N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

METHODE DE PONDERATION DE L'ENQUETE FRANÇAISE « CONSTRUCTION DES IDENTITES »

Pascal Ardilly[1]

RÉSUME

L'enquête « Construction des identités » porte sur 13 500 individus, issus de sources multiples, et sur représente les populations de personnes handicapées, immigrées ou enfants d'immigrés à la date du recensement. Lorsque l'individu tiré n'est pas retrouvé, on échantillonne au hasard une personne dans le ménage visité par l'enquêteur à la date de collecte. Le système de pondération pose de sérieux problèmes du fait du recouvrement des bases de sondage et de cette procédure de tirage au hasard de certains individus. Il y a en particulier des difficultés liées à la non-disponibilité de certaines informations intervenant dans la détermination de poids « sans biais ».

MOTS CLÉS : Bases multiples, enquête en deux phases, information manquante, pondération, sur représentation.

1. OBJECTIF DE L'ENQUETE

L'enquête « Construction des identités », menée en France en mars-avril 2003, a comme objectif de décrire et d'analyser les différents types de liens sociaux qui permettent aux individus de s'intégrer dans la société française du début du XXIème siècle. Elle porte sur 13 500 individus âgés de 18 ans et plus, issus de sources multiples, et sur représente trois populations plus particulièrement susceptibles de vivre des difficultés d'insertion : les personnes immigrées, les personnes nées de parents immigrés et les personnes handicapées de moins de 60 ans. Néanmoins, la question des identités, entendue au sens large comme la façon qu'a un individu de se construire une place dans la société qui permette à la fois son intégration et l'affirmation de son individualité propre, se pose aussi pour les personnes qui n'ont pas d'origine hors de France et ne sont touchées par aucun handicap physique ni mental. C'est pourquoi plus de la moitié de l'échantillon concerne la population qui ne possède a priori aucune des trois caractéristiques précédentes.

2. CONSTITUTION DE L'ECHANTILLON

2.1 Tirages initiaux dans les bases de sondage

Une première fraction de l'échantillon provient d'une enquête dite « Vie quotidienne et santé » (VQS) de très grosse taille (400 000 individus environ), complémentaire au recensement de 1999 et qui a permis de cibler les personnes handicapées de moins de 60 ans. Cette enquête est stratifiée, aréolaire et touche l'intégralité de certains secteurs d'agent recenseur préalablement échantillonnés par sondage aléatoire simple au sein de zones de délégué[2] elles-mêmes issues d'un sondage. Ces zones de délégué sont tirées proportionnellement à leur taille en nombre de personnes physiques.

Dans « Identités », l'échantillon de personnes handicapées constitue une seconde phase de tirage au sein de l'échantillon VQS : partant de zones de délégué de VQS que l'on aura préalablement sous-échantillonnées[3], on effectue un tirage systématique de 2040 personnes handicapées dans l'ensemble de ces zones, puis on examine la taille de l'échantillon recoupant chacune des zones de délégué. S'il y a 6 personnes ou moins dans une zone donnée,

[1] Pascal Ardilly, INSEE, 18 Bd Adolphe Pinard 75675 Paris Cedex 14, France (pascal.ardilly@insee.fr).
[2] Un délégué au recensement a une zone de compétence regroupant de l'ordre de 40 agents recenseurs.
[3] Pour 2 raisons : d'une part on veut limiter la taille de l'échantillon de zones dans « Identités » et d'autre part VQS sur-représente certaines régions et on souhaite rétablir l'équilibre des taux de sondage pour « Identités ».

on les enquête toutes, sinon on effectue de nouveau un tirage aléatoire de personnes afin de ne pas dépasser ce seuil de 6 enquêtes par zone. In fine, on obtient 1215 personnes formant l'échantillon S1. A première vue, les probabilités de sélection finales des individus devraient être voisines, car d'une part les zones de délégué ont été initialement tirées proportionnellement à leur taille, et d'autre part les tirages au second degré ont été conçus de manière à assurer des allocations par zones constantes, ou inférieures au seuil de six en cas de réserve insuffisante.

Une seconde fraction d'échantillon concerne deux populations repérées lors de l'enquête « Etude de l'histoire familiale » (EHF), elle aussi de très grande taille et complémentaire au recensement de 1999, à savoir les immigrés et les enfants d'immigrés. Comme VQS, l'enquête EHF est stratifiée, aréolaire et touche l'intégralité de certains secteurs d'agents recenseurs situés dans les mêmes zones de délégué que VQS, mais néanmoins différents des secteurs d'agents recenseurs utilisés par VQS. Le plan de sondage de EHF est par ailleurs conçu de manière à sur représenter les femmes.

L'échantillonnage « Identité » des immigrés et des enfants d'immigrés s'interprète comme une seconde phase de tirage dans EHF et s'effectue au sein des zones de délégué tirées pour constituer l'échantillon S1 (afin de limiter le nombre de zones de délégué sollicitées, et par là même le nombre d'enquêteurs). On distingue alors les zones de délégué géographiquement étendues des petites zones[4]. Les traitements ultérieurs s'organisent en 2 blocs.

D'une part, on considère l'ensemble constitué des petites zones et des grandes zones comprenant 3 communes ou moins, et on s'intéresse à la population des immigrés qui y résident. On effectue alors un tirage systématique dans cette population. Le principe rejoint ensuite celui de l'échantillonnage des handicapés : on constate l'effectif échantillonné par zone de délégué, et s'il est inférieur à un certain seuil, on retient tout le monde, sinon on pratique un nouvel échantillonnage de personnes, systématique et de taille convenue (égale à 10 personnes).

D'autre part, on considère les grandes zones comprenant plus de 3 communes. On commence par tirer 3 communes dans chaque zone, proportionnellement à leur taille, puis on échantillonne commune par commune un nombre donné d'individus (5 personnes) - sauf si l'effectif disponible est insuffisant (ce qui est assez courant avec les populations « rares » ici concernées), auquel cas on enquête toutes les personnes disponibles.

On effectue en dernier lieu un échantillonnage aléatoire simple de femmes pour « gommer » la sur représentation introduite par la première phase EHF.

In fine, on obtient un échantillon de 2025 immigrés, noté S2. La même procédure complète est appliquée à la population des enfants d'immigrés, conduisant à un échantillon S3 de 2025 personnes également. Pour S2 comme pour S3, avec le même argument que pour S1, on devrait avoir une dispersion des poids relativement limitée, parce que si on fait abstraction des cas où la réserve tirable est inférieure au seuil, les tirages impliqués sont à plusieurs degrés et proportionnels à la taille à chaque degré, sauf au dernier où l'effectif tiré est constant.

Un quatrième échantillon, noté S4, comprenant 7830 individus, est tiré directement dans les fichiers du recensement 1999, indépendamment des échantillonnages précédents si ce n'est qu'il se situe intégralement dans les zones de délégué préalablement échantillonnées pour S1, S2 et S3 mais qu'il évite les secteurs d'agent recenseur dans lesquels se concentrent les enquêtes VQS et EHF (ces secteurs sont déterminés par un tirage à probabilités égales dans la zone de délégué). On distingue de nouveau petites et grandes zones, on échantillonne des communes dans les grandes zones si ces grandes zones comprennent 4 communes ou plus, et on tire in fine des individus par sondage aléatoire simple dans les zones ou dans les communes retenues, selon le cas. Les allocations sont convenues par avance, soit 21 individus dans les petites zones et 8 individus dans chacune des 3 communes retenues pour les grandes zones. S'il y a moins de 3 communes dans une grande zone, on essaie de se rapprocher d'un effectif de 21 personnes tirées dans la zone, afin de limiter la dispersion des poids.

Enfin, on a échantillonné, dans une base spécifique tout à fait disjointe des précédentes, 405 logements neufs dans les zones de délégués retenues, là encore selon un plan complexe, pour prendre en compte la construction achevée après 1999. On a obtenu ainsi un échantillon S5.

[4] On dispose, dans la base de sondage des zones, d'un indicateur de surface totale.

L'échantillon global est donc la réunion de cinq échantillons (S1 à S5) issus de quatre sources différentes, et comprenant 13 500 individus. Il traduit une sur représentation des handicapés de moins de 60 ans, des immigrés et des enfants d'immigrés. Chacun des échantillonnages est stratifié, à plusieurs degrés, et constitue un plan de sondage complexe où l'on a tenté de limiter la dispersion des poids des individus. En fin d'opération, on effectue un repérage puis une élimination de doublons, c'est-à-dire que l'on garde un seul individu dans les ménages qui ont été sollicités au moins deux fois lors des tirages S1 à S4. Ces doublons ne sont pas anecdotiques car les populations concernées par S1, S2 et S3 présentent des effets de grappe assez naturels.

2.2 Sélection des individus finalement interrogés

L'individu tiré étant identifié avec ses caractéristiques à la date du recensement, en particulier l'adresse du logement dans lequel il résidait au moment du recensement, il y a évidemment de nombreux cas où l'on ne retrouve pas la personne tirée, la plupart du temps parce qu'elle a déménagé depuis mars 1999, date du recensement. Comme il n'y a pas de suivi des individus jusqu'à leur nouvelle adresse, la technique consiste dans ce cas de figure à échantillonner au hasard une personne parmi celles qui composent le ménage visité par l'enquêteur à la date de collecte.

3 . DIFFICULTES RENCONTREES CONCERNANT LA PONDERATION

Le plan de sondage adopté crée plusieurs obstacles à un calcul simple de pondérations, susceptibles d'entraîner une certaine perte d'efficacité :

- concernant la sélection proprement dite :

1) Par principe, il convient de souligner la complexité de l'échantillonnage, même si on s'en tient à une population donnée. Pour S1, S2 et S3, le processus est celui d'une enquête en 2 phases où chaque phase résulte d'un plan à probabilités inégales et à plusieurs degrés. Les allocations finalement retenues au sein des différentes unités géographiques - zones de délégués et communes - ne sont pas seulement liées à l'unité en question mais dépendent de l'ensemble du processus d'échantillonnage : par exemple, l'utilisation d'un tirage systématique de taille donnée sur un ensemble de zones de délégué (de tailles variables) rend l'allocation dans chacune d'entre elles totalement dépendante de la liste intégrale des zones échantillonnées. Cet élément de complexité paraît incontournable à partir du moment où les effectifs par échantillon sont fixés et que l'on doit nécessairement s'insérer dans un cadre de tirage en deux phases dont l'échantillon de première phase est de taille variable.

2) La détermination pratique des échantillons s'écarte parfois de la théorie pour tenir compte de nécessités de terrain qui n'ont pas toujours laissé de traces ou du fait de problèmes informatiques d'origine probablement perverse, concomitamment à une insuffisance de la documentation sur la constitution des bases de sondage. En particulier l'examen des échantillons finaux des enquêtes VQS et EHF met en évidence des « curiosités » diverses qui rendent fragiles les calculs de pondération, parce que par exemple la population affichée de certaines zones de délégué apparaît plus éloignée de la taille standard (15 000 habitants) qu'on ne pourrait le penser a priori, ou encore parce que le nombre même de ces zones de délégué découpées sur un territoire donné n'est pas connu de façon certaine. Autre exemple : on a eu à faire à quelques zones de délégué qui donnaient lieu après certaines éliminations d'usage (préalablement à chaque tirage, on enlève les échantillons des logements sollicités dans les enquêtes antérieures) à un effectif tirable sensiblement inférieur à l'effectif attendu.

3) La dispersion des poids bruts, échantillon par échantillon, est finalement apparue plus importante que prévue, puisqu'elle variait dans un rapport de 1 à 10 si on excepte quelques poids extrêmes. Cela est à relier en partie aux problèmes signalés au point précédent, en particulier au fait que les tailles de population des zones de délégué utilisées pour le calcul des poids étaient sensiblement plus dispersées que prévu : le sous échantillonnage initial des zones au début de la deuxième phase s'appuyait sur l'idée que les zones de délégué devaient être calibrées afin de se rapprocher d'une taille moyenne idéale, mais cela s'est avéré faux

à l'usage et une inégalité de pondération a trouvé sa source dans cette erreur d'appréciation. Une autre explication est liée à la dispersion des allocations par zone ou par commune : la pondération uniforme ne s'obtient en effet que si l'allocation par zone ou par commune est constante, ce qui n'était pas assuré loin de là s'agissant de populations relativement peu nombreuses. L'effet de seuil n'avait donc pas non plus été apprécié à son juste impact. Enfin, il faut rappeler que le tirage des zones de délégués s'effectue en proportion de la taille totale en nombre d'habitants mais que les tirages aux degrés ultérieurs ont, eux, utilisé les tailles des populations d'intérêt (handicapés / immigrés / enfants d'immigrés) : or le poids relatif de ces populations varie sensiblement d'une commune à l'autre.

Le tableau qui suit fournit des statistiques sur les poids bruts obtenus à l'issue des tirages initiaux :

	Minimum	Décile 1	Moyenne	Décile 9	Maximum
Handicapé (S1)	253	560	1 403	2 523	6 886
Immigré (S2)	256	975	2 640	5 279	11 976
Enfant d'immigré (S3)	252	952	2 332	3 933	19 270
Autre (S4)	927	3 371	5 628	7 993	21 451

- concernant les opérations postérieures au tirage :

1) Un individu handicapé, immigré ou enfant d'immigré peut être tiré via plusieurs bases de sondage : en effet, il peut provenir de l'un des échantillons S1, S2 ou S3 mais aussi de S4. S'il a plusieurs caractéristiques conjointes, par exemple être à la fois immigré et enfant d'immigré, il peut être joint au travers de 3 bases, voire même de 4 (puisqu'on peut être à la fois immigré, enfant d'immigré et handicapé de moins de 60 ans).

2) Considérons un individu handicapé, ou immigré ou enfant d'immigré. S'il est tiré au travers de l'un des échantillons S1, S2 ou S3, il est possible de calculer une probabilité de sélection de nature conditionnelle, comme on sait le faire dans les tirages en plusieurs phases parce qu'on maîtrise la succession des étapes qui ont permis de le sélectionner. Cette probabilité de sélection n'est pas « la » probabilité de sélection utilisée dans l'estimateur de Horvitz-Thompson, mais elle permet néanmoins d'obtenir un estimateur sans biais. Par ailleurs, on peut aussi calculer sa probabilité de sélection classique *a priori* dans S4, parce qu'il s'agit cette fois d'un plan à plusieurs degrés somme toutes assez simple et en une seule phase. En revanche, si l'individu que nous considérons provient de S4, il est totalement illusoire d'espérer obtenir « quelque chose » de la nature d'une probabilité de sélection relative aux plans menant à S1, S2 ou S3, compte tenu de la complexité de ces plans et surtout de la nécessité, nous semble-t-il, de passer par une approche conditionnelle. Il y a donc une très sérieuse difficulté pour calculer une pondération brute pour les personnes étant soit handicapées, soit immigrées, soit enfants d'immigré si elles proviennent d'un échantillonnage dans S4.

3) Le statut d'un individu - qui peut prendre 4 modalités : handicapé, immigré, enfant d'immigré ou « autre » - n'est pas connu au moment du tirage, mais seulement après la phase de collecte. Il faut donc reconstituer la situation de mars 1999 en interrogeant de manière adéquate les individus (des questions ont été prévues à cet effet), ce qui entraîne nécessairement des erreurs d'observation. En effet, on imagine aisément qu'il peut y avoir un flou autour de la notion de handicap (c'est moins vrai pour les statuts d'immigré et d'enfant d'immigré, bien que des incohérences portant sur des réponses données à 4 ans d'intervalle puissent exister aussi sur des points pourtant factuels). Le calcul des poids, qui dépend fondamentalement du statut, ne peut donc être effectué que tardivement.

4) Le processus de sélection au hasard d'un individu dans le ménage lorsque l'individu initialement désigné a quitté le logement brouille fortement les cartes : en effet, la pondération de l'individu finalement interrogé - qu'il ait été recensé ou non - doit dépendre d'une certaine façon des probabilités de sélection des personnes qui ont quitté le ménage depuis le recensement. Malheureusement, pour ces dernières il est vraiment impossible d'obtenir toute information sur leur statut en 1999.

De plus, si l'individu finalement interrogé était présent au moment du recensement mais qu'il a été désigné comme résultante d'un tirage au sort au sein du ménage, il a aussi une probabilité de sélection initiale qui n'est pas raisonnablement calculable (en effet, cet individu aurait pu être tiré au moins au travers de S4, peut-être même au travers de S1, S2 ou S3). Quant au cas des logements éclatés (il y en a une quarantaine), il est à peu près inextricable.

Ce type de difficulté est en soi un des plus pénalisants dans cette opération, il « détruit » d'une certaine façon tous les efforts faits aux étapes antérieures pour attribuer un poids initial à peu près conforme à l'échantillonnage adopté.

5) La sélection des logements neufs s'appuie sur une base de sondage qui permet de localiser les logements au niveau communal seulement. Or les tirages ont dû être concentrés dans certaines zones de délégué, comme expliqué au 1). Or de nombreuses zones de délégué sont infra communales. Ce décalage d'information n'a pu conduire qu'à des calculs très simplificateurs, ne reflétant pas la réalité du plan de sondage.

4. PROCESSUS MIS EN ŒUVRE POUR OBTENIR UNE PONDERATION

4.1 Au niveau du tirage des individus dans leurs bases respectives

Nous n'avons donc pas cherché à calculer une probabilité de sélection qui tienne compte explicitement des différents plans de sondage associés aux différentes bases par l'intermédiaire desquelles on pourrait sélectionner un individu. Nous avons cependant calculé un poids brut associé à chaque individu tiré au sein de chacun des échantillons S1 à S4, cette étape préliminaire de base apparaissant incontournable ! Puis nous avons utilisé le principe simple suivant : si Sa et Sb sont 2 échantillons indépendants tirés dans une même population, avec les jeux de probabilité de sélection respectives $\Pi_k^{(a)}$ et $\Pi_k^{(b)}$ pour chaque individu k, un estimateur naturel sans biais du total est :

$$\theta \cdot \sum_{k \in S_a} \frac{Y_k}{\Pi_k^{(a)}} + (1-\theta) \sum_{k \in S_b} \frac{Y_k}{\Pi_k^{(b)}} \qquad (1)$$

où θ est un réel au choix entre 0 et 1. La valeur optimale de θ est liée à la variance de chacun des estimateurs respectifs obtenus sur S_a et sur S_b. Le rapport de la taille de l'échantillon S_a sur la taille totale de l'échantillon $S_a \cup S_b$ constitue un choix raisonnable. Cette approche s'étend facilement au cas de 3 échantillons et plus.

Ainsi, pour chacun des 3 statuts handicapé « exclusif » / immigré « exclusif » / enfant d'immigré « exclusif », et pour chaque croisement possible entre ces statuts (ce qui fait au total 7 configurations), nous avons obtenu des poids en enchaînant les étapes suivantes :

- repérage des échantillons susceptibles de contenir les individus ayant le statut donné ;
- calcul d'un coefficient θ (ou de plusieurs coefficients) proportionnel(s) à la taille de l'échantillon associé à θ. Lorsque S4 est impliqué, c'est-à-dire toujours, la taille prise en compte est bien celle de la fraction de S4 qui possède le statut en question ;
- Fixation des probabilités de sélection en partant des probabilités calculées à partir du plan de sondage initial et en les divisant par le coefficient θ, pour la fraction de l'échantillon concernée par le statut

On n'obtient pas vraiment des probabilités de tirage, mais des pseudo probabilités dont l'inverse représente un poids acceptable. Cette opération part du principe que l'on peut toujours obtenir l'information sur le statut 1999, puisqu'il faut être capable d'attribuer un statut à tout individu tiré. Comme signalé précédemment, les erreurs d'observation constituent une source de biais certainement non négligeable.

Prenons un exemple : on repère les personnes handicapées « exclusives » en 1999 (selon leurs déclarations…) : certaines proviennent de S1, d'autres de S4, mais elles ne peuvent pas provenir de S2 ni de S3 (sinon elles ne seraient pas exclusivement handicapées). Celles qui proviennent de S1 ont un poids w_k calculé à partir du plan de sondage de S1, qui concerne mettons 2 000 individus. Celles qui proviennent de S4 ont un poids w_k calculé à partir du plan de sondage de S4 et totalisent par exemple 1 000 individus. On donnera simplement à tout individu de S1 le poids w_k * 2/3 et à chacun des 1000 individus concernés dans S4 le poids w_k /3.

4.2 Au niveau de la sélection de l'individu enquêté dans le logement

Ensuite, il a fallu passer de ces pseudo probabilités de sélection d'individus initialement tirés à des poids associés à chaque individu finalement interrogé. Pour cela, on considère un ménage donné. On appelle A l'ensemble des individus présents à la fois au recensement 1999 et à la date de l'enquête. On appelle B l'ensemble des individus présents au moment du recensement 1999 mais qui ont quitté définitivement le ménage à la date de l'enquête. Enfin, on appelle C l'ensemble des individus du ménage qui n'ont pas été recensés en 1999 dans le logement désigné. Au sein du ménage **m**, on note Π_k^I la pseudo probabilité de sélection de l'individu **k** (obtenue selon 4.1) et Π_k^F la probabilité de sélection finale de l'individu **k**, telle qu'elle résulte du processus terrain.

On néglige la probabilité que deux individus distincts soient tirés dans le même ménage[5]. Dans ce cas, on a :

- Pour tout **k** dans A : $$\Pi_k^F = \Pi_k^I + \frac{1}{N_m} \cdot \sum_{j \in B} \Pi_j^I$$
- Pour tout **k** dans C : $$\Pi_k^F = \frac{1}{N_m} \cdot \sum_{j \in B} \Pi_j^I$$

où N_m est la taille du ménage **m** à la date de l'enquête (nombre d'individus du champ). On notera en particulier que si B est vide mais que C ne l'est pas, il y a existence de biais parce que les individus nouveaux ne sont pas couverts par la méthodologie adoptée.

Comme nous l'avons déjà déploré, on ne sait absolument rien sur les Π_k^I des individus de B, et on a choisi d'imputer une valeur égale à la probabilité de sélection moyenne $\overline{\Pi}$. Cette valeur moyenne - numériquement voisine de 4500 - est obtenue selon

$$\sum_c \frac{\hat{N}_c}{\hat{N}} \overline{\Pi}_c = \overline{\Pi} \qquad (2)$$

où **c** décrit les quatre statuts possibles, $\overline{\Pi}_c$ désigne la moyenne arithmétique des Π_k^I de tous les individus tirés dont on sait qu'ils ont le statut **c** (il s'agit donc nécessairement d'individus du groupe A initialement désignés par le sort), et $\hat{N}_c$ estime la taille de la population de statut **c** (obtenue en utilisant les probabilités Π_k^I). Les calculs sont effectués France entière. Seuls les individus tirés dans des ménages qui n'ont perdu aucun individu depuis le recensement échappent à cette imputation. Cette approche requiert alors la connaissance du cardinal de B, ce qui est acquis car une question a été spécifiquement introduite pour savoir si chaque individu composant le ménage à la date de l'enquête était ou non présent dans ce même logement au moment de recensement (comme on connaît par ailleurs la taille du ménage au moment du recensement, on en déduit la taille de B par différence - néanmoins il faut reconnaître que l'on trouve à ce niveau des incohérences, par exemple des tailles de ménages recensés en mars 1999 inférieures au cardinal de A).

[5] Elle est faible -mais non nulle. Techniquement, on ne sait d'ailleurs pas la calculer de manière exacte, et l'approximation est très complexe.

Le problème de la non-connaissance des Π_j^I des individus **j** de A désignés par un tirage au sort dans le ménage, est traité de la manière suivante - étant bien entendu qu'il faut croire à la pertinence de ces imputations, qui restent sujettes à discussion et ne prétendent que donner des ordres de grandeur rationnels :

On note **k** l'individu (appartenant nécessairement à B) initialement désigné :

- **j** n'est ni handicapé, ni immigré ni enfant d'immigré : il ne peut provenir que de S4 et on recalcule une probabilité de tirage à partir du plan de sondage de S4 que l'on affecte comme valeur de Π_j^I.

- **j** est soit handicapé, soit immigré, soit enfant d'immigré, et il se trouve que **k** a le même statut 1999 que **j**, et que **k** provient de S1 ou de S2 ou de S3. On prend Π_k^I comme valeur de Π_j^I (cela ne traite pas bien les cas d'intersection de statut - par exemple les handicapés et immigrés - parce qu'on n'a qu'une information incomplète sur **k** mais ces cas doivent être rares et pour éviter trop de complications il n'en a pas été tenu compte).

- **j** est soit handicapé, soit immigré, soit enfant d'immigré, soit cumule deux ou trois de ces statuts, et il se trouve que **k** provient de S4. On prend $\overline{\Pi}_c$ comme valeur de Π_j^I où **c** désigne le statut de **j**.

- autres cas : on impute $\overline{\Pi}$.

4.3 Au niveau des logements neufs

Un mot sur les logements neufs, afin de signaler que le problème évoqué au 3. a été brutalement résolu : une estimation de construction neuve a été faite au niveau France entière, et on a affecté d'autorité un poids égal à l'inverse du taux de sondage, soit uniformément égal à 3162.

4.4 Redressements et correction de non réponse

Environ 8 500 individus ont répondu à l'enquête. On part du principe qu'un tirage au sort d'un individu dans un ménage n'est jamais déclenché par la non réponse d'un individu initialement désigné. A partir des poids bruts obtenus selon le processus précédent, un calage utilisant le logiciel CALMAR a été réalisé sur des données externes susceptibles à la fois de corriger le biais de non-réponse totale et de réduire la variance d'échantillonnage.

Tableau 1 : Comparaison entre les marges tirées de l'échantillon (avec la pondération initiale) et les marges dans la population (marges du calage)

	Marge échantillon	Marge population	% échantillon	% population
Né à l'étranger				
Oui	5251,84	5629,50	11,66	12,50
Non	39784,16	39406,50	88,34	87,50
Nombre de personnes dans le ménage				
1	8039,00	7701,16	17,85	17,10
2	16149,24	15447,35	35,86	34,30
3	7826,20	8827,06	17,38	19,60
4	7933,87	8106,48	17,62	18,00
5 et +	5087,69	4953,96	11,30	11,00
Occupation professionnelle de la personne interrogée				
Actif occupé salarié	21575,33	21076,85	47,91	46,80
Actif occupé indépendant	2328,94	2522,02	5,17	5,60
Chômeur	2768,50	2296,84	6,15	5,10
Elève, étudiant	2550,87	3107,48	5,66	6,90
Retraité	9378,48	8151,52	20,82	18,10
Retiré des affaires	1682,44	2386,91	3,74	5,30
Inactif	4751,44	5494,39	10,55	12,20
Classe d'âge par sexe				
Homme de moins de 29 ans	3329,21	4503,60	7,39	10,00
Homme de 29 à 38 ans	3830,01	4233,38	8,50	9,40
Homme de 39 à 48 ans	4160,26	4098,28	9,24	9,10
Homme de 49 à 58 ans	3671,00	3692,95	8,15	8,20
Homme de 59 à 68 ans	2484,00	2431,94	5,52	5,40
Homme de 69 et plus	2993,59	2657,12	6,65	5,90
Femme de moins de 29 ans	4085,82	4458,56	9,07	9,90
Femme de 29 à 38 ans	4427,67	4278,42	9,83	9,50
Femme de 39 à 48 ans	4811,55	4233,38	10,68	9,40
Femme de 49 à 58 ans	4197,02	3783,02	9,32	8,40
Femme de 59 à 68 ans	3061,22	2702,16	6,80	6,00
Femme de 69 ans et plus	3984,64	3963,17	8,85	8,80
Type d'habitat et de commune				
Commune rurale	11724,21	12114,68	26,03	26,90
Individuel, com. de moins de 50 000 hab	8465,82	6980,58	18,80	15,50
Collectif, com. de moins de 50 000 hab	2604,18	3332,66	5,78	7,40
Individuel, com. de 50 à - de 200 000 hab	3796,75	2882,30	8,43	6,40
Collectif, com. de 50 à - de 200 000 hab	2741,47	2792,23	6,09	6,20
Individuel, com. de 200 000 hab et +	5565,80	3873,10	12,36	8,60
Collectif, com. de 200 000 hab et +	5441,11	5629,50	12,08	12,50
Individuel, en agglomération parisienne	1590,92	1891,51	3,53	4,20
Collectif, en agglomération parisienne	1808,41	3873,10	4,02	8,60
Paris « intra muros »	1297,31	1666,33	2,88	3,70

5. CONCLUSION

Ce qui est intéressant - et utile - au travers de cette expérience est tout autant dans la nature des difficultés rencontrées que dans les solutions techniques apportées.

1) Il y a une tendance générale, qui n'est pas nouvelle, à considérer que les questions techniques autour des sondages ne sont pas fondamentales dans le processus d'information, en tout cas elles ne sont pas de nature à impacter les décisions majeures fixant les principales caractéristiques de l'enquête. Bien sûr, il serait difficile à faire admettre à la société civile que la technique représente un obstacle capable de la priver d'une connaissance de certains aspects de son propre fonctionnement, mais dans toute décision le risque d'une dégradation de la qualité doit aussi être bien pesé en contrepartie. Dans le même ordre d'idée, l'argument de simplicité nous semble être un argument fort : cela parait une évidence, mais finalement la quasi-totalité des difficultés rencontrées pour attribuer une pondération acceptable ne sont que la conséquence d'un décalage entre les contraintes de l'opération et la capacité au sens large de l'équipe chargée de la mener.

2) La pondération a été conçue afin de privilégier autant que possible la simplicité, dans le cadre d'un échantillonnage et de procédures informatiques qui étaient des modèles de complexité. Peut-être y a t il eu un traumatisme lié à la multiplicité des bases et aux difficultés de traitement informatique des données, mais il nous apparaît maintenant, à l'expérience, que la méthode de partage des poids aurait probablement constitué une piste plus rigoureuse pour justifier que l'on n'ait pas calculé des probabilités de sélection des individus à partir de la (ou des) base(s) dans laquelle (lesquelles) ils figurent mais sans y avoir été échantillonnés. Elle n'aurait néanmoins pas réglé les autres questions.

3) En l'occurrence, dans le cas présent, les procédures mises en œuvre ont finalement conduit à un système de pondérations tout à fait satisfaisant avant redressement, respectant correctement les principales structures socio-démographiques (voir tableau 1). Doit-on s'attribuer un satisfecit, les traitements ''acrobatiques'' n'ont-ils pour l'occasion qu'un impact numérique modéré, y a t il une part de chance … ? en tout cas cela donne finalement raison, pour cette fois, à tous ceux qui considèrent que tout problème technique possède une solution acceptable.

REFERENCES

Lavallée, P. (2003), *Le sondage indirect ou la méthode généralisée de partage des poids*, Ellipses.

Särndal, C.-E., Swensson, B., Wretman, J. (1992), *Model assisted survey sampling*, New York : Springer-Verlag.