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

Pascal ARDILLY(*), Pierre LAVALLÉE(**)

(*) Insee, (**) Statistique Canada

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. C'est une enquête régulière, 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.

1. Principes généraux

1.1. Le plan de sondage

L'échantillon 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). Chaque panel entrant réunit l'ensemble des individus résidant dans l'ensemble des logements tirés (néanmoins, certaines estimations concernent 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 va sortir et un sous-échantillon va entrer pour le remplacer. La première année où l'enquête a eu lieu, soit en 2004, chaque sous-échantillon comprenait 1780 logements (à quelques unités près, du fait des procédures d'arrondi). A partir de la seconde année, donc dès 2005, la taille du sous-échantillon entrant de l'année a été fixée à 3000 logements[1].

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[2]. 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.
La procédure d'échantillonnage proprement dite est la procédure standard utilisée lorsqu'on tire dans l'échantillon-maître et dans la BSLN. Dans le cas présent, il n'y a aucune sur-représentation

[1] L'initialisation, en 2004, a conduit à un échantillon de 16 000 logements, tronçonné en neuf parties égales. Le régime stationnaire débute en 2005.
[2] Concrètement, en 2004, on a fractionné en neuf parties égales un gros échantillon de 16 000 logements. 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.

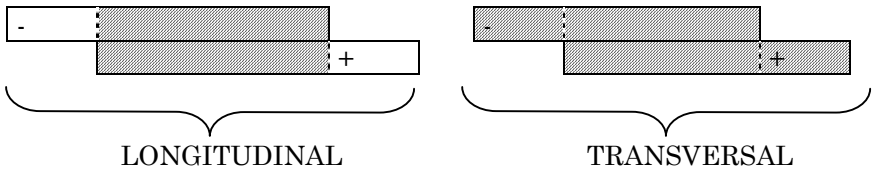
de catégories d'individus (enquête à taux uniforme - aux arrondis près - à l'exception des logements vacants ruraux et des résidences secondaires au Recensement de la population (RP) de 1999 qui sont devenus principales à la date de l'enquête, qui sont comme de tradition sous-représentés).

1.2. Deux approches : la vision longitudinale et la vision transversale

Chaque année, on dispose donc 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 (ou leurs satellites), ou des évolutions de totaux 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, l'objectif peut être d'estimer la différence entre le total à $t+1$ sur Ω_{t+1} et le total à t sur Ω_t (vision transversale), ou ce peut être d'estimer la différence 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.(vision longitudinale).

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.



1.3. 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 :

- Un échantillonnage « indépendant » chaque année. En fait, compte tenu de l'existence d'un échantillon-maître, 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.
- 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.
- 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 :

- 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 »).
- 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 (mais il a plus de force dans le scénario préconisé par Eurostat, qui donne lieu à une enquête annuelle durant quatre années consécutives).
- 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).
- Elle permet de réduire les erreurs d'observation (comme les panels).

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

- Elle nécessite un suivi des individus dans le temps, ce qui occasionne des coûts de dépistage et des non-réponses du fait des déménagements.
- 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.
- La technique de pondération longitudinale / transversale n'est pas simple…

2. 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.

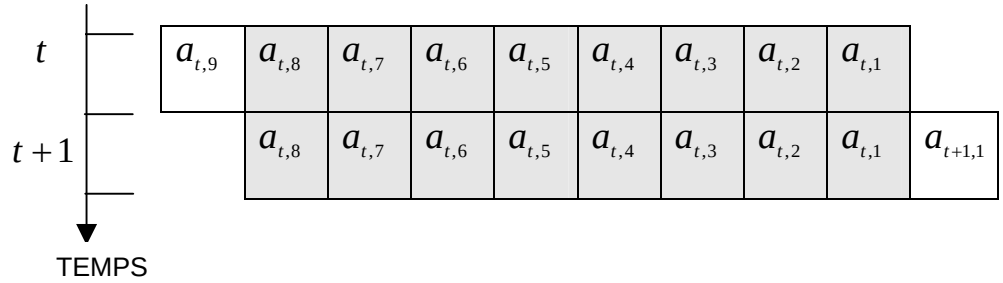
2.1. Le principe de pondération initiale, en l'absence de non-réponse totale

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 $t+1$.

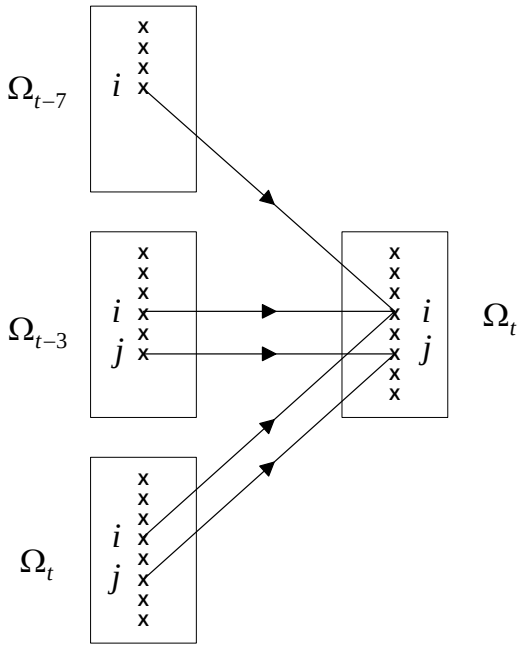
Soit un individu i quelconque de Ω_t, dans le champ de l'enquête à t. On note L_i le nombre d'années parmi $\{t-7, t-6, ..., t-1, t\}$ durant lesquelles l'individu i se trouvait dans le champ de l'enquête, donc était susceptible d'être tiré dans un panel « entrant »[3]. On a $L_i \in \{1,2,3,\ldots,8\}$. 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[4]. 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} (soit un ordre de grandeur de 60 millions pour SILC). Le poids longitudinal à affecter à tout individu i de $s_{t,t+1}$ sera *in fine* :

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

[3] On suppose que chaque année la base de sondage couvre exactement le champ de l'enquête.

[4] Notre contexte exclut qu'un logement donné soit tiré deux fois, parce qu'il y a un principe de non ré-interrogation 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.

Cette expression découle de l'application de la méthode de partage des poids, en définissant 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[5]. 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 précédente 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

$$W_i^{t,t+1} = \frac{W_i}{L_i}$$

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 trop simplifié dans notre contexte[6] - où la population n'évolue pas dans le temps, on aura $L_i = 8$ pour tout i. Si, de plus, les panels sont tirés à probabilités égales[7], W_i sera égal à une constante W et alors

$$W_i^{t,t+1} = \frac{W}{8}$$

[5] Pour plus de clarté, on ne reproduit pas les huit sous-populations initiales sur ce schéma, mais seulement trois d'entres-elles.

[6] La population évolue beaucoup en neuf années, mais avec des durées de panélisation plus courtes, ce cas idéal peut être une approximation acceptable.

[7] 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 1.1). D'autre part, on aura toujours tendance à concevoir l'échantillonnage en se 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.

Ce résultat simple est intuitif : 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 portera 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 suffira 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 : soit l'individu-panel i était dans la base de sondage utilisée en 2004 (donc potentiellement tirable en 2004) et alors $L_i = 8$ (cela 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), soit il n'y était pas (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.

2.2. Le traitement de la non-réponse partielle

Il s'effectue plutôt par imputation, parce qu'on peut ainsi profiter d'informations auxiliaires pour améliorer la qualité de la prédiction des Y_i manquants. On peut utiliser la batterie traditionnelle des méthodes d'imputation, en l'occurrence plutôt des méthodes axées sur l'utilisation de variables auxiliaires. Ce qui est particulièrement intéressant et spécifique aux panels, c'est la possibilité d'utiliser comme variables auxiliaires des séries de données individuelles : si Y_i^t est manquant, on peut utiliser la série $Y_i^{t-1}, Y_i^{t-2}, Y_i^{t-3}$, ..., dont le pouvoir explicatif est généralement fort, pour mieux prédire Y_i^t. Dans ces conditions, on optera pour une imputation de type aléatoire (hot deck, résidus simulés,...), certes au prix d'une augmentation de variance, mais cela permettra de préserver les distributions, ce qui est un objectif essentiel dans une enquête centrée sur la mesure des inégalités de ressources. Cela étant, il y a de nombreux cas de figure à envisager parce que les séries peuvent comprendre plusieurs valeurs manquantes relatives au passé, et cela selon tous les scénarios imaginables[8]. La valeur prédite $\tilde{Y}_i^t$ est alors une fonction, *primo* des valeurs passées des Y, *secondo* de valeurs passées ou présentes de variables auxiliaires X issues du questionnaire ou non, et *tertio* d'une composante aléatoire.

2.3. Le traitement de la non-réponse totale

Il s'effectue généralement par repondération, et nous présentons ici une approche traditionnelle basée sur un modèle explicite d'estimation des probabilités de réponse (Särndal, Swensson et Wretman, 1992, ainsi que Lock Ho et Scheuren, 1983). Il y a, par ailleurs, une autre option qui consiste à pratiquer simultanément la correction de non-réponse et le redressement (ce qu'on appelle le « calage en une étape » (Deville, 1998)), très peu différente de la première dans les enquêtes ponctuelles dès lors qu'on travaille sur des bases de sondage actualisées (comme ce sera

[8] Dans une telle configuration, l'introduction d'hypothèses sur les lois des Y_i^t permettrait d'utiliser des algorithmes sophistiqués, comme par exemple l'algorithme EM (Schafer, 1997).

le cas en France à partir de 2008, grâce aux enquêtes annuelles de recensement). En la circonstance, il ne s'agit pas d'une enquête ponctuelle et l'estimation explicite de probabilités de réponse peut présenter un intérêt spécifique.

Si l'individu répond plusieurs années de suite et que brusquement il devient non-répondant total, il est vrai qu'on peut hésiter entre, d'une part une repondération (comme suggéré ici) et d'autre part une imputation de l'ensemble des valeurs du questionnaire (méthode abordée au 2.2.). En effet, la richesse des séries des Y_i^t en terme d'information peut rendre l'imputation très attractive. C'est en particulier le scénario de la non-réponse dite « de vague », c'est-à-dire de la non-réponse « accidentelle » (individu absent durant le mois de collecte d'une année donnée, par exemple). Si le phénomène de non-réponse persiste dans le temps (en fait, dès la seconde année de non-réponse consécutive), on a à faire à de l'érosion (on dit aussi attrition) et le traitement s'effectuera alors plus naturellement par repondération.

Comme dans toutes les enquêtes, une des principales difficultés consiste à bien faire la distinction entre hors-champ et non-réponse. Un ménage qui ne répond pas mais qui est hors-champ ne doit pas participer à l'estimation des probabilités de réponse. Obtenir cette information est toujours difficile en pratique, parfois même impossible, mais en tout état de cause, il faut faire un effort maximum pour distinguer les deux cas de figure, c'est-à-dire essentiellement pour savoir si la résidence tirée est ou non une résidence principale au moment de l'enquête[9].

Le contexte longitudinal a la particularité d'offrir deux méthodes (au moins) d'estimation des probabilités de réponse.

Méthode 1 :

C'est l'approche traditionnelle. Elle ne tire pas avantage du caractère longitudinal mais elle a le gros atout d'être sensiblement plus simple à mettre en œuvre que la méthode alternative (méthode 2) et est applicable en toutes circonstances. Il s'agit d'utiliser un modèle logistique (par exemple) pour estimer directement, pour tout t et pour tout k de 1 à 8 :

$$P(i \text{ répond à } t+1 \mid i \in a_{t,k})$$

C'est, en effet, la réponse à $t+1$ qui apparaît comme la caractéristique la plus naturelle à modéliser (par opposition à la réponse à t), parce que l'échantillon répondant à $t+1$ est chronologiquement le dernier dont on dispose et parce qu'on s'attend à ce qu'il constitue « presque » un sous-ensemble de l'échantillon répondant à t. Le point 2.5 évoquera la question connexe de la connaissance ou non des valeurs Y_i^t, ces valeurs individuelles entrant dans l'expression de l'estimateur d'évolution basé sur les répondants à $t+1$.

Le conditionnement utilisé vient du fait que le processus probabiliste qui modélise la non-réponse est déterminé par l'ancienneté de i dans l'échantillon interrogé une année donnée[10]. Cela traduit le phénomène d'érosion (la probabilité de réponse décroît avec l'ancienneté). Cette probabilité est aussi

$$P(i \text{ répond à } t+1 \mid i \text{ interrogé pour la première fois à } t-k+1)$$

ou encore $P(i \text{ répond au } k+1^{ème} \text{ interview})$ si on considère que le processus n'est pas sensible à t (hypothèse de type « processus stationnaire »).

[9] Un redressement final sur le nombre total de résidences principales estimé à la date de l'enquête permettra néanmoins d'atténuer les conséquences d'éventuelles erreurs sur la catégorie du logement.
[10] Si par hasard l'individu est tiré dans deux panels, on retiendra seulement le plus ancien.

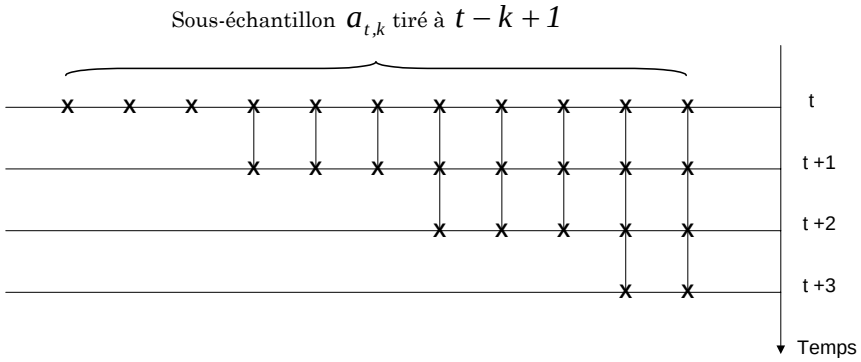
Cette approche ne se conçoit donc concrètement que sous-échantillon par sous-échantillon, afin de pouvoir estimer les probabilités de réponse en tenant compte de l'ancienneté de présence de l'individu dans le panel. On considère donc l'ensemble des individus constituant la panel $a_{t,k}$ et on modélise la probabilité (conditionnelle) précédente, *a priori* en utilisant les informations auxiliaires de la base de sondage parce qu'il faut mobiliser des variables connues sur les non-répondants. Il y a donc une dégradation de la qualité des estimations des probabilités de réponse au fur et à mesure que l'ancienneté k augmente. La suite relève des techniques classiques, étant entendu qu'on utilise souvent une modélisation logistique. Clairement, cette approche ne prend pas en compte « l'histoire » de l'individu sur la période $[t-k+1,t]$.

Méthode 2 (Naud, 2004)

Dans certaines conditions, on peut tirer parti du caractère longitudinal de l'enquête en cherchant à utiliser les séries de données individuelles collectées durant les années antérieures à $t+1$ pour modéliser plus efficacement les probabilités de réponse. Pour cela, on peut exploiter la relation suivante :

$$P(i \text{ répond à } t+1 \mid i \in a_{t,k}) = \prod_{\theta=t-k+2}^{t+1} P(i \text{ répond à } \theta \mid i \text{ répond à } \theta-1) P(i \text{ répond à } t-k+1 \mid i \in a_{t,k})$$

Cette relation n'est valable en théorie que si la non-réponse totale une année donnée devient définitive les années suivantes, selon le schéma qui suit. Cela la fragilise par rapport à l'approche précédente.



En pratique, elle peut être appliquée si la probabilité qu'un non-répondant une année donnée devienne répondant l'année suivante est (très) faible. Dans le cas présent, c'est à juger à l'expérience, mais il est possible que ce soit le cas, ne serait-ce que parce que le protocole de collecte stipule qu'un refus une année donnée devient non-réponse définitive (il reste donc à étudier le comportement des « impossibles à joindre »).

L'intérêt essentiel de cette décomposition tient à la procédure d'estimation des termes $P(i \text{ répond à } \theta \mid i \text{ répond à } \theta-1)$ qui peut s'appuyer sur l'information collectée en $\theta-1$ et aux dates antérieures puisque la procédure logistique s'applique sur l'ensemble des répondants à $\theta-1$. En contrepartie, les calculs sont plus contraignants qu'avec la première méthode. Numériquement, parce qu'on néglige certaines situations « raisonnablement négligeables », il y aura une légère sous--estimation chronique des probabilités de réponse, que l'on imagine pouvoir être corrigée de manière satisfaisante par les redressements.

2.4. Les redressements

Il s'agit ici de modifier le poids longitudinal « brut » corrigé de la non-réponse totale afin de produire, pour un ensemble de variables auxiliaires, des estimations égales par construction aux vrais totaux, issus pour leur part de sources externes (Deville et Särndal, 1992). La problématique est classique, mais la population d'inférence reste Ω_t (et non Ω_{t+1}) privée des « morts » entre les dates des collectes t et $t+1$: on se trouve donc dans le cadre d'une estimation sur un domaine puisque les individus de Ω_t encore enquêtables à $t+1$ constituent un domaine de Ω_t).

Une première opération de calage peut toujours être effectuée très en amont, sous-échantillon par sous-échantillon, à partir des variables de la base de sondage. L'échantillon est alors un échantillon de logements et on retrouve un contexte classique. Ce processus de calage ne peut qu'améliorer la qualité, mais probablement de manière mineure parce que l'information auxiliaire est celle de la base de sondage (donc une information plutôt obsolète pour les sous-échantillons les plus anciens qui se trouvent en fin de vie) et en aucun cas un ensemble de variables actualisées. Dans ce cas, les opérations d'adaptation des poids exposées au 2.1. et au 2.3. doivent être postérieures à ce calage initial, c'est-à-dire que les $W_i(t,k)$ introduits au 2.1. sont des poids déjà redressés en ce sens.

Une seconde opération, plus naturelle, plus commune, et surtout plus efficace lorsqu'elle peut être mise en œuvre (mais qui n'empêche en rien l'opération précédente), consiste à utiliser les structures externes provenant d'une grosse enquête régulière, comme l'enquête sur l'emploi ou comme les enquêtes annuelles de recensement. Ainsi, on bénéficie d'une information actualisée et on peut réduire plus efficacement la variance d'échantillonnage. L'information individuelle provient du questionnaire SILC, tandis que l'information agrégée est, soit une moyenne annuelle de diverses structures trimestrielles estimées par l'enquête sur l'emploi, soit une estimation annuelle issue du recensement.

L'échantillon longitudinal est fondamentalement un échantillon d'individus physiques, et de ce fait on a spontanément tendance à faire porter le calage exclusivement sur des variables individuelles. En réalité, il est possible d'inclure dans le processus de calage des structures portant sur les logements. En effet, il suffit de se souvenir que chaque individu-panel conduit naturellement à un logement (celui dans lequel il réside) et on peut ainsi utiliser des variables auxiliaires se rapportant soit aux individus, soit aux logements, ou sinon aux deux unités. Ce type de calage est relativement courant et est appliquée dans des enquêtes majeures comme l'Enquête sur la population active de Statistique Canada (Singh et coll., 1990). De notre côté, nous nous limiterons à un calage sur des informations individuelles seulement.

Dans ces conditions, partant du poids de l'individu i après traitement de la non-réponse totale à la date $t+1$ (selon l'une des techniques suggérées au 2.3.), le calage devrait assurer en théorie :

$$\sum_{r_{t+1}} \dot{W}_i^{t,t+1} \cdot X_i^t = \sum_{\tilde{\Omega}_t} X_i^t$$

où $\tilde{\Omega}_t$ est la population Ω_t privée des individus morts entre t et $t+1$, tandis que X_i^t est un vecteur d'informations auxiliaires relatives à i et collectées l'année t. L'échantillon r_{t+1} des répondants à $t+1$ est l'échantillon « central » qui, correctement pondéré, représente $\tilde{\Omega}_t$ et sur lequel on construit l'estimation longitudinale (voir 2.5). C'est donc sur cet échantillon qu'il faut appliquer le redressement, par exemple en utilisant le logiciel de redressement CALMAR (Sautory, 1991).

En pratique, autant on estime (assez) facilement et avec une (très) bonne précision le total naturel $\sum_{\Omega_t} X_i^t$, autant obtenir $\sum_{\tilde{\Omega}_t} X_i^t$ est un objectif insurmontable parce qu'on ne sait pas, sauf exception, tenir compte des morts entre t et $t+1$ pour établir une telle statistique. Cependant, à l'exception du nombre total d'individus, on peut considérer que les totaux marginaux reflètent des structures, si bien qu'un décalage d'une seule année n'aura probablement pas d'impact significatif. L'argument est d'ailleurs renforcé par le fait que les totaux sur lesquels on se cale sont eux-mêmes des estimations entachées d'une erreur d'échantillonnage. En revanche, il serait souhaitable de pouvoir ajuster le nombre total d'individus, qui représente un effectif. Une solution consiste à estimer *a priori*, via des sources démographiques, un rapport μ qui représente la part de population du champ à t encore vivante à $t+1$, et à effectuer le calage sur les marges

$$\mu \cdot \sum_{\Omega_t} X_i^t \text{ au lieu de } \sum_{\tilde{\Omega}_t} X_i^t .$$

Il est à noter qu'il existe une piste alternative plus rigoureuse, mais dont on peut craindre *in fine* une moindre efficacité. En effet, on peut envisager d'effectuer un calage qui prenne en compte les morts entre t et $t+1$, en assurant :

$$\sum_{i \in \delta_{t+1}} \dot{W}_i^{t,t+1} \cdot X_i^t = \sum_{\Omega_t} X_i^t$$

où δ_{t+1} est la réunion de l'échantillon r_{t+1} et des individus de r_t morts entre t et $t+1$. Le poids des individus de r_{t+1} en entrée du calage est le poids d'échantillonnage (issu de la partie 2.1) divisé par la probabilité de réponse estimée à $t+1$ (voir partie 2.3), tandis que le poids à affecter en entrée de calage aux individus de r_t morts entre t et $t+1$ est leur poids d'échantillonnage (issu de la partie 2.1) divisé par la probabilité de réponse à t. L'opération s'achève, après le calage, en abandonnant purement et simplement tout ce qui concerne les morts entre t et $t+1$. Cette approche évite de faire des hypothèses plus ou moins acceptables sur les marges (contrairement à la première méthode), mais elle a le défaut de générer une augmentation de variance parce que l'estimation finale porte sur un domaine (les individus du champ vivant à $t+1$, soit $\tilde{\Omega}_t$) alors que le calage a été conçu sur une population plus vaste (les individus du champ vivant à t, soit Ω_t). Il n'est donc pas évident qu'il s'agisse de la meilleure méthode.

2.5. Estimation longitudinale finale

On rencontre d'emblée une difficulté particulière avec les non-répondants à t qui répondent à $t+1$. Ce contexte s'applique à certaines personnes impossibles à joindre à t, mais néanmoins joignables en $t+1$ (on rappelle que ce ne peut pas être le cas pour des refus à t puisque le protocole de collecte conduit à ne pas re-interroger en $t+1$ un individu refusant de répondre à t). Dans ce contexte, il faut bien comprendre que l'individu a nécessairement répondu en $t-1$, faute de quoi il ne peut pas être interrogé en $t+1$. En effet, après deux années de non-réponse consécutives, quelle que soit la cause de la non-réponse, on abandonne définitivement le suivi de l'individu-panel. Néanmoins, on peut penser qu'il devrait s'agir d'une situation peu fréquente (cela reste néanmoins à vérifier, à l'expérience…) et, surtout, dans laquelle une imputation de Y à la date t s'avère particulièrement favorable, puisqu'on dispose par construction d'une valeur Y à $t-1$ et d'une valeur Y à $t+1$. On peut alors imputer Y_i^t pour les individus i de r_{t+1} qui ne sont pas dans r_t.

Passée cette étape préliminaire, on part des poids individuels calés $\dot{W}_i^{t,t+1}$ à l'issue de l'étape précédente (voir 2.4.) et on forme

$$\hat{\Delta}_{t,t+1} = \sum_{r_{t+1}} \dot{W}_i^{t,t+1} \cdot \left(Y_i^{t+1} - \tilde{Y}_i^{t} \right)$$

où $\tilde{Y}_i^{t}$ est égal, soit à la vraie valeur Y_i^t, soit à la valeur imputée lorsque ça s'impose. $\hat{\Delta}_{t,t+1}$ est donc l'estimateur de la différence entre les totaux de Y considérés respectivement à $t+1$ et à t, tous deux définis sur $\tilde{\Omega}_t$, population à t encore vivante à $t+1$. *A priori*, les poids $\dot{W}_i^{t,t+1}$ ne sont utilisés que dans la 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.

Si les individus de r_{t+1} qui ne sont pas dans r_t sont jugés trop nombreux, ou que l'imputation des Y_i^t n'apparaît pas souhaitable, il faut renoncer à travailler sur des différences de valeurs individuelles, et utiliser par exemple l'alternative

$$\Delta^*_{t,t+1} = \sum_{r_{t+1}} \dot{W}_i^{t+1} Y_i^{t+1} - \sum_{r_t^*} \dot{W}_i^{t} Y_i^{t}$$

où r_t^* est l'échantillon répondant r_t privé des morts entre t et $t+1$. Le poids $\dot{W}_i^t$ (respectivement $\dot{W}_i^{t+1}$) est le poids obtenu à partir de $W_i(t,k)$, après correction de la non-réponse à t (respectivement à $t+1$) et calage éventuel sur des totaux définis sur $\tilde{\Omega}_t$.

3. La pondération transversale

Il s'agit de pratiquer une inférence sur la population globale du champ de l'enquête à la date courante, soit Ω_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 cette enquête, le complément en question 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 couvre 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. Ce défaut de couverture est en général considéré comme négligeable parce qu'il est en partie corrigé par l'utilisation du redressement (voir 3.4).

3.1. Le principe de pondération initiale, en l'absence de non-réponse totale

La technique centrale utilisée pour produire les poids transversaux est la méthode de partage des poids (Lavallée, 2002). L'année t, on dispose de neuf sous-échantillons panels $a_{t,k}\left(1 \le k \le 9\right)$. On suppose dans un premier temps qu'il n'y a pas de non-réponse.

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). Pour cela, il faut 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 2.1). Dans ces conditions et en régime stationnaire, le poids transversal W_i^t 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 :

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

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[11]).

Comme dans le cas longitudinal (voir 2.1), ce calcul de poids 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-panel 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 au 1.3. En contrepartie, la formule de pondération a un

[11] L'échantillonnage de logements dans l'échantillon-maître et la BSLN s'appuie sur un principe de non ré-interrogation des logements déjà tirés.

inconvénient qui est sa (relative) complexité, à la fois sur le plan théorique et lors de la phase de programmation informatique.
En phase de montée en charge (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 :

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

En 2006, ce sera :

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

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

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

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 puisque 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 $a_{t,k}$ $(k \geq 2)$, on ne couvre pas la population des ménages constitués exclusivement d'« immigrants » (au sens large) entre $t-k+1$ et t. Notons que raisonner en année civile n'est pas ici très rigoureux et il faudrait considérer des périodes de mai à mai, donc couvrant une année complète mais entre deux campagnes de collecte. Pour formaliser le contexte et aboutir au poids transversal final, on notera $\Omega_{\alpha,t}^{immig}$ la population d'« immigrants » (au sens large) présente à t dans des ménages ne comprenant que des immigrants échantillonnables après l'année α [12], avec $t-8 \leq \alpha \leq t-1$.

A la date t, la population complète Ω_t est partitionnée en neuf composantes : les huit sous-populations $\Omega_{\alpha,t}^{immig}$, 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[13] à $t-8$. Par ailleurs, on note $\tilde{u}_{t,k}$ l'échantillon transversal à t issu du panel $a_{t,k}$ (soit $\bigcup_{k=1}^{9} \tilde{u}_{t,k} = \tilde{u}_t$) et $Y_{\alpha,t}^{immig}$, le vrai total des Y_i^t défini sur $\Omega_{\alpha,t}^{immig}$. On a alors, suite au partage des poids effectué pour tout k variant de 2 à 9 :

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

et

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

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

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

$$\frac{1}{9}\sum_{k=1}^{9}\sum_{j\in\tilde{u}_{t,k}} \tilde{W}_j(t,k)\cdot Y_j^t = \frac{1}{9}\sum_{j\in\tilde{u}_t} \tilde{W}_j(t,k)\cdot Y_j^t$$

Cependant, les panels utilisés en France ont 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 $\Omega_{\alpha,t}^{immig}$. Dans ces conditions, on vérifie que tout individu j de $\Omega_{\alpha,t}^{immig}$ qui se trouve finalement dans l'échantillon transversal $\tilde{u}_t$ aura un poids transversal brut $W_j(t)$ égal à la valeur issue du partage des poids (c'est-à-dire $\tilde{W}_j(t,k)$) 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 $\Omega_{\alpha,t}^{immig}$ (donc la grande majorité des cas) aura un poids final égal à $\tilde{W}_j(t,k)/9$. On remarquera par ailleurs que si j se trouve dans $\Omega_{\alpha,t}^{immig}$, il ne peut être enquêté qu'au travers de $a_{t,1}, a_{t,2}, \ldots, a_{t,t-\alpha}$.

[12] Plus précisément, il faudrait dire « échantillonnables à partir d'une date strictement postérieure à la date de collecte de l'année α ».

[13] 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.

Durant la période de montée en charge, il faut adapter les pondérations. En 2005, les individus de $\Omega_{2004,2005}^{immig}$ 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 $\Omega_{2005,2006}^{immig}$ auront un poids égal à celui du logement dans lequel ils résident et qui reflète directement le tirage de $a_{2006,1}$, ceux de $\Omega_{2004,2006}^{immig}$ 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.

3.2. Le traitement de la non-réponse partielle

Pour les individus-panels, il faut utiliser l'imputation conçue dans le cadre longitudinal. En effet, cette imputation est *a priori* plus pertinente parce qu'elle utilise davantage d'information, et il est hautement souhaitable de ne manipuler qu'une seule valeur imputée par combinaison individu-variable.

Pour les cohabitants, c'est la problématique habituelle, sachant que l'on ne dispose pas, par définition, de séries chronologiques et que l'information auxiliaire utilisable ne peut donc être qu'issue du questionnaire de l'année en cours, voire de la base de sondage.

3.3. Le traitement de la non-réponse totale

Il y a deux types de non-réponse totale : la non-réponse des individus-panels au moment du tirage « initial » à t_0, c'est-à-dire lorsqu'on échantillonne le panel entrant, et la non-réponse à la date courante t (qui touche tous les individus). Cette vision dichotomique de la non-réponse n'est peut-être pas la seule possible, mais elle est pratique parce qu'elle permet de distinguer le traitement des individus-panels non répondants de celui des cohabitants non répondants. On rappelle que le protocole de collecte a été fixé de façon à ce qu'un non-répondant pour cause de refus ne soit pas ré enquêté alors qu'un non-répondant pour cause d'absence de longue durée ou classé « impossible à joindre » fait l'objet d'une seconde tentative d'enquête l'année suivante. Notons que si cette nouvelle tentative se solde par une nouvelle non-réponse, le suivi de l'individu-panel est définitivement abandonné, quelle que soit la cause de cette nouvelle non-réponse.

Pour se placer dans un contexte simplifié sans perdre en rigueur, on va accepter de résumer l'histoire du suivi d'un individu-panel par ses comportements (en matière de réponse) à deux dates, respectivement la date initiale t_0 (définie comme la date du tirage du panel dans lequel il se trouve) et la date courante t. En effet, ce qui se passe entre ces deux dates n'a pas d'importance pour la pondération. La cause de la non-réponse à la date t est la dernière cause connue. Par exemple, si l' individu-panel refuse de répondre « une année quelconque » entre t_0 et

t, il est classé en refus à t. S'il est deux années de suite impossible à joindre, il restera perpétuellement considéré comme impossible à joindre. Mais s'il est impossible à joindre une année donnée seulement et qu'il recommence à répondre ensuite, on oubliera définitivement l'année où il y a eu non-réponse (la non-réponse ponctuelle est donc « transparente » s'il ne s'agit pas de l'année durant laquelle elle se manifeste). En particulier, on convient du principe suivant : si un individu-panel est non répondant à t_0 (donc lorsqu'il est enquêté pour la première fois) mais qu'il est encore suivi sur le terrain à la date t en tant qu' individu-panel (ce qui signifie que la cause de non-réponse initiale n'est pas un refus mais une absence de contact), alors on le traitera par la suite en tous points comme s'il était répondant initial. On procède alors par étapes de la manière suivante :

a) On estime pour chaque individu j du panel $a_{t,k}$, sa probabilité de réponse « initiale ». Pour cela, on peut utiliser l'approche classique - généralement un modèle logistique - ou une procédure de calage « en une étape »[14]. On note $\Phi_j(t,k)$ cette probabilité estimée.

b) Chaque échantillon $a_{t,k}$ est composé de deux parties : les répondants à la date initiale $t-k+1$, constituant le sous-échantillon noté $r_{t,k}$, et les non-répondants à cette même date, constituant le complémentaire. On calcule alors les poids individuels suivants, pour tout j de $a_{t,k}$:

$$W_j^*(t,k) = \frac{W_j(t,k)}{\Phi_j(t,k)} \cdot 1_{j \in r_{t,k}}$$

où $W_j(t,k)$ est le poids (introduit au 2.1) qui découle de l'échantillonnage $a_{t,k}$. et $1_{j \in r_{t,k}}$ désigne la variable aléatoire qui vaut 1 si j répond et 0 sinon. Autrement dit, si l'individu-panel est répondant initial (après application du principe exposé précédemment), on divise son poids de tirage par sa probabilité de réponse (estimée), mais s'il est non-répondant initial on lui impose un poids nul.

c) On applique un partage des poids exactement comme exposé au 3.1, en remplaçant $W_j(t,k)$ par $W_j^*(t,k)$. On peut évidemment choisir l'une ou l'autre des deux options présentées au 3.1.

Il est fondamental de noter que le dénombrement des liens n'est en rien modifié par la non-réponse. D'ailleurs, la définition même du lien n'est pas remise en cause par la non-réponse.

Si on opte, par exemple, pour l'option 1 (qui raisonne directement sur l'ensemble des neuf sous-échantillons), on calculera, pour tout i du ménage m :

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

[14] À condition de disposer des vrais totaux des variables explicatives de la non-réponse.

d) On s'intéresse alors à l'échantillon transversal $\tilde{u}_t$ (obtenu par le suivi des individus-panels des sous-échantillons $r_{t,k}$ et après application du protocole exposé au début de la partie 3). Il comprend des individus répondants - formant un échantillon noté $\tilde{r}_t$ - et des individus non répondants. On peut estimer, par un modèle logistique par exemple, la probabilité de réponse θ_i de chaque individu i de $\tilde{u}_t$. Pour cette phase, peu importe qui est individu-panel et qui est cohabitant. Comme de tradition, le poids de tout individu non répondant est nul, mais le poids final de tout individu i répondant à t est

$$\frac{W_i^{*t}}{\theta_i}$$

Ce poids final permet de construire un estimateur sans biais du total à la date courante.

Deux points sont à souligner :

- Le calcul des liens doit se faire de manière « habituelle », ce qui veut dire qu'il faut obtenir l'information nécessaire et suffisante pour cela, individu par individu, dans tout ménage répondant à la date courante[15], que ces individus soient ou non répondants. Le fait qu'il y ait non-réponse détériore certainement la qualité de la mesure des liens, mais pas nécessairement de manière forte, d'une part, parce que l'information sur l'existence ou non d'un lien s'obtient par définition auprès des individus échantillonnés à la date courante et non pas auprès de ceux qui sont tirés et, d'autre part, parce que si un individu est non-répondant à la date courante, l'information le concernant peut très bien être fournie par une autre personne du ménage. Il est, en effet, assez fréquent d'interroger des « proxy », et on peut penser que les données qui permettent de quantifier les liens s'obtiennent facilement et avec une fiabilité satisfaisante.

- A l'issue de l'étape c), les poids individuels restent égaux pour tous les individus d'un même ménage, pour tout ménage répondant à la date courante t. Un ménage entièrement non répondant à t n'a pas besoin d'être pondéré (en fait, son poids vaut 0). Cela étant, un individu-panel suivi jusqu'en t et non répondant à t participe « normalement » au partage des poids, c'est-à-dire qu'il est pris en compte dans le numérateur de W_i^{*t} exactement comme s'il était répondant (on rappelle que $r_{t,k}$ désigne les individus-panels répondants à t_0 et non ceux qui répondent à t). Cette précision a son importance parce qu'elle signifie en pratique qu'il faut être capable d'identifier, dans tout ménage répondant à la date courante, tous les individus (répondant ou non) de ce ménage qui sont des individus-panels.

Notons, pour information, que si à l'avenir le protocole de traitement des non-répondants est modifié en ce sens où il impose la ré interrogation systématique chaque année des non-répondants (ce qui serait cohérent avec un objectif de limitation du biais, mais évidemment une pratique peu compatible avec la recherche d'économie), cette technique de repondération aura un atout important. En effet, en présence d'un ménage non répondant à t_0 (c'est-à-dire dans lequel chaque individu-panel est non répondant), il est inutile d'assurer un suivi des individus de ce ménage puisque ces derniers n'apporteront aucune contribution à la pondération finale. Cela est très appréciable parce qu'on imagine facilement que le suivi de ménages non répondants (autrement dit, l'activation des liens) soit difficile, voire impossible, dans certains contextes[16]. Le protocole actuel de collecte de SILC, qui limite les cas de ré interrogation de non-répondants, affaiblit évidemment cet argument.

[15] Si le ménage est entièrement non-répondant à t, la question ne se posera pas. L'étape (d) explique en effet que ce ménage ne participera pas du tout à la pondération courante.
[16] Si un ménage refuse de répondre dès la première année, on peut douter d'être en mesure de le suivre s'il déménage par la suite.

3.4. Les redressements

Une approche standard consiste à considérer l'intégralité de l'échantillon répondant transversal $\widetilde{r}_t$ avec les poids corrigés de la non-réponse totale selon les principes du 3.3. On pratique alors un redressement « classique » en se calant sur des structures obtenues à partir de données auxiliaires caractérisant la population du champ à la date transversale t. Une alternative est offerte par le calage « en une étape », qui s'appuie directement sur l'échantillon $\widetilde{r}_t$ et les poids W_i^{*t} en évitant de dérouler l'étape d) de la partie 3.3.

Concernant les sources, dans l'attente de pouvoir utiliser les informations des enquêtes annuelles de recensement, on se fonde sur des marges annualisées produites par l'enquête trimestrielle « Emploi »[17]. A noter que le calage simultané individus-ménage constitue la règle à suivre. On aboutit *in fine* à des poids redressés $\dot{W}_i^t$ (concrètement, ces poids sont fournis par le logiciel de calage CALMAR).

On rappelle qu'il est possible de produire des poids qui réalisent un calage simultané de deux échantillons sur deux systèmes d'informations indépendants : d'une part un calage de l'échantillon d'individus-panels sur des totaux produits à partir de la base de sondage (actuellement le recensement général de 1999) et, d'autre part, un calage de l'échantillon transversal sur les totaux issus de l'enquête trimestrielle emploi. Si elle était appliquée, cette méthode serait en théorie plus performante que la précédente. Cela étant, elle est sensiblement plus compliquée et on peut s'attendre à ce que le gain que générerait sa mise en oeuvre soit numériquement faible. Aussi, nous ne la préconisons pas durant cette période de rodage du système de pondération, et suggérons d'y réfléchir ultérieurement.

3.5. Estimation transversale finale

On part des poids corrigés de la non-réponse totale et redressés $\dot{W}_i^t$, et on forme

$$\hat{T}_t = \sum_{i \in \widetilde{r}_t} \dot{W}_i^t \cdot Y_i^t$$

pour estimer un total à t, et

$$\hat{\hat{\Delta}}_{t,t+1} = \sum_{\widetilde{r}_{t+1}} \dot{W}_i^{t+1} \cdot Y_i^{t+1} - \sum_{\widetilde{r}_t} \dot{W}_i^t \cdot Y_i^t$$

pour estimer une évolution (selon les critères « transversaux ») entre les dates t et $t+1$.

[17] Voir note interne à la Direction des Statistiques Démographiques et Sociales de l'INSEE, n°11/F401 du 31 janvier 2005.

Bibliographie

[1] Ardilly, P. (2006). *Les techniques de sondage, 2nde édition*. Éditions Technip, Paris.

[2] Deville, J.-C., Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, Vol. 87, No. 418, juin 1992, pp. 376-382.

[3] Deville, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. *Recueil 1998 de la Section des méthodes d'enquête*, Société statistique du Canada, pp. 103-110.

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

[5] Lock Oh, H., Scheuren, F.J. (1983). *Weighting Adjustment for Unit Nonresponse. In Incomplete Data in Sample Surveys* (Madow, W.G., Olkin, I., Rubin, D.B., Éditeurs), Vol. 2, Academic Press, New York, pp. 143-184.

[6] Merkouris, T. (2001). Estimation transversale dans le cas des enquêtes auprès des ménages à panels multiples. *Techniques d'enquêtes*, Vol. 27, No. 2, pp. 189-200.

[7] Naud, J.-F. (2004). *Etude sur des modifications possibles à l'ajustement pour la non-réponse*. Document interne de l'enquête sur la dynamique du travail et du revenu, Statistique Canada, 13 juillet 2004.

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

[9] Sautory, O. (1991). *La macro SAS : CALMAR (redressement d'un échantillon par calage sur marges)*. Document interne de l'INSEE, Paris.

[10] Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

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