PANEL SOCIO-ECONOMIQUE

"LIEWEN ZU LËTZEBUERG"

DOCUMENT PSELL N° 102 AVRIL 1997

REPRESENTATIVITE ET PONDERATION DES ECHANTILLONS DU PSELL.2 1994-1995

par

Bernard GAILLY

Docteur en sociologie

CEPS/Instead
Differdange

Grand-Duché de Luxembourg

1997

Présentation du programme P S E L L

Les informations présentées dans ce cahier proviennent du programme PSELL développé par la Division "Ménages" du C.E.P.S./Instead. Grâce à ce programme, le Grand-Duché de Luxembourg dispose d'un instrument exceptionnel permettant de connaître les conditions d'existence des personnes et des ménages qui y vivent : le panel socio-économique "Liewen zu Lëtzebuerg" (PSELL).

Dans le cadre de ce programme, de nombreuses informations sont récoltées chaque année sur les principaux aspects de la vie de la population du pays :

- conditions de logement, équipement et composition des ménages
- principales dépenses
- précarité
- endettement
- position scolaire des enfants
- position socioprofessionnelle des adultes
- revenus, ...

Cette recherche a débuté en 1985 par des interviews auprès d'un échantillon de 6110 personnes réparties dans 2012 ménages. Chaque année, cette enquête est reprise et le même échantillon est suivi année après année. Bien sûr, cet échantillon évolue, tout comme la population du pays (naissances, mariages, décès, émigration, ...). En 1994, il était composé de 4966 personnes vivant dans 1809 ménages.

En 1994, cette étude a fêté son dixième anniversaire. Sur le plan scientifique, cet événement représentait certainement un succès parce qu'il est très rare qu'un même programme de recherche puisse être développé sur une période aussi longue. Une large part de ce succès revient toutefois aux milliers de personnes qui, au fil des années, ont accepté de recevoir chez elles nos enquêteurs et de participer à ce vaste programme ; par leur contribution, elles ont permis de réunir un capital de connaissances inestimable, couvrant dix ans de la vie de la population de notre pays.

Les données récoltées ont déjà fait l'objet de nombreuses études publiées pour la plupart au CEPS/Instead dans les séries suivantes :

- Documents PSELL (voir liste en annexe)
- Notes de Recherche
- PSELL INFO
- ECOCEPS.
- Population & Emploi - Série “Conditions de vie”

Pour plus d’informations
(I. BOUVY)
Tel: (00 352) 58 58 55- 513
Fax: (00 352) 58 55 60

Document produit par le

CEPS/Instead

Centre d'études de populations, de pauvreté et de politiques socio-économiques
B.P. 48 - L 4501 Differdange

Président : Gaston Schaber

Document PSELL N°102 - ISBN 2 - 87987 - 147 - 6 - 1997

SOMMAIRE

ATTENTION !

Les utilisateurs du PSELL.2 sont invités à lire attentivement les indications qui suivent surtout s'ils ont déjà utilisé les données du PSELL.1 en recourant au système de pondération. Un certain nombre de modifications ont été apportées à la technique de pondération et les résultats de ces modifications peuvent paraître troublantes si l'on saisi mal leurs origines : certains concepts de base ont été soit clarifiés soit modifiés et certains dispositifs purement techniques ont été mieux adaptés à la nature de l'évolution des échantillons.

1. Qu'est-ce qu'une population et qu'est ce qu'un échantillon ?
2. Qu'est-ce qu'un échantillon représentatif ?
3. Qu'est ce qu'un échantillon longitudinal et qu'est-ce qu'un échantillon transversal ?
4. Qu'est-ce qu'un taux de réponse correctif ?
5. A quoi sert-il de "pondérer" les unités d'observation d'un échantillon ?
6. Faut-il repondérer l'échantillon longitudinal et comment ?
7. Comment sauvegarder la représentativité annuelle des échantillons transversaux d'un panel ?

Nous n'envisagerons pas les aspects trop techniques de ces différentes procédures. Nous essayerons simplement de montrer que ce n'est pas en vain que l'on s'efforce d'apporter le plus de soin possible à ces différents aspects de la gestion des échantillons générés par un panel.

CHAPITRE I

QU'EST-CE QU'UNE POPULATION ET QU'EST-CE QU'UN ECHANTILLON ?

1. QU'EST-CE QU'UNE POPULATION ET QU'EST-CE QU'UN ECHANTILLON ?

La population est un univers auquel on souhaite étendre l'ensemble des conclusions d'une étude. Ces conclusions ne pourront être étendues au-delà de cet univers.

La population visée par le PSELL se limite aux personnes liées directement ou indirectement au système de sécurité et de protection sociale luxembourgeois.

Par exemple, les conclusions des études relatives aux "revenus" ou à la "pauvreté" ne concernent donc en aucun cas "l'ensemble des personnes qui résident au Luxembourg". Les fonctionnaires étrangers, les agents de sociétés étrangères installés provisoirement au Luxembourg et liés au système de sécurité sociale de leur pays ainsi que les Luxembourgeois (frontaliers) travaillant en dehors du pays ne sont pas visés par cette étude.

L'échantillon (s) extrait de cette population (U) est formé par l'ensemble de toutes les unités sélectionnées avant toute observation et quel que soit le statut de ces unités après l'observation.

S est composé de m personnes appartenant à U composé de M personnes. Il se présente généralement sous la forme suivante après l'observation. Pour m adresses distribuées, s comprend :

- **m1** "hors champ"

  dont :
  - décès
  - émigrés
  - en ménages collectifs
  - autres erreurs liées à la nature de la population

- **m2** "enquêtes réalisées"

- **m3** "échecs"

  dont :
  - refus
  - adresses inexactes ou inconnues
  - impossible à réaliser (maladie, handicap, ...)
  - impossible à contacter (absences, ...)
  - autres impossibilités concernant des personnes dans le champ.

L'échantillon est composé de toutes les unités m sélectionnées. Soit les enquêtes réalisées, les hors champ et les échecs.

Les répondants (soit m2 + m1) comprennent non seulement les enquêtes réalisées (m2) mais aussi les "hors champ" (m1).

Les "hors champ observés", soit m1, étaient bien recensés dans la population au moment où l'échantillon a été tiré[1]. Ils ont donc influencé le tirage de l'échantillon. Ce sont des répondants pour lesquels toutes les variables mesurées prendront une valeur "0" ou "non-réponse". Nous verrons par la suite qu'il n'y a pas lieu de les écarter de la définition et de la composition de l'échantillon.

Le premier échantillon du PSELL.2 est sélectionné au sein d'une population exhaustive de "**titulaires principaux de revenus"** correspondant au fichier de l'Inspection Générale de la Sécurité sociale soit 154 534 titulaires principaux. Ces titulaires principaux sont aussi bien des personnes qui travaillent que des pensionnés ou des enfants titulaires d'une pension d'orphelin.

Cet échantillon compte 5 713 titulaires de revenus sélectionnés de manière aléatoire :

- chaque unité a *la même probabilité d'être choisie* (pas de sur-représentation, pas de stratification)
- et cette *probabilité est différente de 0* (sans biais).

Aucun sous-groupe n'est sur-représenté ni sous-représenté de manière volontaire. La sous-représentation de certains groupes pourrait être regrettable si leur intérêt particulier pour l'étude ne se révélait que plus tard.

Les titulaires de revenus ne sont que des **unités de sélection** conduisant à des **unités d'observation** : les ménages et les membres de ces ménages.

Chaque titulaire de revenu conduit à une adresse. A chaque adresse correspond un ménage. Les 5 713 titulaires de revenus conduisent donc à 5 713 ménages.

Ces 5 713 ménages forment un échantillon qui se répartit de la manière suivante :

[1] Il est important de noter qu'aucune correction ne peut être introduite dans la base de sondage suite à l'observation sous peine de fausser les probabilités de sondage.

STATUT	*FREQUENCES*	*POURCENT*
Enquêtes réalisées	2978	52.1
"Hors Champ"	269	4.7
- décédés	*16*	
- émigrés	*61*	
- ménages collectifs	*192*	
"Echecs"	2466	43.2
- refus	*1922*	
- toujours absents	*410*	
- mauvaises adresses	*134*	
TOTAL	***5713***	***100.0***

La probabilité d'inclusion des ménages est de 3.69% ou, en d'autres termes, chaque titulaire principal en représente 27 (154 534 / 5 713).

La probabilité d'inclusion des répondants est de 2.1 % ou, en d'autres termes, chaque titulaire principal en représente 50 (154 534 / 3 247).

La probabilité de sélection des répondants (y compris les "hors champ") est de 56.8 %.

Si un titulaire principal conduit à une et à une seule adresse, donc à un et à un seul ménage, l'inverse n'est pas nécessairement vrai. Un ménage peut être composé de plusieurs titulaires principaux. En d'autres termes, à partir de la base de sondage (le fichier de l'I.G.S.S.) plusieurs titulaires peuvent conduire au même ménage. L'échantillon des titulaires principaux était un échantillon aléatoire simple. La sélection de chaque unité était équiprobable. L'échantillon des ménages ne répond plus à cette condition.

La probabilité d'inclusion des ménages est inégale. Plus le nombre de titulaires principaux conduisant à un même ménage est élevé, plus le ménage a une probabilité d'inclusion élevée dans l'échantillon.

Afin de rétablir l'égalité des probabilités d'inclusion des ménages (équiprobabilité), ces derniers reçoivent un poids inverse au nombre de titulaires principaux susceptibles de conduire à leur adresse. On notera le fait que ce nombre n'est connu qu'au terme de l'observation des ménages.

Au sein de chaque ménage, tous les membres sont observés. Lorsque l'équiprobabilité de sélection des ménages est rétablie, tous les individus appartenant à ces ménages peuvent retrouver eux aussi une probabilité de sélection égale. Chaque membre reçoit le poids du ménage et l'équiprobabilité de ces unités d'observation est rétablie.

L'ensemble des membres sélectionnés au cours du tirage du premier échantillon sont des **membres longitudinaux** : nous reviendrons sur ce point de manière plus détaillée par la suite.

Ces 10 967 individus appartenant aux 5 713 ménages forment un échantillon qui se répartit de la manière suivante :

STATUT	***FREQUENCES***	***POURCENT***
Enquêtes réalisées	8 232	75.1
"Hors Champ"	269	2.4
"Echecs"	2 466	22.5
TOTAL	***10 967***	***100.0***

Les enquêtes réalisées et les hors champ forment l'ensemble des répondants.

- Les membres observés ont reçu un poids égal au poids de leur ménage.

- Les "titulaires hors champ" conduisent aussi à des "ménages hors champ". Par définition, les ménages "hors champ" ne peuvent être observés. Nous avons estimé qu'un titulaire "hors champ" correspond à un ménage "hors champ" et à un individu "hors champ". Il reçoit donc un poids unitaire.

  Ces individus "hors champ" entrent également dans la composition de l'échantillon des individus. Ils y apportent donc leur poids.

- Les ménages ayant refusé de répondre ne peuvent être observés. Nous avons estimé qu'un refus correspond à un titulaire, à un ménage et à un seul individu. Il reçoit un poids unitaire et entre à ce titre dans le calcul de la taille de l'échantillon des individus.

CHAPITRE II

QU'EST-CE QU'UN ECHANTILLON REPRESENTATIF ?

2. QU'EST-CE QU'UN ECHANTILLON REPRESENTATIF ?

Un échantillon est représentatif de la population dont il est extrait **si et seulement si toutes les unités de la base de sondage ont une probabilité strictement supérieure à zéro d'être sélectionnées**.

Si ce critère est respecté, l'échantillon peut être utilisé **pour produire des estimations sans biais de quantités reliées à la population** qui lui correspond.

Si le fichier contient des erreurs ou des unités qui ne devraient plus s'y trouver (les hors champ), il est exclu d'écarter volontairement ces unités de l'échantillon. Ces unités appartiennent à la base de sondage. Ecarter ces personnes de l'échantillon après les avoir sélectionnées revient à faire "comme si" leur probabilité de sondage était nulle ce qui n'est manifestement pas le cas.

CHAPITRE III

QU'EST-CE QU'UN ECHANTILLON LONGITUDINAL ET QU'EST-CE QU'UN ECHANTILLON TRANSVERSAL ?

3. QU'EST-CE QU'UN ECHANTILLON LONGITUDINAL ET QU'EST-CE QU'UN ECHANTILLON TRANSVERSAL ?

L'échantillon sélectionné en 1995 est destiné à la réalisation d'un panel. Un panel est une étude qui a la particularité de poursuivre deux objectifs simultanément. D'une part, il permet de suivre la trajectoire d'un certain nombre d'individus au cours d'un certain nombre d'années. D'autre part, il doit permettre de produire chaque année des estimations sans biais de quantités relatives à la population au sein de laquelle il évolue.

1. UN ECHANTILLON LONGITUDINAL

"Suivre la trajectoire d'un certain nombre d'individus au cours d'un certain nombre d'années" est une opération strictement longitudinale qui s'effectue sur la base d'un **échantillon d'unités longitudinales** : chaque année les mêmes individus sont observés et les informations disponibles pour chacun d'eux augmente ainsi progressivement.

L'échantillon longitudinal est composé et n'est composé que des individus sélectionnés la première année. Cet échantillon **est représentatif et n'est représentatif que de la population en l'année où il a été tiré**.

L'évolution de la population de référence au fil des années n'a strictement aucun effet sur l'échantillon longitudinal ni sur son évolution.

2. DES ECHANTILLONS TRANSVERSAUX

"Produire chaque année des estimations sans biais de quantités relatives à la population au sein de laquelle l'échantillon évolue" est une opération strictement transversale ou annuelle (si les observations ont lieu chaque année). Cette opération s'effectue sur la base **d'échantillons transversaux**. Un échantillon transversal est un échantillon ponctuel dont la fonction est de **représenter l'état de la population au moment où il est observé**. Cette représentativité transversale suppose notamment que chaque année de nouveaux individus entrent dans l'échantillon longitudinal.

.

Les échantillons transversaux sont donc composés

- des individus longitudinaux
- d'individus présents dans le pays au moment du tirage du premier échantillon, qui n'ont pas été sélectionnés et qui rejoignent les individus longitudinaux dans leurs ménages
- et d'individus absents du pays au moment du tirage du premier échantillon et qui sont "nés" dans le pays depuis lors soit dans les ménages formés par les individus longitudinaux, soit par le biais d'une sélection annuelle volontaire d'un sous-échantillon représentatif de ces nouvelles "naissances" ("nouveau-nés" et immigrants).

3. UNITES LONGITUDINALES ET UNITES TRANSVERSALES

Seuls les individus sont des **unités longitudinales** parce que ce sont des unités qui ne peuvent ni se diviser ni se fusionner.

Ces unités doivent appartenir à un échantillon représentatif d'une population à un moment donné. Ce ne sont, par exemple, que les individus sélectionnés la première année du panel. Ce pourrait être aussi la combinaison d'un sous-ensemble aléatoire d'individus appartenant au premier échantillon sélectionné pour le panel et d'un sous-ensemble aléatoire d'individus sélectionnés dans la population de référence quatre années plus tard (renouvellement de l'échantillon par rotation).

Les **unités transversales** sont

- les ménages (unités qui peuvent se diviser ou se fusionner),
- les individus présents dans le pays au moment du tirage du premier échantillon qui rejoignent les individus longitudinaux dans leurs ménages
- et les individus "nés" dans le pays depuis le tirage du premier échantillon.

Leur existence dans les échantillons annuels est ponctuelle et maintient la représentativité transversale des échantillons successifs.

Si les ménages ne sont que des unités transversales, on notera toutefois le fait que les caractéristiques de ces ménages peuvent toujours être attribuées aux individus longitudinaux : les caractéristiques des ménages peuvent évoluer, les ménages peuvent changer de structure, se diviser, se fusionner mais l'unité individuelle d'observation longitudinale reste identique à elle-même et se caractérise simplement par les modifications du contexte familial dans lequel elle évolue.

CHAPITRE IV

QU'EST-CE QU'UN TAUX DE REPONSE CORRECTIF ?

4. QU'EST-CE QU'UN TAUX DE REPONSE CORRECTIF ?

Si l'on se réfère au tableau de présentation de la distribution des individus dans le premier échantillon, on peut considérer que 75.1% des individus ont participé à l'observation. Dans ce cas, on parlera d'un taux de réponse descriptif.

On peut aussi affirmer que le taux de réponse est de 77.5% des individus appartenant à l'échantillon. Dans ce second cas, il s'agit d'un **taux de réponse correctif**.

Un taux de réponse descriptif fournit des informations concernant le déroulement du processus de sélection : "269 individus hors champ dans un échantillon de 10 967 individus représentent un taux de 2.4% de hors champ". Un tel taux est aussi un taux descriptif.

Un taux de réponse correctif prend en compte toute l'information nécessaire pour corriger l'estimateur des poids des individus en fonction de la non-réponse. Il correspond au rapport entre les répondants et le nombre d'unités dans l'échantillon.

Dans ce calcul, les hors champ ne peuvent pas être considérés comme des "non-répondants".

1. TAUX DE REPONSE CORRECTIF TRANSVERSAL

La première année d'observation d'un panel fournit un et un seul échantillon purement transversal et identique à tout échantillon utilisé pour n'importe quel type d'enquête. Le taux de réponse correctif transversal des individus de la première vague du panel est donc

$$Rtr\,(t) = \frac{\text{nombre de répondants (au temps } t)}{\text{nombre d'unités dans l'échantillon (au temps } t)}$$

soit :

$$\frac{\mathbf{8232 + 269}}{\mathbf{10\ 967}} = \mathbf{77.5\%}$$

2. TAUX DE REPONSE CORRECTIF LONGITUDINAL

On obtient le taux de réponse correctif longitudinal en rapportant *le nombre d'unités longitudinales* ayant répondu à la vague *t* par rapport au *nombre d'unités dans la première vague*. Soit :

$$R_{lg}(t) = \frac{\text{nombre d'unités longitudinales qui répondent (au temps } t)}{\text{nombre d'unités dans l'échantillon de la vague 1}}$$

Ce taux de réponse peut se calculer pour l'ensemble de l'échantillon ou au niveau de différentes strates si l'on soupçonne que certaines catégories de la population ont des raisons de répondre moins souvent que d'autres.

CHAPITRE V

A QUOI SERT-IL DE "PONDERER" LES UNITES D'OBSERVATION D'UN ECHANTILLON ?

5. A QUOI SERT-IL DE "PONDERER" LES UNITES D'OBSERVATION D'UN ECHANTILLON ?

1. REPRESENTATIVITE ET PONDERATION

Il est impératif de comprendre que la pondération des unités d'un échantillon ne peut rien changer au caractère représentatif ou non-représentatif de l'échantillon. La condition sous laquelle la représentativité d'un échantillon est assurée a été clairement énoncée.

Si l'échantillon est représentatif, il permet de produire des estimations sans biais de quantités relatives à la population dont il est extrait. **La pondération sert à améliorer la qualité des résultats** en remédiant aux perturbations introduites éventuellement par d'autres sources de biais telles que, par exemple, la non-réponse.

Si un échantillon n'est pas représentatif (choix volontaire des unités observées, échantillon de "prototypes", échantillons longitudinal "cylindré"), la pondération des unités de cet échantillon ne le rendra jamais représentatif de quoi que ce soit (si ce n'est d'un univers mental qui peut, par ailleurs, avoir sa propre cohérence interne).

2. PONDERER L'ECHANTILLON INITIAL

L'échantillon initial est un échantillon transversal.

Nous avons déjà évoqué le fait qu'un poids avait été attribué aux membres de l'échantillon initial (inversement proportionnel au nombre de titulaires principaux de revenus). Ce poids a permis de rendre aux unités d'observation un caractère équiprobable que les unités de sélection de l'échantillon ne permettait pas d'assurer.

Cette technique suppose que l'on connaisse très exactement la clé de passage des unités de sélection aux unités d'observation. D'une manière générale, elle suppose que l'on connaisse la structure des liens qui unissent les unités appartenant à un univers quelconque "U1" et les unités appartenant à un autre univers "U2".

Un retour en arrière s'impose, ici, car le rétablissement de l'équiprobabilité entre les unités d'observation supposait une opération préalable dont il n'a pas encore été fait état.

Première étape

Tous les titulaires principaux de revenus n'avaient pas répondu aux sollicitations des enquêteurs (2 466 refus). La non-réponse de ces titulaires principaux entraînait la non-réponse d'un certain nombre de ménages et donc d'un certain nombre d'individus. Ce phénomène risquait donc d'introduire certains biais dans les estimations des quantités mesurées par l'étude dans la mesure où la non-réponse est "non-ignorable" et l'on dit de la non-réponse qu'elle est "non-ignorable" lorsque la probabilité de répondre ou de ne pas répondre est liée à des variables mesurées par l'enquête.

En l'occurrence, la non-réponse était liée à la distribution géographique des titulaires principaux (par canton).

Un premier poids devait donc être calculé pour les titulaires de revenus afin de compenser ces taux de non-réponse inégaux selon les cantons de résidence.

Le parallélisme entre la distribution des titulaires principaux dans le fichier de référence et dans l'échantillon qui nous a été fourni est tel qu'on peut exclure toute sur-représentation ou sous-représentation géographique des ménages au sein de l'échantillon des adresses qui ont été mises à notre disposition.

Par contre, le taux de réponse obtenu suite à l'enquête présente des disparités selon les cantons. Un taux de correction ou "poids de sélection" peut être calculé pour chaque titulaire résidant dans chaque canton.

Soit, pour chaque titulaire, **l'inverse du taux de réponse correctif transversal** dans son canton :

$$Rtr\,(t) = \frac{\text{nombre de répondants dans le canton}}{\text{nombre d'unités dans l'échantillon du canton}}$$

Une fois pondérés, les titulaires répondants se distribuent dans les cantons de la même manière que les titulaires de l'ensemble de l'échantillon. L'inégalité des probabilités de sélection selon les cantons est compensée.

Les titulaires ayant refusé de répondre sont remplacés par les poids accordés aux titulaires ayant répondu et, ceci, de manière inversement proportionnelle au taux de réponse dans chaque canton. La taille initiale de l'échantillon n'est donc pas modifiée bien que les refus n'aient pas été observés (5 713 titulaires et 5 713 ménages).

Cette pondération a un double résultat. Elle maintient le degré de précision des estimations qui seront produites : n (taille de l'échantillon) reste constante. Elle améliore la qualité des résultats en corrigeant la non-réponse non-ignorable.

Deuxième étape

Les ménages héritent de ces poids de sélection attribués aux titulaires de revenus puisque chaque titulaire conduit à un et un seul ménage.

Troisième étape

Ce n'est qu'ensuite que les ménages sont pondérés en raison inverse du nombre de titulaires principaux susceptibles de conduire à chaque ménage.

Le poids d'un ménage est donc égal à l'inverse du nombre des **titulaires principaux pondérés** susceptibles de conduire à ce ménage

C'est ce poids des ménages qui est attribué à leurs membres. La physionomie de l'échantillon initial s'en trouve bien entendu modifiée.

Comparaison des distributions des titulaires principaux selon les cantons dans la population et dans l'échantillon total

CANTON	***POPULATION FREQUENCE***	***ECHANTILLON FREQUENCE***
LUXEMBOURG	19.6	20.2
CAPELLEN	7.7	7.4
ESCH / ALZETTE	32.2	31.7
LUX. CAMPAGNE	9.3	9.3
MERSCH	5.1	4.9
CLERVAUX	2.6	2.5
DIEKIRCH	5.9	5.7
REDANGE	2.8	3.1
VIANDEN	0.6	0.6
WILTZ	2.7	2.6
ECHTERNACH	3.4	3.6
GREVENMACHER	4.6	5.0
REMICH	3.5	3.3
Donnée manquante	0.0	0.0
TOTAL	***154 534***	***5713***

Distribution des taux de réponse des titulaires selon les cantons

CANTON	*Taux de réponse*
LUXEMBOURG	.599
CAPELLEN	.574
ESCH / ALZETTE	.583
LUX. CAMPAGNE	.617
MERSCH	.523
CLERVAUX	.549
DIEKIRCH	.421
REDANGE	.588
VIANDEN	.500
WILTZ	.583
ECHTERNACH	.504
GREVENMACHER	.528
REMICH	.534
TOTAL	***.568***

Structure de l'échantillon des individus avant et après pondération

STATUT	*FREQ. 1994 avant pondération*	*FREQ. 1994 après pondération*
Enquêtes réalisées	8 232	10 497
"Hors Champ"	269	470
"Echecs"	2 466	-
TOTAL	***10 967***	***10 967***

CHAPITRE VI

FAUT-IL REPONDERER L'ECHANTILLON LONGITUDINAL ET COMMENT ?

1. INDICATION DE REPONDERATION

Le PSELL.2, comme tous les panels, est soumis à un phénomène inévitable : l'érosion de l'échantillon lors du passage de la première vague à la seconde (et ainsi pour chaque nouvelle vague d'observation). Cette érosion (ou attrition) de l'échantillon provient de la lassitude des personnes interrogées à de multiples reprises qui refusent de poursuivre leur collaboration à une étude de longue durée[1].

Cette érosion liée à la non-réponse annuelle des membres **longitudinaux réduit le nombre des répondants** et **peut être une source de biais si la non-réponse est non-ignorable**. La repondération annuelle des unités de l'échantillon longitudinal est destinée à réduire ces biais et à améliorer la qualité des résultats.

Les taux de réponse des individus longitudinaux font donc l'objet d'une analyse très intensive. Chaque année, on examine la répartition des taux de réponse selon un grand nombre de variables mesurées en 1994 (première vague d'observation de l'échantillon). L'objectif de cette procédure est de repérer les catégories d'individus importantes pour l'étude qui manifesteraient une plus faible propension à répondre que les autres catégories d'individus. Si aucune concentration de la non-réponse n'apparaît dans aucune catégorie de population stratégique pour l'étude, on est en droit de considérer que la non-réponse est un phénomène "général" sans effet sur la qualité des résultats de l'étude. Dans le cas contraire, il y a lieu de repondérer les membres longitudinaux en s'efforçant de compenser cette distribution inégale des taux de réponse.

Il s'agit d'"expliquer" le mieux possible la non-réponse. La part de la non-réponse qui reste "inexpliquée" par ces variables est "ignorable" parce qu'elle n'est pas liée à des variables mesurées par l'enquête et que ces personnes qui se lassent ne se distinguent des autres personnes de l'échantillon que par le seul fait ... qu'elles se lassent de répondre.

2. LE TAUX DE REPONSE CORRECTIF LONGITUDINAL

Le taux de réponse correctif longitudinal se calcule selon la formule

$$R_{lg}(t) = \frac{\text{nombre d'unités longitudinales qui répondent (au temps } t)}{\text{nombre d'unités dans l'échantillon de la vague 1}}$$

[1] On distingue ce type de "non-réponse totale de vague" ou "chronique" de la "non-réponse partielle" où certaines observations visées par l'enquête sont manquantes pour certaines unités d'observation (dans ce dernier cas, on aura généralement recours à différentes procédures "d'imputation").

La variable dépendante à prendre en compte est "le taux de réponse correctif" qui inclut les "hors champ" dans les répondants. Ces "hors champ" ne sont pas seulement les "hors champ" enregistrés comme tels en 1994 mais aussi les nouveaux "hors champ" apparus comme tels au cours de la nouvelle observation (les nouvelles personnes décédées, émigrées et en ménage collectif).

Ce taux de réponse est de 83.7% soit 8 477 enquêtes réalisées et 700 "hors champ" qui sont les hors champ de 1994 augmentés de nouveaux décès, de nouvelles émigrations et de nouvelles personnes passées en ménages collectifs.

Distribution de l'échantillon longitudinal en 1994 et en 1995 pondéré par le poids de sélection (wt94)

STATUT	***FREQ. 1994 pondérées***	***FREQ. 1995 (poids 1994)***	***Fréq. relative 1995***
Enquêtes réalisées	10 497	8 477	77.3%
"Hors Champ"	470	700	6.4%
"Echecs"	-	1 790	16.3%
TOTAL	***10 967***	***10 967***	***100.0%***

Les "refus" de 1994 (1ère vague) ont "disparu". Ils ont été absorbés par les poids des répondants de 1994. Seuls les "nouveaux refus" enregistrés en 1995 apparaissent. Ils glissent du numérateur vers le dénominateur de la formule de calcul du taux de réponse. Soit 1 790 individus.

La taille de l'échantillon longitudinal n'est donc pas modifiée. Les 10 967 individus "circulent" entre les trois groupes. Les nouveaux "hors champ" et les nouveaux "refus" ont emporté avec eux le poids qui leur a été attribué en 1994.

Désormais, à chaque vague, le taux de réponse correctif longitudinal se rapportera toujours à l'échantillon pondéré de la première vague (1994). Le nombre (pondéré) des refus et des "hors champ" sera donc cumulatif.

Des "refus" de 1995 pourront être récupérés par la suite; des émigrés "hors champ" pourront éventuellement revenir dans leur ménage d'origine mais, inexorablement, la proportion des "enquêtes réalisées" et des "hors champ" (les "répondants") aura tendance à diminuer et la proportion des échecs aura tendance à augmenter. Les poids des répondants auront donc tendance à augmenter pour compenser ces échecs.

En tout état de cause, la taille de l'échantillon longitudinal restera constante pendant toute la durée de vie de cet échantillon longitudinal.

3. LA NON-REPONSE EST-ELLE IGNORABLE ?

La non-réponse est-elle ignorable ou risque-t-elle d'introduire certains biais parce qu'elle est liée au moins partiellement à des variables mesurées par l'enquête ? Pour tenter d'expliquer cette non-réponse, il est fait appel à des variables qui couvrent différents volets de l'étude :

- Le niveau de formation le plus élevé atteint par chaque personne.
- Le nombre d'emplois dans le ménage.
- La diversité des revenus existants dans le ménage.
- Le montant de ces revenus.
- La taille du ménage.
- Le nombre d'adultes.
- Le nombre d'enfants.
- Le revenu disponible pour chaque unité de consommation.
- Le fait d'être propriétaire ou locataire de son logement.
- Le type d'habitation (individuel, individuel en série, appartement dans un immeuble plus ou moins grand).
- Des caractéristiques individuelles : âge, sexe, nationalité, ancienneté dans le pays, localisation géographique (canton).
- Des caractéristiques du chef de ménage : âge, sexe ; a-t-il des revenus propres ? a-t-il un emploi ?
- L'effet "enquêteur" : la distribution des adresses aux enquêteurs s'effectue selon des zones géographiques plus détaillées que les cantons.

Toutes ces variables sont transformées en "dummies" et entrées dans une analyse discriminante[1] "pas à pas".

Un "arbre" des taux de réponse est constitué à partir des variables les plus discriminantes (c'est-à-dire celles qui accroissent significativement le pourcentage de variance expliquée, aussi longtemps qu'elles ne conduisent pas à distinguer des sous-catégories de trop petite taille).

[1] Pour mémoire, la variable dépendante est une variable dichotomique. Les probabilités de réponse peuvent être modélisées à l'aide d'une régression logistique. L'analyse discriminante, bien qu'elle soit fondée sur un autre modèle, donne des résultats équivalents selon le point de vue qui nous occupe.

Les variables les plus discriminantes sont : la distribution géographique (reflétée au travers des groupes d'enquêteurs), le nombre d'enfants dans le ménage, certaines catégories de revenus (sans effet linéaire) et l'âge du chef de ménage.

La combinaison de ces variables ne permet d'expliquer que 6% de la non-réponse. Nous émettons l'hypothèse selon laquelle la part non-expliquée de la non-réponse est "négligeable" en ce sens qu'elle n'est pas liée aux variables importantes de l'étude.

4. LES POIDS LONGITUDINAUX AJUSTES

Les poids des individus longitudinaux ajustés en 1995 (**w*il* 95**) sont égaux

au produit

- des poids des individus longitudinaux en 1994 (**wt94**)
- et de l'inverse des taux de réponse en 1995 (**1 / R*lg* (1995)**)

soit :

$$\mathbf{w}il\ \mathbf{95} = \mathbf{wt\ (94)} * \frac{\mathbf{1}}{\mathbf{R}lg\ \mathbf{(95)}}$$ soit :

$$\frac{\mathbf{1}}{\mathbf{probabilité\ de\ sélection\ en\ 94}} * \frac{\mathbf{n\ unités\ en\ 94}}{\mathbf{n\ rép.\ longitud.\ en\ 1995}}$$

sachant que le rapport entre le nombre d'individus longitudinaux ayant répondu en 1995 et le nombre d'unités dans l'échantillon en 1994 peut varier selon les strates définies par les variables discriminantes.

Cet ajustement produit des estimations sans biais si **R*lg* (95)** est égal à la vraie probabilité de réponse des unités à la vague 1995.

Distribution de l'échantillon longitudinal pondéré en 1994 (wt94) et en 1995 (wil95)

STATUT	***FREQ. 1994 pondérées***	***FREQ. 1995 pondérées***
Enquêtes réalisées	10 497	10 177
"Hors Champ"	470	790
TOTAL	***10 967***	***10 967***

CHAPITRE VII

SAUVEGARDER LA REPRESENTATIVITE ANNUELLE DES ECHANTILLONS TRANSVERSAUX

7. SAUVEGARDER LA REPRESENTATIVITE ANNUELLE DES ECHANTILLONS TRANSVERSAUX

Chaque nouvelle vague d'enquêtes correspond non seulement à une nouvelle étape dans l'élaboration de l'échantillon longitudinal mais aussi à la création d'un nouvel échantillon transversal. Ce nouvel échantillon transversal pose deux problèmes.

D'une part, chaque échantillon transversal doit rester **représentatif de la population au moment de la nouvelle observation**. D'autre part, il doit prendre en compte l'ajustement des **poids des individus longitudinaux** en fonction du taux de réponse longitudinal.

1. CONCERNANT LA REPRESENTATIVITE

Le nouvel échantillon resterait représentatif de la population s'il reproduisait intégralement tous les phénomènes démographiques qui se sont produits dans la population. Or, il ne les reproduit que partiellement.

- Il génère des "naissances" : de nouveaux ménages se forment (par mariage), des nouveaux nés apparaissent dans les ménages et des cohabitants rejoignent les ménages formés par les individus longitudinaux.

- Il génère des "morts" : des ménages se défont (par divorce qui génèrent à leur tour de nouveaux ménages), des personnes décèdent, d'autres quittent le pays ou entrent dans des ménages collectifs.

- Il ne reproduit pratiquement pas les "naissances" par immigration.

Pour prendre en compte les naissances un nouvel échantillon supplémentaire doit être introduit chaque année (si possible). Toutefois, le panel produit par lui-même les nouveau-nés et seules les "naissances par immigration" doivent être absolument introduites volontairement.

Composition des échantillons transversaux en 1994 et en 1995

Statut dans l'échantillon	***1994***	***1995***
Indiv. longitud. : "échecs 1994"	2466	2466
Indiv. longitud. : "échecs 1995"		1462
Indiv. longitudinaux : "réalisés"	8232	6597
Nouveaux indiv. transversaux		132
"nouveau-nés"		56
"nouveaux immigrants"		22
Hors champ 1994	269	269
Hors champ 1995		174
Total	***10967***	***11178***

2. CONCERNANT LA REPONDERATION DES ECHANTILLONS TRANSVERSAUX

Pour la clarté nous désignons dorénavant par "**poids initial"**, le poids des individus longitudinaux (présents dans chaque fichier transversal).

Nous considérons des **ménages entiers** et non pas seulement les personnes. Prendre ainsi en compte tous les membres de chaque ménage est une procédure efficace lorsqu'on souhaite produire des estimations au niveau des ménages.

Par contre, cette approche pose le **problème de l'entrée incessante de nouveaux membres** dans les ménages. Ces nouveaux membres pouvaient être présents dans la population au moment de la sélection de l'échantillon longitudinal mais elles n'ont pas été sélectionnées. Dans ces conditions, comment faut-il les pondérer ?

2.1. Repondération de l'ensemble des personnes qui forment les ménages

Admettons dans un premier temps que la population de référence ne s'est pas modifiée. Il ne s'y est produit ni "naissances", ni "décès".

L'observation de ménages entiers montre malgré tout que les ménages peuvent s'être modifiés d'une année à l'autre : des nouvelles personnes sont venues s'ajouter aux membres du ménage observés l'année précédente.

Ces "membres ajoutés" étaient présents dans la population de référence (U) au moment de la sélection de l'échantillon (s) en 1994 mais ils n'ont pas été sélectionnés. Ils sont observés au cours d'une vague ultérieure (en 1995) parce qu'ils se retrouvent dans un ménage avec un membre (m) de l'échantillon (s), soit un membre longitudinal.

Comment obtenir un poids pour ces membres non-longitudinaux sans biaiser l'estimateur des poids ?

On peut montrer que la méthode dite du "partage des poids" permet d'obtenir un poids pour ces membres sans introduire aucun biais dans l'estimateur des poids à condition que ces membres "transversaux" reçoivent dans un premier temps un poids initial de "0".

Dans chaque échantillon transversal

- les individus longitudinaux ont donc un **poids initial** déterminé par la repondération annuelle des individus longitudinaux (**cf. 6.4**)

- et les individus non-longitudinaux ont un **poids initial** égal à "0".

Le poids du ménage est égal à la moyenne de ces poids, soit la somme des poids des membres divisée par le nombre de membres dans le ménage.

Chaque personne reçoit ensuite **un poids final** : la moyenne des poids des membres du ménage.

On observera le fait que les membres non-longitudinaux ont simplement "reçu" un poids qui n'est qu'une partie des poids des membres longitudinaux. Cette procédure permet de prendre en compte toute la richesse de l'information apportée par ces membres non-longitudinaux. Elle contribue à maintenir la représentativité de l'échantillon par rapport à la population de référence dans la mesure où **les ménages ont évolué.**

Mais cette population n'est pas constante. D'autres phénomènes doivent être pris en compte.

2.2. Pondération des naissances : nouveau-nés[1]

La population n'est pas constante. Nous supposons, dans un second temps, qu'elle n'évolue qu'en fonction des naissances d'enfants et des décès.

Il n'est pas nécessaire de sélectionner chaque année un échantillon de nouveau-nés puisqu'ils se retrouvent inévitablement dans les ménages contenant les personnes de l'échantillon initial.

La méthode du "partage des poids" s'applique ici aussi. Toutefois, à la différence des membres non-longitudinaux qui pouvaient être présents dans la population au moment de la sélection de l'échantillon initial (s), ces nouveaux membres non-longitudinaux doivent être pris en compte en tant qu'individus supplémentaires dans l'échantillon puisque ce sont aussi des individus supplémentaires dans la population.

Ils reçoivent aussi un **poids initial** égal à "0".

Le **poids du ménage** est égal à la somme des poids des individus divisée par le nombre de membres dans le ménage excluant les nouveau-nés.

Le poids final des nouveau-nés est donc égal à la moyenne des poids des autres membres du ménage.

On notera que ces nouveau-nés sont des individus transversaux. Ils ne sont pas nécessairement inclus dans l'échantillon de manière permanente. On pourrait aussi bien sélectionner tous les ans un nouvel échantillon de "nouveau-nés" entrés dans la population depuis la sélection de l'échantillon initial. Cette procédure pose certains problèmes que nous ne développerons pas ici. En outre, elle implique nécessairement que les nouveau-nés apparus spontanément dans les ménages du panel soient pondérés différemment.

[1] Concernant ce point et le suivant voir P. Lavallée: « Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode des poids partagés », Techniques d'enquête, Vol. 21, n°1, pp. 27-35, Statistique Canada, juin 1995,.

2.3. Pondération des naissances : les nouveaux immigrants

La population n'est pas constante. Nous avons supposé jusqu'ici qu'elle n'évolue qu'en fonction des naissances d'enfants et des décès. Dans un troisième temps, il nous faut encore admettre que la population évolue en fonction de la "naissance d'immigrants".

Il n'est pas nécessaire de sélectionner chaque année un échantillon de nouveau-nés, puisqu'ils se retrouvent dans les ménages du panel mais rien n'interdit de le faire.

De la même manière, des nouveaux immigrants apparaissent spontanément dans le panel. Ils y forment un sous-groupe de l'ensemble des nouveaux venus qui rejoignent les ménages à cette différence près qu'ils n'étaient pas présents dans la population au moment du tirage du premier échantillon.

Cette génération de nouveaux immigrants dans les ménages du panel n'est évidemment qu'un phénomène relativement marginal. La plupart des nouveaux immigrants forment de nouveaux ménages et le panel ne peut pas rendre compte spontanément de cette modification de la structure de la population.

Il est donc indispensable de sélectionner un nouvel échantillon représentatif des immigrants entrés dans le pays depuis la sélection du premier échantillon. Idéalement, cette procédure devrait être effectuée chaque année. En pratique, elle ne sera pratiquée que tous les deux ans[1].

1. **En 1995**, seuls les nouveaux immigrants entrés "par la porte" des ménages du panel seront pris en compte. Ils sont pondérés de la même manière que les nouveau-nés.

   Ils reçoivent un **poids initial** égal à "0".

   Le **poids du ménage** est égal à la somme des poids des individus divisée par le nombre de membres dans le ménage excluant les nouveaux immigrants.

   Le poids final des nouveaux immigrants est donc égal à la moyenne des poids des autres membres du ménage.

2. **En 1996**, un échantillon de nouveaux immigrants entrés au pays depuis 1994 sera sélectionné sur la base du fichier de la Sécurité Sociale. Ces "nouveaux immigrants" sont des "titulaires principaux de revenus" qui conduisent à des ménages. La probabilité de sélection de ces ménages est inégale. L'équiprobabilité doit être rétablie.

   Dans l'échantillon initial, le poids du ménage était attribué à l'ensemble de ses membres. Dans les échantillons longitudinaux suivants, cette procédure doit prendre en compte le fait que les unités de sélection de cet échantillon supplémentaire sont limitées aux nouveaux *immigrants*. En sélectionnant ces nouveaux immigrants par l'intermédiaire des ménages, on risque d'inclure également d'autres individus : individus déjà présents dans la population au moment de la sélection du premier échantillon, nouveau-nés, individus longitudinaux.

[1] Afin de ne pas surcharger les personnes du service de l'I.G.S.S. qui nous prêtent très aimablement leur concours.

Ces autres individus sont sélectionnés de manière irrégulière. Leur probabilité de sélection est mal définie. Ils feront donc l'objet d'une procédure de pondération particulière que nous ne développerons pas ici.

En outre, en l'année où un échantillon supplémentaire de nouveaux immigrants est inséré dans l'échantillon transversal du panel, les immigrants entrés dans l'échantillon depuis 1994 par le truchement des ménages du panel voient leur poids modifié. Ce ne sont plus des individus "supplémentaires", sans quoi la probabilité de sélection des nouveaux immigrants serait faussée ce qui introduirait inexorablement un biais.

A noter ici encore le fait que cet échantillon supplémentaire est purement transversal sans existence permanente dans l'échantillon.

3. PROFILS DES ECHANTILLONS TRANSVERSAUX PONDERES

Composition des échantillons des individus transversaux en 1994 et en 1995

STATUT	***FREQ. 94 pondérées (wt94)***	***FREQ. 95 pondérées (wtf95)***
1. Indiv. longitudinaux "réalisés"	10 497	9 844
2. Nouveaux individ. transversaux		70
3. "nouveau-nés"		23
4. "nouveaux immigrants"		10
Total 1	10 497	9 947
5. (hors champ)	(470)	(790)
Total 2	***10 967***	***10 737***

Remarques

1. Les hors champ (groupe #5) sont considérés comme des répondants pour lesquels les variables mesurées prennent la valeur "0".

2. "9 947" est en soi une valeur "résultante". Elle résulte de quatre composantes :

- le nombre d'unités longitudinales ayant répondu, pondérées en vue de compenser la non-réponse de certains individus longitudinaux (groupe #1)
- parmi lesquelles certaines ont partagé leur poids avec des "nouveaux individus transversaux cohabitants" (groupe #2).
- Ce nombre est augmenté des nouveau-nés, unités supplémentaires ayant reçu un poids égal à celui des autres membres du ménage (groupe #3)

- et du nombre des nouveaux immigrants, unités supplémentaires ayant reçu un poids égal à celui des autres membres du ménage (en l'absence d'une sélection d'un nouvel échantillon de nouveaux immigrants) (groupe #4).

Composition des échantillons des ménages transversaux pondérés en 1994 et en 1995

STATUT	*FREQ. 94 (mwt94)*	*FREQ. 95 (mwt95)*
1. Ménages "observés"	5 120	3 886
2. Hors champ	593	741
Total après pondération	***5 713***	***4 627***

Les 2 472 ménages observés plus les 77 nouveaux ménages formés en 1995 "pèsent" **3 886 unités transversales pondérées** soit

- la somme sur l'ensemble des ménages
- de la somme dans chaque ménage des poids des membres divisée par la taille du ménage.

Pour mémoire :

wt94 est le poids des individus en 1994

mwt94 est le poids des ménages en 1994

wil95 est le poids longitudinal des individus en 1995

wtf95 est le poids transversal des individus en 1995

mwt95 est le poids des ménages en 1995.