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

PROBLEMES THEORIQUES ET PRATIQUES DE LA CONSTRUCTION DE L'« EMEX » : COMMENT AMELIORER LA PRECISION DES EXTENSIONS REGIONALES DES ENQUETES NATIONALES GRACE A UN ECHANTILLONNAGE ADDITIONNEL ?

Marc Christine et Laurent Wilms[1]

RÉSUMÉ

Ce papier présente une méthode de constitution d'un échantillon destiné aux extensions régionales d'enquêtes nationales. Elle s'appuie sur une technique d'équilibrage dans un cadre conditionnel.

MOTS CLÉS : Équilibrage, équilibrage inverse, précision régionale, probabilités d'inclusion, stratification, tirage conditionnel.

1. INTRODUCTION

Les enquêtes nationales auprès des ménages réalisées en France, la plupart du temps en face à face, reposent sur un échantillon de logements tiré d'un échantillon-maître (EM). Cet échantillon est constitué, pour simplifier, d'un premier degré où l'on tire des unités primaires (UP) et d'un second degré où l'on tire les logements. Cet échantillon-maître a été conçu pour assurer une précision acceptable pour ces enquêtes nationales tout en limitant les coûts d'enquête et, notamment, les coûts de déplacement, grâce à la concentration des enquêtes dans les UP tirées au 1er degré.

Régulièrement, et ce de manière croissante dans le cadre des lois successives de décentralisation en France, les utilisateurs régionaux souhaitent disposer de résultats régionaux, dans le but d'améliorer la précision et de permettre des comparaisons d'une région à l'autre ou entre le niveau régional et le niveau national.

Pour répondre à cette demande, l'Institut national statistique français (Insee[2]) est amené à proposer la mise en place d'extensions régionales[3] pour les enquêtes nationales auprès des ménages qu'il réalise.

Dans le passé, ces extensions d'enquêtes étaient la plupart du temps gérées directement par les Directions régionales de l'Insee (DR), de manière déconnectée avec l'enquête nationale. Il a donc été jugé nécessaire, au niveau central, de construire un cadre standard permettant un traitement homogène des demandes d'extension régionale et une prise en charge globale des questions méthodologiques.

Ce cadre nécessite tout d'abord de définir une méthodologie de constitution des échantillons de logements pour ces extensions régionales. A cette fin, un système d'échantillonnage complémentaire au niveau régional a été mis en

[1] Marc Christine, INSEE, Unité Méthodes Statistiques F402 18, Boulevard Adolphe Pinard, F-75675 Paris cedex 14, marc.christine@insee.fr et Laurent Wilms, INSEE, Direction du Programme de Rénovation du Recensement L112 18, Boulevard Adolphe Pinard, F-75675 Paris cedex 14, laurent.wilms@insee.fr

[2] Institut national de la statistique et des études économiques.

[3] Il faut distinguer ici les « enquêtes purement locales, qui ***ne sont pas concernées par la présente problématique*** et les extensions régionales d'enquêtes nationales, qui ont des objectifs similaires, pour la région, à ceux du niveau national.

place. Cet outil qui va être décrit ci-après a été baptisé « **E**chantillon **M**aître pour les **Ex**tensions régionales » (**EMEX**). Il a été utilisé pour la première fois en 2002 pour l'enquête Santé.

Le papier exposera les difficultés théoriques et pratiques qui ont été rencontrées dans la mise en œuvre de cet échantillon et les différentes voies pour y répondre.

2. PRINCIPES GÉNÉRAUX DE LA CONSTITUTION DE L'EMEX

Comme son nom l'indique, cet outil procède de la même philosophie d'ensemble que celle de l'échantillon-maître, mis en place en 2000-2001. ***Mais la question essentielle est le souci d'une bonne « représentativité » régionale EM + EMEX, garante de la qualité des résultats.***

Le tirage des extensions qui a été mis en place se déroule donc selon un mode proche de celui du tirage des logements dans l'échantillon-maître. A l'intérieur de strates définies par le croisement de la région et du degré d'urbanisation, il est constitué par :

- un premier degré consistant en un tirage d'unités primaires spécifiques (dites UP-EMEX), en nombre fixé à l'intérieur de chaque strate, celles-ci ayant été constituées une fois pour toutes pour l'ensemble des extensions régionales.

- et un second degré dans lequel, pour chaque enquête concernée, les logements sont tirés dans les UP-EMEX des régions bénéficiant d'une extension[4] (en général par sondage aléatoire simple[5]).

Les principes généraux retenus sont les suivants :

- les UP-EMEX sont *disjointes* de celles retenues pour l'EM national : en cela, l'EMEX est un complément à l'EM.

- en zone rurale, les UP-EMEX (comme d'ailleurs les UP-EM) sont des regroupements de communes et, dans les agglomérations de plus de 100.000 habitants, il s'agit de regroupements d'immeubles contigus, baptisés *districts*. Dans les petites agglomérations, c'est l'agglomération elle-même qui est une UP[6].

- le nombre d'UP à tirer dans chaque strate a été défini a priori en anticipant la taille moyenne probable des extensions dans chaque région (de l'ordre de la fraction régionale de l'échantillon national).

La difficulté principale, dont la résolution fait l'objet du présent papier, est de définir le mode de tirage adéquat des UP-EMEX en vue d'assurer une bonne représentativité régionale de l'échantillon issu de l'EM et de l'EMEX.

[4] In fine, on dispose donc, dans une région bénéficiant d'une extension, de l'échantillon de logements issu des UP-EMEX qui s'ajoute à celui issu des UP-EM. Cependant, en général, seules quelques régions sont concernées par une extension (5 dans le cas de l'enquête Santé 2002).

[5] Dans la pratique, la chaîne informatique de tirage des logements dans les UP fonctionne ***de manière globale et intégrée*** : elle calcule les allocations de logements à tirer dans chaque UP (EM ou EMEX indistinctement) et réalise ce tirage sur l'ensemble des UP éligibles de chaque région (EM seul s'il n'y a pas d'extension, EM + EMEX s'il y en a une).

[6] L'EM, dans cette strate, a donc consisté en un tirage d'agglomérations.

3. LE TIRAGE DES UNITES PRIMAIRES DE L'EMEX : LA QUESTION DE L'EQUILIBRAGE ET LES PROBLEMES THEORIQUES QU'ELLE POSE

3.1 Rappels sur l'équilibrage

La précision résultant du tirage d'échantillons selon un plan de sondage donné peut être améliorée grâce à la technique de l'*équilibrage*.

On rappelle que l'équilibrage consiste à imposer qu'un estimateur de type HORVITZ-THOMPSON (H-T) du total de certaines variables d'intérêt, fabriqué à partir des unités tirées, prenne une valeur identique à celle (connue) de ce total sur l'ensemble de la population.

L'objectif est de faire en sorte que l'échantillon tiré constitue un meilleur modèle réduit de la population de référence au vu des variables d'équilibrage considérées, supposées bien corrélées avec les variables d'intérêt de l'enquête. Par abus de langage, on dira que ce tirage assure une meilleure « *représentativité* ».

On dispose à l'heure actuelle d'algorithmes de tirage équilibré permettant de sélectionner dans une population un échantillon d'unités statistiques vérifiant les contraintes d'équilibrage au niveau de cette population, *pour un jeu de probabilités d'inclusion fixé a priori.* L'un d'entre eux est la méthode dite du CUBE, mise au point et développée récemment par Jean-Claude DEVILLE et Yves TILLE.

En ce qui concerne l'EM, les UP avaient été tirées strate par strate.

- *Pour les UP rurales,* on s'était contenté d'assurer un équilibrage au niveau de *groupes de régions* (de façon à augmenter le nombre de degrés de liberté au moment du tirage), les *probabilités d'inclusion des UP étant proportionnelles à leur taille.* Concrètement, cela signifie qu'une UP rurale d'une région donnée, dotée de certaines caractéristiques socio-démographiques, peut être représentée statistiquement par une UP de caractéristiques voisines, *mais appartenant à une autre région au sein du même groupe de régions.* Les variables retenues pour l'équilibrage étaient : tranches d'âge, revenu net imposable, nombres de logements (principaux ou non). Il en résulte que la représentativité régionale (et, notamment, la couverture de l'ensemble de la région) n'a pas été assurée a priori dans l'EM[7].

- *Pour les districts des grosses agglomérations,* on s'est contenté d'un tirage aléatoire simple, également avec des conditions d'équilibrage au niveau de chaque agglomération [8].

La question de l'équilibrage se repose dans les mêmes termes pour le tirage des UP-EMEX mais conduit à une difficulté théorique importante qui va être explicitée dans le présent paragraphe.

3.2 Le problème de l'EMEX : un tirage « ex-post »

Comme il a été dit ci-dessus, les UP de l'EM ont été tirées sans qu'il soit possible d'assurer une vraie représentativité régionale. Cette opération a eu lieu en 1999-2000. Ce n'est qu'ensuite que la décision a été prise de concevoir et de tirer l'EMEX (2001-2002). A ce moment, la mise en place de l'EM, l'organisation des enquêtes et la constitution du réseau d'enquêteurs étant arrêtées une fois connues les UP-EM, *il n'était plus possible de remettre en cause le résultat du tirage de ces dernières.*

[7] On rappelle que l'enquête Emploi fait figure d'exception : elle est dotée d'un échantillon spécial (aréolaire) distinct de l'échantillon-maître et ce dernier a été construit sous la contrainte explicite de satisfaire des conditions de précision régionale, édictées par Eurostat.

[8] Les districts sont en général de petite taille et en très grand nombre. Ce sont des « grains » relativement homogènes. Ceci explique que l'on n'avait guère besoin de raffiner leur mode de tirage : on a donc fait un tirage aléatoire simple. A l'opposé, les unités primaires rurales sont en faible nombre, de grande taille (de 1800 à 3600 logements principaux), plus ou moins hétérogènes.

La question qui s'est alors posée était la suivante : ***peut-on tirer les Unités primaires de l'EMEX, une fois celles de l'EM tirées, tout en assurant des conditions d'équilibrage pour l'ensemble EM + EMEX (et pas pour le seul EMEX) au niveau de chaque région ?***

Pour mettre en œuvre des conditions d'équilibrage, il convient d'examiner comment construire des estimateurs sans biais de totaux régionaux à partir de la réunion des deux échantillons : EM et EMEX.

Il est clair que, disposant de deux échantillons relatifs à la région, l'un correspondant à la partie régionale de l'EM, l'autre issue de l'EMEX, on accroît la précision de l'estimation d'un total régional en combinant *toutes les observations disponibles*[9].

Le contexte spécifique de l'EMEX est que les nouvelles UP doivent être tirées *conditionnellement à la réalisation du tirage de l'EM.* Pour réaliser ce tirage, il va donc falloir se donner une loi conditionnelle de tirage. A priori, on doit avoir une maîtrise au moins partielle de cette loi. Elle sera résumée (avec perte d'information cependant) par les *probabilités d'inclusion conditionnelles d'ordre 1* des Unités primaires tirées pour l'EMEX.

Dans ce cadre, *on considère comme donnée intangible la loi de tirage des Unités primaires de l'EM* ; cette loi n'est connue également que partiellement par l'intermédiaire des probabilités d'inclusion d'ordre 1 de ces dernières.

3.3 Le cadre théorique

Pour définir un cadre théorique, notons ***P*** l'ensemble des UP d'une région donnée (population de référence), S_1 et S_2 respectivement, la fraction régionale de l'EM et la partie de l'EMEX relative à la région considérée, sélectionnées dans ***P***. La réunion de S_1 et S_2 est notée S. Pour une unité i de la population, nous noterons :

Π_i^1 la probabilité d'inclusion de cette unité dans l'échantillon S_1, Π_i^{2/S_1} la probabilité d'inclusion d'ordre 1 de l'unité i dans l'échantillon S_2 *conditionnellement à la réalisation de l'échantillon* S_1 (il s'agit d'une fonction de S_1 dont les valeurs prises dépendent de la réalisation s_1 de S_1). Π_i la probabilité d'inclusion finale de l'unité i dans l'échantillon global S. On démontre que ces différentes probabilités sont reliées entre elles par la relation

$$\Pi_i = \Pi_i^1 + E\,(\Pi_i^{2/S_1} 1_{i \notin S_1}) \qquad (3.1)$$

(L'espérance est ici prise par rapport à la loi de tirage du 1er échantillon).

Dans la pratique, les échantillons S_1 et S_2 seront *disjoints*. Puisque les UP de l'EMEX doivent être distinctes de celles de l'EM. Cela entraîne que

$$\Pi_i^{2/S_1} = 0 \text{ pour i} \in S_1 .$$

Les Π_i^1 sont donc les probabilités d'inclusion dans le 1er échantillon (EM) : *ces probabilités sont données une fois pour toutes.* Pour les UP rurales et les petites agglomérations, elles sont proportionnelles à la taille en termes de nombre de logements principaux au RP de 1999 ; pour les districts tirés dans les grosses UP, elles sont constantes (tirage à probabilités égales).

Les probabilités d'inclusion finales Π_i seront imposées également par les conditions déduites du mode de tirage aux degrés ultérieurs et par celles que l'on souhaitera imposer pour la pondération des unités finales (logements) : par exemple, *l'équiprobabilité finale des logements* (sondage « autopondéré ») (cf. § 6.2).

[9] L'objectif est de construire, à partir de la réunion EM + EMEX, des estimateurs *sans biais* de totaux relatifs à la région. L'EM seul répond aussi à cet objectif (théorie des domaines), même s'il n'assure pas une représentativité régionale correcte. En revanche, l'EMEX seul ne peut y parvenir car il est tiré dans un complément à l'EM, « déséquilibré » du fait du tirage des UP de l'EM à probabilités inégales.

Les paramètres sur lesquels on peut agir sont ceux qui définissent le tirage du 2ème échantillon conditionnellement au premier, résumé par les probabilités d'inclusion conditionnelles.

3.4 Les difficultés théoriques

L'une des difficultés réside dans le fait que la relation (3.1) ci-dessus entre les différentes probabilités est complexe et qu'il n'est pas simple (voire impossible), pour des probabilités finales données, de déterminer des probabilités conditionnelles qui conduisent à ces probabilités finales. Cela supposerait en pratique de connaître la loi complète de tirage du 1er échantillon et pas seulement les probabilités d'inclusion d'ordre 1 (sauf dans certains cas particuliers, comme celui où le tirage de l'échantillon S_1 est aléatoire simple).

La seconde difficulté que l'on rencontre tient à la nature des contraintes d'équilibrage que l'on veut assurer.

- En effet, dans l'approche classique de l'équilibrage, la condition d'équilibrage qu'on souhaite atteindre sur une variable X quand on tire un échantillon S dans la population ***P*** est de la forme : $\hat{T} = T$, où T est le vrai total de la variable X, connu sur la population au sein de laquelle on tire l'échantillon, et $\hat{T}$ l'estimateur de ce total à partir de l'échantillon.

Dans le cadre d'un estimateur classique de Horvitz-Thompson, cette équation s'écrit

$$\sum_{i \in S} \frac{x_i}{\Pi_i} = \sum_{i \in P} x_i \ .$$

Le 2ème terme de cette équation représente le total de la variable X sur la région considérée, c'est-à-dire que *le total sur lequel on équilibre doit être le total des observations, sur la population, de la variable (éventuellement vectorielle) d'équilibrage.*

On notera que la forme de l'estimateur utilisé (ici Horvitz-Thompson) *ne dépend pas fonctionnellement des variables d'équilibrage* : elle est calculée de telle sorte que l'estimateur obtenu soit un *estimateur sans biais du total de n'importe quelle variable Y.* On est alors dans la configuration où : $\forall$ Y : E [$\sum_{i \in s} \frac{Y_i}{\Pi_i}$] = $\sum_{i \in P} Y_i$, l'espérance étant prise par rapport à la loi de tirage de l'échantillon.

Pour la variable d'équilibrage X, l'égalité ci-dessus est vraie rigoureusement, et non plus seulement en espérance.

Dans le cas qui nous intéresse, S étant la réunion des deux échantillons disjoints S_1 et S_2, cette condition s'écrit aussi

$$\sum_{i \in S_2} \frac{x_i}{\Pi_i} = \sum_{i \in P} x_i - \sum_{i \in S_1} \frac{x_i}{\Pi_i} . \qquad (3.2)$$

Or, l'échantillon S_1 ayant été tiré une fois pour toutes, le tirage de S_2 doit s'effectuer *conditionnellement à celui de* S_1, dans ***P*** - S_1. Les conditions d'équilibrage ne peuvent donc être maîtrisées que dans le cadre de ce tirage conditionnel, et vis-à-vis d'une population-mère qui est ici ***P*** - S_1.

Aussi, les seules conditions d'équilibrage que l'on peut atteindre lors de ce tirage s'écrivent-elles nécessairement (pour une variable d'équilibrage Z quelconque) sous la forme

$$\sum_{i \in S_2} \frac{Z_i}{\Pi_i^{2/S_1}} = \sum_{i \in P - S_1} Z_i \qquad (3.3)$$

C'est ainsi que se transpose le cadre classique et qu'on peut, pour assurer ce dernier équilibrage, utiliser des outils standards.

Dans le cadre classique, l'obtention d'échantillons équilibrés peut en effet être assurée dans la pratique en utilisant l'algorithme issu de la théorie du CUBE, en utilisant un estimateur de type Horvitz-Thompson.

Transposée au cas présent, dans le cadre d'un tirage conditionnel, cette méthode, pour être employée, nécessiterait que l'on ait la relation

$$E\left[\sum_{i\in S_2}\frac{Y_i}{\Pi_i^{2/S_1}} \,/\, S_1\right] = \sum_{i\in P-S_1} Y_i \ ,$$

pour toute variable Y dont on estimerait le total à partir du tirage conditionnel du 2ème échantillon (l'espérance est ici prise *vis-à-vis de la loi conditionnelle de S_2 sachant S_1*), l'échantillon S_2 *étant tiré de telle sorte qu'il y ait égalité vraie (et non plus seulement en espérance) pour la variable d'équilibrage Z.*

Mais les Π_i, probabilités d'inclusion de l'unité i dans l'échantillon global S, étant imposées a priori, il n'est en général pas possible de trouver les Π_i^{2/S_1} assurant à la fois l'identité des équations (3.2) et (3.3) et la relation (3.1) écrite ci-dessus.

Clairement, et sauf cas très particuliers qui seront examinés dans le §4, le problème posé n'entre donc pas dans les conditions d'application du tirage CUBE pour un équilibrage « classique ».

Pour résoudre ce problème de manière générale, on sera donc obligé de recourir à des solutions approchées.

4. CONTEXTES PARTICULIERS OÙ LE PROBLÈME POSSÈDE UNE SOLUTION EXACTE.

Trois cas particuliers permettent de résoudre effectivement le problème posé.

4.1 Tirages simultanés de l'EM et de l'EMEX

Si la problématique de l'EMEX avait été prise en compte au moment du tirage de l'EM et que l'on eût tiré simultanément les deux échantillons, il y aurait alors eu un moyen très simple de les équilibrer simultanément sur une même variable X, selon la procédure suivante.

a) Tirer un échantillon S de k unités dans ***P*** avec des probabilités d'inclusion Π_i en équilibrant sur X

$$\sum_{i\in S}\frac{x_i}{\Pi_i} = \sum_{i\in P} x_i \ .$$

Cet échantillon S correspond à la réunion de l'EM et de l'EMEX.

b) Enlever de S, par sondage aléatoire simple au taux f, un échantillon S_2 de taille fixe (k - a), qui représentera l'EMEX, en l'équilibrant sur la variable $Z_i = \frac{x_i}{\Pi_i}$

$$\sum_{i\in S_2}\frac{1}{f}\frac{x_i}{\Pi_i} = \sum_{i\in S}\frac{x_i}{\Pi_i} \ .$$

c) Alors, l'échantillon $S - S_2$, qui représenterait l'EM seul, assure à chaque unité i la probabilité de sélection $\Pi_i^1 = (1 - f)\ \Pi_i$. En particulier, si Π_i est proportionnelle à la taille de l'unité i, il en va de même de Π_i^1. De surcroît, on vérifie facilement que

$$\sum_{i \in S - S_2} \frac{x_i}{(1-f)\Pi_i} = \sum_{i \in P} x_i \ ,$$

c'est-à-dire que $S - S_2$ est également équilibré sur la même variable X au niveau de la population ***P***.

Concrètement, cela veut dire qu'on tire systématiquement d'emblée un échantillon plus gros, équilibré sur la variable d'intérêt ; dans les régions pourvues d'une extension régionale, c'est cet échantillon qui est retenu (qui représente EM + EMEX) ; dans les autres, on procède à une seconde phase par sondage aléatoire simple pour sélectionner le sous-échantillon utile (l'EM).

L'inconvénient de la procédure est double :

- d'une part, elle n'aurait pu être mise en jeu qu'en définissant à l'avance le tirage des deux échantillons (alors qu'en réalité la nécessité et l'intérêt de l'EMEX ne sont apparus que *postérieurement* au tirage de l'EM).

- d'autre part, pour les régions sans extension, on risque de perdre en variance du fait de l'échantillonnage en deux phases au lieu d'un tirage direct de l'EM.

4.2 Cas de tirages successifs équilibrés et à probabilités égales

Comme cela a été évoqué ci-dessus (§ 3.1), ce cas s'est rencontré avec le tirage des districts dans les grosses agglomérations.

Ce cas est justiciable du résultat théorique suivant (dont la démonstration est laissée au lecteur) :

Proposition :

Si l'on tire un échantillon S_1 dans une population ***P*** selon un tirage équilibré sur une variable X et que le tirage soit à *probabilités égales* ; et si l'on tire un deuxième échantillon (S_2) dans ***P*** – S_1 tel que le tirage conditionnel de S_2 sachant S_1 soit à probabilités égales et équilibré sur la même variable X (au niveau de ***P*** – S_1) ;

Alors $S = S_1 \cup S_2$est équilibré sur X au niveau de la population totale ***P*** (vis-à-vis de la loi de tirage finale de S).

4.3 Utilisation d'une forme particulière d'estimateur d'Horvitz-Thompson

Prenons comme estimateur du total d'une variable Y quelconque $\hat{T} = \sum_{i \in S_1} Y_i + \sum_{i \in S_2} \frac{Y_i}{\Pi_i^{2/S_1}}$. Alors, on obtient un estimateur sans biais. De surcroît, à partir du moment où S_2 est sélectionné dans ***P*** – S_1 selon un sondage équilibré sur X, cet estimateur assure la propriété d'équilibrage sur X au niveau de la population ***P***, pour l'échantillon global S.

Tout se passe comme si S_1 était une population de référence : on prend le total non pondéré sur cette sous-population et on rajoute une estimation de type Horvitz-Thompson en utilisant la loi conditionnelle de tirage du $2^{ème}$ échantillon. On ne tient donc pas compte des propriétés inférentielles de l'échantillon S_1, ce qui n'est certainement pas optimal.

5. RETOUR AU CAS GÉNÉRAL : UNE SOLUTION APPROCHÉE, *L'ÉQUILIBRAGE INVERSE.*

On va se placer dans le cadre où l'on utilise un estimateur de Horvitz-Thompson sur la réunion des deux échantillons, avec des probabilités finales d'inclusion Π_i *fixées*.

L'équation d'équilibrage sur une variable X devrait s'écrire

$$\sum_{i\in S_1} \frac{x_i}{\Pi_i} + \sum_{i\in S_2} \frac{x_i}{\Pi_i} = \sum_{i\in P} x_i \ . \qquad (5.1)$$

Comme cette équation ne peut être assurée a priori, pour une loi du 1er échantillon donnée, on va relâcher une contrainte : on va admettre que l'on n'imposera plus de respecter exactement la valeur souhaitée des probabilités finales d'inclusion Π_i, mais qu'on s'en écartera « de manière acceptable » : on cherchera donc de nouvelles probabilités d'inclusion finales, notées $\tilde{\Pi}_i$, proches des Π_i au sens d'une certaine distance, de telle sorte que, en introduisant une loi de tirage conditionnel ad hoc, on satisfasse l'équation d'équilibrage global (5.1). Concrètement, on cherchera à minimiser un critère de la forme : $\sum_{i\in P} d(\tilde{\Pi}_i, \Pi_i)$, tout en respectant l'équation d'équilibrage (5.1), où ***P*** est l'ensemble des UP constituant une strate (croisement d'une région et d'une catégorie d'agglomération) et où d désigne une distance (distance euclidienne ou du χ^2 en pratique). L'équation d'équilibrage doit être assurée dans le cadre du tirage conditionnel de S_2 sachant S_1 (car les UP-EM ont déjà été tirées). On cherche donc une forme équivalente à l'équation (5.1), qui s'interprète effectivement comme une condition d'équilibrage à respecter lors du tirage conditionnel de S_2 sachant S_1. Pour cela, on va écrire l'équation (1) sous la forme

$$\sum_{i\in S_2} \frac{z_i}{\tilde{\Pi}_i^{2/S_1}} = \sum_{i\in P} x_i \ - \sum_{i\in S_1} \frac{x_i}{\tilde{\Pi}_i}$$

où l'on a posé : $z_i = x_i \frac{\tilde{\Pi}_i^{2/S_1}}{\tilde{\Pi}_i}$ et où $\tilde{\Pi}_i^{2/S_1}$ désigne une nouvelle probabilité conditionnelle de tirage de l'UP n° i sachant S_1.

Pour que cette équation s'interprète comme une véritable contrainte d'équilibrage sur le total des variables z_i sur la population au sein de laquelle on réalise ce tirage conditionnel (soit ***P*** – S_1), il faut qu'elle s'écrive sous la forme :

$$\sum_{i\in S_2} \frac{z_i}{\tilde{\Pi}_i^{2/S_1}} = \sum_{i\in P-S_1} z_i \ . \qquad (5.2)$$

Ces conditions sont satisfaites si et seulement si

$$\sum_{i\in P} x_i \ - \sum_{i\in S_1} \frac{x_i}{\tilde{\Pi}_i} = \sum_{i\in P-S_1} z_i = \sum_{i\in P-S_1} x_i \frac{\tilde{\Pi}_i^{2/S_1}}{\tilde{\Pi}_i} \ . \qquad (5.3)$$

On va prendre un <u>cas simple de détermination de la loi conditionnelle</u> de tirage de S_2 sachant S_1 définie comme suit :

$$\pi_i^{2/s_1} = \begin{cases} 0 \text{ si } i \text{ appartient à } S_1 \\ \mu_i \text{ si } i \text{ n'appartient pas à } S_1, \end{cases}$$

c'est-à-dire que Π_i^{2/S_1} ne dépend de S_1 que par l'intermédiaire de l'indicatrice $\mathbf{1}_{i\in S_1}$. Les coefficients μ_i dépendent de l'unité i mais leur valeur est fixée ex-ante, indépendamment de la réalisation de l'échantillon S_1.

On obtient alors, d'après l'équation (3.1)

$$\Pi_i = \Pi_i^1 + \mu_i \sum_{S_1 / i \notin S_1} P(S_1 = s_1)$$
$$= \Pi_i^1 + \mu_i \left(1 - \Pi_i^1\right)$$

ou encore

$$\Pi_i = (1 - \mu_i)\ \Pi_i^1 + \mu_i$$

On en tire

$$\mu_i = \frac{\Pi_i - \Pi_i^1}{1 - \Pi_i^1} \tag{5.4}$$

Π_i^1 étant fixé, il y a donc, dans le cadre de cette détermination de la loi conditionnelle, une correspondance biunivoque entre les Π_i et les Π_i^{2/S_1}.

Ainsi, par exemple, dans le cas où l'on souhaiterait que les probabilités *finales* d'inclusion des unités primaires dans la réunion des deux échantillons soient, comme elles le sont pour le seul EM, proportionnelles à leur taille, avec un effectif d'échantillon fixé k (à l'intérieur d'une strate donnée), cela imposerait la condition :

$$\forall\ i \in \boldsymbol{P} :\ \Pi_i = k \frac{T_i}{T}$$

où T_i est la taille de l'unité i et T la taille totale de la population (somme des tailles des unités).

Avec une condition analogue pour l'EM : $\Pi_i^1 = a \frac{T_i}{T}$, où a est le nombre d'unités tirées dans l'EM, on obtiendra au final la relation

$$\mu_i = \frac{(k-a)T_i}{T - aT_i}.$$

On obtient en général des probabilités admissibles, à valeurs dans [0,1][10].

Lorsqu'on remplace les vraies probabilités finales d'inclusion Π_i par les probabilités approchées $\tilde{\Pi}_i$, les probabilités conditionnelles qui permettent d'obtenir ces $\tilde{\Pi}_i$ doivent être modifiées de manière cohérente en respectant la relation (5.4) précédente.

On aura alors la relation, pour tout i n'appartenant pas à S_1

$$\tilde{\Pi}_i^{2/S_1} = \frac{\tilde{\Pi}_i - \Pi_i^1}{1 - \Pi_i^1}.$$

[10] Sauf dans certains cas particuliers où des T_i seraient très grands.

L'équation de contrainte (5.3) s'écrira donc exclusivement en fonction des $\tilde{\Pi}_i$ comme

$$\sum_{i\in P} x_i \; - \sum_{i\in S_1} \frac{x_i}{\tilde{\Pi}_i} \; = \; \sum_{i\in P-S_1} x_i \frac{1-\frac{\Pi_i^1}{\tilde{\Pi}_i}}{1-\Pi_i^1}\,.$$

Finalement, dans ce cas, le problème revient à résoudre le programme (P) de recherche des $\tilde{\Pi}_i$

$$\text{Min} \sum_{i\in P} d(\tilde{\Pi}_i, \Pi_i)$$

sous la contrainte

$$\sum_{i\in P} x_i \; - \sum_{i\in S_1} \frac{x_i}{\tilde{\Pi}_i} \; = \; \sum_{i\in P-S_1} x_i \frac{1-\frac{\Pi_i^1}{\tilde{\Pi}_i}}{1-\Pi_i^1}\,.$$

On ajoutera également éventuellement la contrainte de taille fixe donnée par $\sum_{i\in P} \tilde{\Pi}_i$ = k. Au total, la résolution de ce programme permet de déterminer de nouvelles probabilités d'inclusion finales $\tilde{\Pi}_i$, d'où l'on déduira des probabilités d'inclusion conditionnelles $\tilde{\Pi}_i^{2/S_1}$ vérifiant les propriétés suivantes.

Si l'on tire le 2ème échantillon selon ces probabilités d'inclusion conditionnelles, en l'équilibrant sur le total de la variable $z_i = x_i \frac{\tilde{\Pi}_i^{2/S_1}}{\tilde{\Pi}_i}$ au niveau de la population $\boldsymbol{P} - S_1$, alors, quand on réunit les deux échantillons, d'une part on assure à chaque unité i une probabilité d'inclusion globale égale à $\tilde{\Pi}_i$, d'autre part on assure un équilibrage sur le total de la variable X au niveau de la population ***P*** tout entière.

Une différence importante à noter par rapport à la procédure classique d'équilibrage est que, dans cette dernière, on peut tirer n'importe quel échantillon avec des probabilités d'inclusion ***fixées ex-ante*** en astreignant l'échantillon à vérifier des contraintes d'équilibrage. Dans la présente procédure, on va procéder de manière « inverse » : on détermine des probabilités d'inclusion assurant l'équation d'équilibrage, qui ***dépendront, de ce fait, des variables d'équilibrage choisies*** et des valeurs qu'elles prennent dans la population. ***D'où le nom « d'équilibrage inverse » que l'on peut attribuer à la procédure.***

Une autre différence est à souligner *avec les procédures de calage*. Dans celles-ci, on modifie également les pondérations initiales (égales aux inverses des probabilités d'inclusion) en cherchant à minimiser une certaine distance entre pondérations initiales et finales, pour forcer les estimateurs à satisfaire des contraintes de calage. Mais les nouvelles pondérations ne sont calculées ex-post que *sur les seules unités tirées*. Au contraire, dans la procédure de l'équilibrage inverse, on recalcule des probabilités finales d'inclusion pour toutes les unités de la population.
L'approche de l'équilibrage inverse pose toutefois la question de l'existence de solutions admissibles (avec tous les $\tilde{\Pi}_i$ appartenant à $[\Pi_i^1, 1]$; quant à la résolution, elle se fait par des méthodes numériques[11].

[11] Par une procédure SAS IML.

6. CONCLUSION : INTÉRÊT DE L'EMEX ET PERSPECTIVES D'AMÉLIORATION

On peut tirer deux types de conclusions de cette étude :

6.1 D'ordre pratique

L'EMEX assure une coordination de l'échantillon national et des échantillons d'extension permettant un calcul unique de pondération, servant pour les estimations nationales et régionales et l'homogénéisation des méthodes de traitement. Mais, mis en place pour la 1ère fois en 2002, l'EMEX est encore un outil trop « jeune » pour qu'on puisse en tirer à l'heure actuelle un bilan complet.

6.2 D'ordre théorique

Le papier a permis de mettre en lumière la difficulté d'obtenir une propriété d'équilibrage dans un cadre d'échantillonnages successifs et apporte un éclairage sur la notion d'échantillonnage conditionnel. Il a mis en évidence une solution approchée admissible pour résoudre cette difficulté.

6.3 Le futur : conception simultanée

On a vu que le problème aurait été très simple a résoudre si l'on avait tiré simultanément les UP de l'EM et de l'EMEX (cf. § 4.1) et qu'au contraire la difficulté provenait du fait que la nécessité des extensions régionales ne s'était dégagée que *postérieurement* au tirage de l'échantillon-maître national.

Dans le cadre des travaux que mène actuellement l'Insee sur la construction des nouveaux échantillons issus du recensement rénové de la population (fondé sur un système d'échantillonnage rotatif), cette préoccupation de pouvoir répondre à des demandes d'extension régionale par la constitution d'un échantillonnage complémentaire sera prise en compte dès l'amont.

Le contexte sera donc différent et ne devrait pas conduire à être confronté aux mêmes problèmes théoriques que ceux mis en lumière dans ce papier.

RÉFÉRENCES

Bourdallé G., Christine M., Wilms L., « Echantillons Maître et Emploi », *Insee-Méthodes : VII èmes JMS, 4-5 décembre 2000*, n° 100 (tome 1), p. 139-241.

Bousabaa A., Lieber J. et Sirolli. R (1999) : la macro Cube, Document de travail, ENSAI, Rennes.

Deville J-C. et Tillé Y. (2001) : variance reduction using balanced sampling : the Cube method.

Favre A-C.et Tillé Y. : coordination, combination and extension of optimal balanced samples, à paraître.

Rousseau S. et Tardieu F. (2004) : la Macro SAS CUBE d'échantillonnage équilibré, documentation de l'utilisateur. *Cette documentation peut être téléchargée sur le site de l'Insee : www.insee.fr, rubrique « nomenclatures, définitions, méthodes ».*