce cas idéal peut être une approximation acceptable. Si, de plus, les panels sont tirés à probabilités égales, W_i sera égal à une constante *W* et alors

$$W_i^{t,t+1} = \frac{W}{8}. \qquad (3)$$

Ce cas de figure reste très peu probable dans le cas de la France. D'une part, jusqu'en 2012, il y a coexistence de sous-échantillons tirés avec des poids bruts nettement distincts (voir l'introduction). D'autre part, on aura tendance à concevoir l'échantillonnage en fixant le nombre total de logements à tirer (alors même que le nombre total de logements augmente) et non pas en raisonnant sur un objectif de taux de sondage constant.

Notons que la formule (3) est intuitive : finalement, tout se passe « comme si » n'importe quel individu de l'échantillon longitudinal $s_{t,t+1}$ avait une probabilité de sélection égale à huit fois celle qui caractérise chaque sous-échantillon panel composant $s_{t,t+1}$.

Ce qui précède s'applique au régime stationnaire et doit être légèrement adapté durant la phase d'initialisation du processus, c'est-à-dire jusqu'en 2012. La première opération de nature longitudinale porte sur les données conjointes 2004-2005, pour estimer des évolutions entre 2004 et 2005 avec la population de référence 2004 (privée des « morts » en 2005). Dans ce contexte, il suffit de diviser tous les poids W_i des huit sous-échantillons $a_{2004,1}$ à $a_{2004,8}$ par huit - autrement dit $L_i = 8$ pour tout *i*. En 2006, lorsqu'on s'intéressera aux évolutions 2005-2006, le dénominateur L_i pourra prendre deux valeurs seulement. Dans le premier scénario, l'individu-panel *i* était dans la base de sondage utilisée en 2004 (donc potentiellement tirable en 2004) et alors $L_i = 8$. Ceci vient du fait que tout se passe comme si, en 2004, on avait effectué les sept tirages des panels $a_{2005,2}$ à $a_{2005,8}$ exactement dans les mêmes conditions Dans le second scénario, l'individu *i* n'était pas dans la base de sondage de 2004 - mais alors il est dans la base 2005 et il se trouve nécessairement dans $a_{2005,1}$ - et $L_i = 1$. Pour mesurer les évolutions 2006-2007, L_i pourra être égal à 1, 2 ou 8, et ainsi de suite. Pour retrouver l'ensemble des valeurs possibles de L_i parmi $\{1, 2, 3, \ldots, 8\}$, il faudra attendre la mesure des évolutions 2011-2012.

Passée cette étape de pondération longitudinale, on obtient des poids longitudinaux $W_i^{t,t+1}$ et on forme l'estimateur de la différence $\Delta_{t,t+1}$ selon

$$\hat{\Delta}_{t,t+1} = \sum_{s_{t,t+1}} W_i^{t,t+1}.(Y_i^{t+1} - Y_i^t). \qquad (4)$$

A priori, les poids $W_i^{t,t+1}$ ne sont utilisés que dans le cadre d'une estimation d'évolution. Pour des estimations ponctuelles, ils apparaissent sans intérêt parce que la population d'inférence n'a pas grande signification à date donnée. Rappelons que jusqu'ici les $W_i^{t,t+1}$ n'ont fait l'objet d'aucune correction pour non-réponse, ni redressement. En pratique, l'estimateur (4) pour l'enquête SILC sera sujet à de tels ajustements.

L'estimation de la différence $\Delta^*_{t,t+1} = Y_{t+1} - Y_t$ correspond à une vision transversale : elle fait ainsi appel à la pondération qui est présentée dans la section suivante.

4. La pondération transversale

Il s'agit de pratiquer une inférence sur la population globale Ω_t du champ de l'enquête à la date courante *t*. La difficulté essentielle tient au fait qu'un sous-échantillon (panélisé) donné ne couvre correctement, en théorie, la population que l'année de son tirage. Passée cette année, le sous-échantillon panel ne représente plus la population nouvelle des « naissances », c'est-à-dire ceux qui entrent dans le champ de l'enquête. Cela concerne en particulier les nouveau-nés, les immigrants, les individus dont l'âge atteint certains seuils, les personnes anciennement sans domicile qui retrouvent un logement ordinaire, les retours de communautés, *etc*. Si en pratique on peut imaginer s'en satisfaire pendant quelque temps, ce défaut de couverture devient assez vite excessif (cela est vrai chaque année pour la plupart des sous-échantillons panels) et il faut d'une façon ou d'une autre obtenir un échantillon complémentaire au panel. Il est à noter que la problématique de l'évolution dans le temps de la population est fortement dissymétrique parce que la sous-population qui disparaît d'une année sur l'autre (les « morts ») ne pose pas de problème particulier en terme de pondération.

Dans l'enquête SILC, l'échantillon complémentaire est obtenu en appliquant la méthodologie suivante : on décide, pour chaque individu-panel enquêté lors du processus de suivi longitudinal, d'interroger l'ensemble des individus du ménage dans lequel se trouve l'individu-panel. Ainsi, tout ménage enquêté dans l'optique transversale est composé de deux types de personnes : des individus panel et des cohabitants (on nomme ainsi toute personne enquêtée qui n'est pas individu-panel). Cette méthodologie permet de couvrir une grande partie des « naissances » (au sens large) au sein de la population. Cependant, elle ne permet pas d'atteindre les ménages constitués seulement de « naissances » comme, par exemple, les ménages contenant seulement des immigrants. Précisons que la détermination

du statut de « naissance » se fait généralement en demandant, pour les nouveaux-nés, la date de naissance, et pour les immigrants, la date d'entrée au pays. On ajoute qu'en pratique, le défaut de couverture des naissances est en général considéré comme négligeable parce qu'il est en partie corrigé par l'utilisation du redressement.

La technique centrale utilisée pour produire les poids transversaux est la méthode de partage des poids (Lavallée 2002). Rappelons qu'à l'année *t*, on dispose de neuf sous-échantillons panels $a_{t,k} (1 \le k \le 9)$. On présente ici deux approches possibles pour l'application de la méthode du partage des poids. Notons que les informations à recueillir dans le questionnaire sont identiques pour mettre en oeuvre les deux méthodes.

4.1 Méthode 1

L'approche la plus rigoureuse consiste à relier l'ensemble des neuf sous-échantillons $a_{t,k}$ à l'échantillon transversal de l'année *t*, que nous noterons $\tilde{u}_t$ (Merkouris 2001). L'échantillon $\tilde{u}_t$ correspond donc à $s_{t,t} = \bigcup_{k=1}^{9} a_{t,k}$. Il faut tout d'abord commencer par définir les liens associés à ce schéma : lorsqu'un individu-panel quelconque de l'un des neuf sous-échantillons $a_{t,k}$ a été désigné par le sort, il pointe sur lui-même en tant qu'individu de l'échantillon transversal à *t* (schéma voisin de celui du 3.1). Dans ces conditions et en régime stationnaire, le poids transversal $W_i^{t(1)}$ d'un individu quelconque *i* de $\tilde{u}_t$ s'obtient de la façon qui suit. On note *m* le ménage auquel appartient *i*. On a

$$W_i^{t(1)} = \frac{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in a_{t,k}}} W_j(t,k)}{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in \Omega_{t-k+1}}} 1} \qquad (5)$$

où $W_j(t,k)$ est le poids de sondage qui découle de l'échantillonnage $a_{t,k}$.

Cette expression montre que tous les individus d'un même ménage ont à la fin le même poids. Au numérateur, on trouve la somme de tous les poids « bruts » (ceux qui reflètent l'échantillonnage) de tous les individus-panels du ménage, étant entendu qu'en général un individu-panel n'apparaît que dans un seul sous-échantillon mais qu'il peut y avoir des cas où un individu-panel a été tiré deux fois ou même davantage sur une période de neuf années consécutives (pour cause de déménagement, essentiellement). Il est à noter que l'échantillonnage de logements dans l'échantillon-maître et la BSLN s'appuie sur un principe de non-réinterrogation des logements déjà tirés et ainsi, dans le cas de SILC, la probabilité qu'un individu n'ayant pas changé de ménage apparaisse dans deux panels distincts est nulle.

Comme dans le cas longitudinal (voir 3.1), la pondération ne peut s'effectuer que si le système informatique de gestion des données est en mesure de rattacher chaque individu-panel de $\tilde{u}_t$ à l'ensemble des échantillons panels $a_{t,k}$ dans lesquels il se trouve. Au dénominateur, on dénombre pour chacune des neuf années $t-8$ à *t* considérées, les individus du ménage (qu'ils soient individus-panels ou cohabitants) qui se trouvent dans la base de sondage utilisée pour le tirage du sous-échantillon panel entrant l'année en question. Ce calcul nécessite évidemment la disponibilité de l'information via le questionnaire.

Cette approche a un double atout : d'une part elle est parfaitement générale, et d'autre part elle donne immédiatement lieu à des poids transversaux sans biais parce que tout ménage transversal est nécessairement relié à l'un quelconque des neuf sous-échantillons considérés. Le fait qu'il y ait chaque année un sous-échantillon entrant permet de représenter l'intégralité de la population transversale Ω_t, c'est-à-dire, dans un langage plus technique, assure l'existence d'au moins un lien pour chaque ménage considéré à *t*. C'est une propriété intéressante de l'échantillonnage rotatif que nous avons déjà mentionnée à la section 2.2. En contrepartie, la formule de pondération a un inconvénient qui est sa (relative) complexité, à la fois sur le plan théorique et lors de la phase de programmation informatique.

Dans la phase d'initialisation (donc jusqu'en 2011 compris), cette expression doit être adaptée : le numérateur ne change pas mais le dénominateur dénombre les individus échantillonnables à partir de 2004, première année de réalisation de l'enquête. En 2004, la pondération est évidente puisqu'il n'y a pas de partage des poids, mais en 2005 on prendra

$$W_i^{t(1)} = \frac{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in a_{t,k}}} W_j(t,k)}{\left(\sum_{\substack{j \in m \\ j \in \Omega_{2005}}} 1 \right) + 8 \cdot \left(\sum_{\substack{j \in m \\ j \in \Omega_{2004}}} 1 \right)}. \qquad (6)$$

En 2006, ce sera

$$W_i^{t(1)} = \frac{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in a_{t,k}}} W_j(t,k)}{\left(\sum_{\substack{j \in m \\ j \in \Omega_{2006}}} 1 \right) + \left(\sum_{\substack{j \in m \\ j \in \Omega_{2005}}} 1 \right) + 7 \cdot \left(\sum_{\substack{j \in m \\ j \in \Omega_{2004}}} 1 \right)}. \qquad (7)$$

4.2 Méthode 2

On peut avoir une vision alternative de la pondération transversale qui conduit à une expression de poids « un peu » plus simple et qui peut se programmer plus

facilement, mais qui se heurte à une difficulté qui n'apparaissait pas dans la méthode précédente et qui risque en pratique de rendre la pondération définitive un peu moins rigoureuse. L'idée est de raisonner non pas sur l'ensemble des sous-échantillons, mais sous-échantillon par sous-échantillon. On considère un quelconque des neuf sous-échantillons $a_{t,k}$ ainsi que l'échantillon de ménages auquel il mène. On applique alors le partage des poids, ce qui donne en régime stationnaire une pondération individuelle égale à

$$\tilde{W}_i(t,k) = \frac{\sum_{\substack{j\in m \\ j\in a_{t,k}}} W_j(t,k)}{\sum_{\substack{j\in m \\ j\in \Omega_{t-k+1}}} 1} \qquad (8)$$

pour tout individu *i* du ménage *m*. On vérifie très facilement que si $k=1$ (cas du sous-échantillon entrant), $\tilde{W}_i(t,1)$ est le poids de tirage du ménage *m*.

La difficulté associée à cette approche est liée à l'existence (a priori) à la date *t* d'individus non enquêtables parce qu'ils appartiennent à des ménages qui ne sont pas du tout « atteignables » au travers de l'échantillonnage $a_{t,k}$ (dès lors que $k \geq 2$), c'est-à-dire qui ont une probabilité nulle d'être enquêtés à *t*. Ce phénomène perturbateur n'existait pas dans la méthode précédente grâce à la prise en compte globale de l'ensemble des sous-échantillons puisqu'à la date *t* tout ménage a une probabilité strictement positive d'être sélectionné, au moins au travers de $a_{t,1}$. C'est une nouvelle occasion de souligner un des atouts essentiels de l'échantillonnage rotatif qui constitue une technique permettant chaque année de couvrir l'intégralité de la population. Dans notre approche, il est clair que si on considère $a_{t,k}$ pour $k \geq 2$, on ne couvre pas la population des ménages constitués exclusivement d'« immigrants » (au sens large) entre $t-k+1$ et *t*. Pour formaliser le contexte et aboutir au poids transversal final, on notera $\Omega_{\alpha,t}^{\text{immig}}$ la population d'« immigrants » (au sens large) présente à *t* dans des ménages ne comprenant que des immigrants échantillonnables après l'année α, avec $t-8 \leq \alpha \leq t-1$. Notons que, plus précisément, il faudrait dire « échantillonnables à partir d'une date strictement postérieure à la date de collecte de l'année α ».

A la date *t*, la population complète Ω_t est partitionnée en neuf composantes : les huit sous-populations $\Omega_{\alpha,t}^{\text{immig}}$, avec α variant de $t-8$ à $t-1$, et la sous-population constituée par les individus, soit qui étaient déjà enquêtables à $t-8$, soit qui sont devenus enquêtables à une date ultérieure à $t-8$ (donc des immigrants au-delà de $t-8$) mais qui sont intégrés à *t* dans un ménage comprenant au moins une personne enquêtable à $t-8$. Notons que l'on considère que si le ménage à *t* comprend au moins une personne échantillonnable à $t-8$, il en sera de même à toute date comprise entre $t-8$ et $t-1$. Cela revient à négliger les situations où un individu dans le champ à une date donnée en sort durant quelque temps (émigration, par exemple), puis y revient ensuite.

Par ailleurs, on note $\tilde{u}_{t,k}$ l'échantillon transversal à *t* issu du panel $a_{t,k}$, ce qui conduit à $\bigcup_{k=1}^{9} \tilde{u}_{t,k} = \tilde{u}_t$. Soit $Y_{\alpha,t}^{\text{immig}}$, le total des Y_i^t défini sur $\Omega_{\alpha,t}^{\text{immig}}$. On a alors, suite au partage des poids effectué pour tout $k = 2, ..., 9$:

$$E\left(\sum_{j\in\tilde{u}_{t,k}} \tilde{W}_j(t,k)\cdot Y_j^t\right) = \sum_{\Omega_t} Y_j^t - \sum_{\alpha=t-k+1}^{t-1} Y_{\alpha,t}^{\text{immig}}$$

$$= Y_t - \sum_{\alpha=t-k+1}^{t-1} Y_{\alpha,t}^{\text{immig}} \qquad (9)$$

et

$$E\left(\sum_{j\in\tilde{u}_{t,1}} \tilde{W}_j(t,1)\cdot Y_j^t\right) = \sum_{\Omega_t} Y_j^t = Y_t \qquad (10)$$

puisque $\tilde{u}_{t,1} = a_{t,1}$.

Avec un système de panels à courte durée, on pourrait peut-être négliger les $Y_{\alpha,t}^{\text{immig}}$ devant le vrai total sur Ω_t et alors le poids transversal final « brut » de tout individu *i* serait $\tilde{W}_i(t,k)/9$ si *i* est issu de $a_{t,k}$, ce qui conduirait à l'estimateur final

$$\hat{Y}_t = \frac{1}{9}\sum_{k=1}^{9}\sum_{i\in\tilde{u}_{t,k}} \tilde{W}_i(t,k)\cdot Y_i^t$$

$$= \frac{1}{9}\sum_{i\in\tilde{u}_t} \tilde{W}_i(t,k)\cdot Y_i^t. \qquad (11)$$

Les panels utilisés en France ont cependant une durée de vie longue, aussi il est fort possible que l'on ne puisse pas raisonner ainsi (l'examen des fichiers de collecte permettra d'en juger) et qu'il soit nécessaire de pondérer spécifiquement les individus des $\Omega_{\alpha,t}^{\text{immig}}$. Dans ces conditions, on vérifie que tout individu *i* de $\Omega_{\alpha,t}^{\text{immig}}$ qui se trouve finalement dans l'échantillon transversal $\tilde{u}_t$ aura un poids transversal brut $W_i^{t(2)}$ égal à la valeur $\tilde{W}_i(t,k)$ issue du partage des poids, divisée par $t-\alpha$ (et donc $1 \leq t-\alpha \leq 8$). Pour sa part, tout individu de Ω_t qui n'appartient à aucun des $\Omega_{\alpha,t}^{\text{immig}}$ (donc la grande majorité des cas) aura un poids final égal à $\tilde{W}_i(t,k)/9$. On remarquera par ailleurs que si *i* se trouve dans $\Omega_{\alpha,t}^{\text{immig}}$, il ne peut être enquêté qu'au travers de $a_{t,1}, a_{t,2}, \ldots, a_{t,t-\alpha}$. On obtient ainsi

$$W_i^{t(2)} = \begin{cases} \tilde{W}_i(t,k)/(t-\alpha) & \text{si } i \in \Omega_{\alpha,t}^{\text{immig}} \\ \tilde{W}_i(t,k)/9 & \text{sinon} \end{cases}. \qquad (12)$$

Durant la phase d'initialisation, il faut adapter les pondérations. En 2005, les individus de $\Omega^{\text{immig}}_{2004,2005}$ auront un poids final transversal directement issu du tirage du logement dans $a_{2005,1}$ (ils ne peuvent être atteints qu'au travers de ce panel entrant). En revanche, tous les autres individus sont « normalement » enquêtables à partir des neufs panels $a_{2005,k}$ $(1 \le k \le 9)$, si bien que leurs poids issus du partage des poids seront tous systématiquement divisés par 9. En 2006, les individus de $\Omega^{\text{immig}}_{2005,2006}$ auront un poids égal à celui du logement dans lequel ils résident et qui reflète directement le tirage de $a_{2006,1}$, ceux de $\Omega^{\text{immig}}_{2004,2006}$ auront leurs poids issus du partage des poids divisés par 2, et tous les autres individus auront leurs poids issus du partage des poids divisés par 9.

Ce traitement s'effectue bien sous-échantillon par sous-échantillon et ne doit pas tenir compte de ce qui survient dans les autres sous-échantillons. Si un individu est enquêté à t par l'intermédiaire de deux (ou plus) sous-échantillons $a_{t,k}$ distincts, on déroule le traitement complet associé à chacun des deux (ou plus) sous-échantillons. Ce peut être le cas, par exemple, d'un ménage composé de deux individus-panels provenant de deux sous-échantillons $a_{t,k}$ différents parce que ces individus se sont mariés et qu'avant leur mariage ils étaient suivis chacun séparément en formant un ménage de taille un. Dans cette configuration, chacun des deux individus est « formellement » enquêté deux fois, une fois en tant qu'individu-panel, une fois en tant que cohabitant.

Finalement, pour l'estimation de la différence $\Delta^*_{t,t+1} = Y_{t+1} - Y_t$, on pourra utiliser les poids $W_i^{t(1)}$ issus de la méthode 1, et ainsi calculer

$$\hat{\Delta}^*_{t,t+1} = \sum_{i \in \tilde{u}_{t+1}} W_i^{t+1(1)} Y_i^{t+1} - \sum_{i \in \tilde{u}_t} W_i^{t(1)} Y_i^t . \qquad (13)$$

Sinon, on pourra utiliser les poids $W_i^{t(2)}$ issus de la méthode 2. L'estimateur de la différence $\Delta^*_{t,t+1}$ sera alors donné par

$$\hat{\Delta}^*_{t,t+1} = \sum_{i \in \tilde{u}_{t+1}} W_i^{t+1(2)} Y_i^{t+1} - \sum_{i \in \tilde{u}_t} W_i^{t(2)} Y_i^t . \qquad (14)$$

Bibliographie

Ardilly, P. (2006). *Les techniques de sondage*, 2ième édition. Éditions Technip, Paris.

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

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

Lévesque, I., et Franklin, S. (2000). Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics, 1997 Reference year. Document de recherche sur le revenu, Statistique Canada, Catalogue No. 75F0002MIE-00004, juin 2000.

Merkouris, T. (2001). Estimation transversale dans le cas des enquêtes auprès des ménages à panels multiples. *Techniques d'enquêtes*, 27, 189-200.

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

Singh, M.P., Drew, J.D., Gambino, J.G. et Mayda, F. (1990). *Méthodologie de l'Enquête sur la population active*. Statistique Canada, Catalogue 71-526.