Constitution de l'échantillon-maître pour les extensions régionales : une procédure de tirage d'échantillons équilibrés emboîtés.

Marc CHRISTINE (*), Emmanuel GROS (**)

(*) Insee, Unité Méthodes Statistiques
(**) Insee, Unité Méthodologie Statistique - Entreprises

Introduction

En France, chaque enquête nationale auprès des ménages repose sur un échantillon de logements spécifique. L'ensemble de ces échantillons[1] est cependant constitué suivant un schéma directeur commun : celui d'un sondage à deux degrés. Dans un premier temps, des unités primaires sont sélectionnées aléatoirement. Dans un second temps, les logements à enquêter sont tirés au sein de ces unités primaires, et ce, pour chaque enquête Ménages. Cette ligne directrice a été retenue afin d'assurer une précision acceptable pour les enquêtes nationales tout en limitant les coûts d'enquête, notamment les coûts de déplacement des enquêteurs. Dans certains cas, l'échantillon de logements est adapté aux particularités de la thématique abordée. A titre d'exemple, l'enquête « Modes de garde », qui s'intéresse aux dispositifs à disposition des parents pour faire garder leurs enfants en bas âge, s'appuie sur un échantillon exclusivement composé de ménages ayant au moins un enfant de moins de sept ans.

Le tirage de premier degré des unités primaires est réalisé une fois pour toutes, traditionnellement après chaque recensement, et les unités ainsi retenues composent l'échantillon-maître (EM). Le système actuel d'échantillonnage des enquêtes Ménages s'articule donc autour de ce concept d'échantillon-maître, qui constitue la base de sondage principale au sein de laquelle sont sélectionnés les logements à enquêter. Cette base de sondage nationale se voit complétée, dans le cadre de l'exploitation régionale des résultats des enquêtes, par une base de sondage supplémentaire : l'échantillon-maître pour les extensions régionales (EMEX).

La mise en place du recensement rénové de la population, désormais réalisé en continu, permet dorénavant l'actualisation annuelle de la base de sondage des enquêtes Ménages et a ainsi conduit à une refonte complète du système d'échantillonnage de ces enquêtes – définition et construction des unités primaires, tirage de nouveaux échantillon-maître et échantillon-maître pour les extensions régionales et procédure de tirage final des logements au sein de l'échantillon-maître – à travers le projet Octopusse[2]. Ce nouveau système d'échantillonnage des enquêtes ménages, fondé sur les fichiers de logements issus du nouveau recensement « rotatif », fait l'objet d'une description complète dans une autre communication aux JMS 2009[3].

Le présent article se focalise sur la construction de l'échantillon-maître pour les extensions régionales mise en œuvre dans Octopusse, qui a abouti au développement d'une procédure spécifique de tirage en deux phases des unités primaires servant à alimenter l'EM et EMEX : elle permet d'obtenir des échantillons équilibrés emboîtés, représentatifs à chaque niveau de tirage pour les variables d'équilibrage introduites.

[1] À l'exception de ceux relatifs à l'enquête Emploi, qui reposent sur un système aréolaire et seront désormais construits à partir des fichiers de la taxe d'habitation.
[2] Organisation Coordonnée de Tirages Optimisés Pour une Utilisation StatiStique des Echantillons.
[3] Sébastien Faivre. *Le projet Octopusse de nouvel échantillon-maître de l'Insee.* JMS 2009

1. Problématique, contexte et notations

Dans le cadre du projet Octopusse, la construction de nouvelles unités primaires a conduit à la spécification d'un nouveau plan de sondage pour l'échantillon-maître, exposé en détail dans la présentation aux JMS 2009 du système Octopusse. Nous rappelons ici les différentes caractéristiques de l'échantillon-maître, pour une région donnée :

- les ZAE[4] exhaustives sont au nombre de k_0 et représentent un nombre de résidences principales $N^{exh} = \sum_{\{i / ZAE\ i\ exhaustive\}} N_i$;

- le nombre de ZAE non exhaustives à tirer au sein de la région est noté k ;

- les probabilités d'inclusion des ZAE sont notées π_i : le tirage des ZAE s'effectuant proportionnellement à leur taille, cette probabilité vaut $\pi_i = k \frac{N_i}{N - N^{exh}}$ pour les ZAE non exhaustives, où N_i désigne la taille, en nombre de résidences principales, de la ZAE i et N la taille totale de la région ;

- enfin l'échantillon-maître obtenu est équilibré sur un vecteur des variables d'équilibrage retenues, noté X par la suite.

Ces caractéristiques de l'échantillon-maître étant posées, la problématique de l'échantillon-maître pour les extensions régionales est la suivante : il s'agit de constituer une base de sondage complémentaire de l'échantillon-maître, c'est-à-dire un ensemble d'unités primaires additionnelles (celles-ci ayant été construites une fois pour toutes pour l'EM), de même taille, de telle sorte que l'ensemble EM+EMEX possède une représentativité régionale suffisante, tout en préservant les caractéristiques de l'échantillon-maître, en particulier en termes de représentativité nationale.

Or la mise en place de l'EMEX précédent s'était heurtée à un problème théorique complexe[5]. En effet, les travaux de conception de l'EMEX ayant abouti après la construction de l'EM, il avait fallu constituer l'EMEX conditionnellement au tirage de l'EM, c'est-à-dire tirer un complément d'unités primaires conciliant les trois objectifs suivants :

- l'EM ayant déjà été construit, les unités primaires de l'EMEX devaient être tirées dans le complémentaire de l'EM ;

- les probabilités de tirage des unités primaires de l'EMEX devaient être – comme c'est le cas dans l'EM – proportionnelles à leur nombre de résidences principales ;

- le tirage des unités de l'EMEX devait assurer un équilibrage pour l'ensemble EM+EMEX sur les variables ayant servi à l'équilibrage de l'EM.

Une telle opération est particulièrement complexe à mettre en œuvre, et il n'avait pas été possible à l'époque de trouver un plan de sondage pour l'EMEX conditionnellement à l'EM conduisant à un ensemble EM+EMEX équilibré et à probabilités proportionnelles à la taille. La condition d'équilibrage avait alors été abandonnée au profit du respect des probabilités d'inclusions proportionnelles à la taille.

Le contexte actuel est plus favorable, dans la mesure où l'impératif de construction de l'EMEX a été pris en compte dès le début du projet Octopusse : l'EM et l'EMEX sont ainsi constitués simultanément. Les caractéristiques de l'EM ayant déjà été définies – probabilités d'inclusion proportionnelles à la taille des ZAE et choix des variables d'équilibrage –, l'objectif est donc de tirer conjointement l'EM et l'EMEX, de telle sorte que les caractéristiques de l'EM soient préservées, que les deux échantillons soient disjoints et que l'ensemble EM+EMEX présente des propriétés statistiques adéquates.

[4] Zones d'Action Enquêteurs, nouveau nom des unités primaires dans Octopusse.
[5] Cf. Marc Christine et Laurent Wilms. *Problèmes théoriques et pratiques de la construction de l'EMEX.* Insee, Actes des Journées Méthodologie Statistique 2002.

Plus précisément, on se place dans un cadre similaire à celui du tirage de l'échantillon-maître : en vertu du principe de stratification régionale, on raisonne donc région par région. L'objectif est alors de définir un plan de sondage permettant de tirer conjointement deux échantillons de ZAE EM et EMEX disjoints vérifiant les propriétés suivantes :

- la taille de l'EMEX doit être semblable à celle de l'EM. En effet, la pratique récente en matière d'extensions régionales montre qu'un doublement du nombre de ZAE EM+EMEX – du moins pour les ZAE non exhaustives – est suffisant pour la majorité des enquêtes : ceci permet d'assurer une précision régionale satisfaisante et la base de logements ainsi constituée est de taille suffisante pour assurer la réalisation d'enquêtes à extension régionale pendant cinq ans minimum sans jamais interroger deux fois un même ménage (principe de disjonction des échantillons des enquêtes ménages) ;

- l'EM ainsi construit doit posséder les mêmes caractéristiques que s'il avait été constitué de manière indépendante : probabilités d'inclusion identiques et équilibrage sur les mêmes variables ;

- enfin l'ensemble EM+EMEX doit être équilibré sur les mêmes variables que l'EM, et ce afin d'assurer une bonne représentativité régionale de l'ensemble.

2. De la nécessité de développer une procédure spécifique au problème

La première solution envisagée consistait à utiliser un mode de tirage préexistant et qui avait été mis en œuvre pour la construction des groupes de rotation du nouveau recensement[6]. Cette méthode permet en effet de sélectionner deux échantillons disjoints, équilibrés sur la même variable vectorielle X et avec des probabilités d'inclusion π_i, selon le principe suivant (dont la démonstration est fournie en annexe 1).

- On tire tout d'abord un premier échantillon s_1 avec les probabilités $\pi_i^1 = \pi_i$, équilibré sur les variables X et $\frac{X}{1-\pi}$: la première variable assure l'équilibrage de s_1 sur X, tandis que la seconde variable permet d'obtenir un équilibrage du complémentaire de s_1 sur X.

- On sélectionne ensuite le second échantillon s_2 dans le complémentaire de s_1, avec les probabilités d'inclusion conditionnelles à s_1 $\pi_i^{2/s_1} = \frac{\pi_i}{1-\pi_i} 1_{i \in \bar{s}_1}$ et équilibré sur la variable $\frac{X}{1-\pi}$. L'échantillon ainsi obtenu est bien équilibré sur la variable X avec des probabilités d'inclusion π_i.

Si l'on applique cette procédure pour le tirage simultané de l'EM et de l'EMEX, on obtient ainsi deux échantillons disjoints, respectant les probabilités d'inclusion π_i voulues et équilibrés tous les deux sur X. En outre, ces deux échantillons étant disjoints, on a :

$$P(i \in EM \cup EMEX) = P(i \in EM) + P(i \in EMEX) = 2\pi_i$$

et

[6] Cf. Guillaume Chauvet et Yves Tillé. *De nouvelles macros SAS d'échantillonnage équilibré*. Insee, Actes des Journées Méthodologie Statistique 2005.

$$\sum_{i\in EM\cup EMEX}\frac{X_i}{2\pi_i}=\sum_{i\in EM}\frac{X_i}{2\pi_i}+\sum_{i\in EMEX}\frac{X_i}{2\pi_i}=\frac{1}{2}(\sum_{i\in U}X_i+\sum_{i\in U}X_i)=\sum_{i\in U}X_i$$

⇒ L'ensemble EM+EMEX est donc bien équilibré sur le vecteur X.

D'un point de vue théorique, cette méthode répond donc parfaitement au problème posé par le tirage simultané de l'EM et de l'EMEX, et présente de plus l'avantage de fournir des échantillons EM et EMEX symétriques et tous les deux équilibrés. Nonobstant, l'application pratique de ce procédé se heurte à deux problèmes fondamentaux :

- en premier lieu, cette procédure nécessite de doubler le nombre de variables d'équilibrage, puisqu'il faut, lors du premier tirage, équilibrer l'échantillon et son complémentaire. Or dans le cadre du tirage de l'EM, le nombre de ZAE constituées par région est relativement faible – de l'ordre de la centaine –, et l'algorithme CUBE ne parvient pas à respecter exactement toutes les contraintes d'équilibrage. En particulier, le complémentaire du premier échantillon n'est jamais correctement équilibré, ce qui conduit à un équilibrage de mauvaise qualité pour le second échantillon ;

- en outre, les quantités $\frac{\pi_i}{1-\pi_i}$ ne sont pas systématiquement inférieures à 1. Le tirage du deuxième échantillon ne peut alors s'effectuer selon les bonnes probabilités, ce qui invalide la procédure.

En conséquence, il n'a pas été possible d'utiliser cette méthode, et il a fallu mettre au point une procédure de tirage spécifique.

3. Un système de tirage en deux phases

3.1. Le plan de sondage retenu

Très rapidement, les recherches se sont orientées vers un sondage en deux phases, avec une première phase de tirage de l'ensemble EM+EMEX de taille 2k, suivi d'une seconde phase de partition de cet échantillon en deux parties de même taille k. Le plan de sondage finalement retenu est le suivant :

- préalablement à tout tirage, on commence par retirer de la base de sondage les k_0 ZAE exhaustives définies dans le cadre de l'EM seul. Ces ZAE seront affectées automatiquement à l'EM, ainsi qu'à l'ensemble EM+EMEX. L'univers dans lequel on va tirer les ZAE est ainsi U = {ZAE de la région hors ZAE exhaustives} ;

- première phase : tirage au sein de U d'un échantillon s_1 de 2k ZAE, avec des probabilités proportionnelles à leur taille, et équilibrage sur le vecteur des variables d'équilibrage X. Ceci conduit à définir k_1 unités exhaustives pour cette phase, dites pseudo-exhaustives dans la suite, de taille cumulée N_1^{exh}. Les probabilités de tirage à cette phase sont alors égales à

$$\pi_i^1=\begin{cases}1 \text{ si la ZAE } i \text{ est pseudo-exhaustive}\\ (2k-k_1)\dfrac{N_i}{N-N^{exh}-N_1^{exh}} \text{ sinon}\end{cases}.$$

L'ensemble EM+EMEX ainsi défini par s_1+{ZAE exhaustives} est équilibré sur X, avec des probabilités d'inclusion proportionnelles à la taille des ZAE ;

- seconde phase : conditionnellement au tirage de s_1, tirage d'un échantillon s_2 de taille k au sein de s_1, selon le plan de sondage suivant :

- ✓ probabilités d'inclusion[7] $\pi_i^{2/1} = \begin{cases} k \dfrac{N_i}{N - N^{exh}} \text{ si la ZAE } i \text{ est pseudo - exhaustive} \\ \dfrac{k}{2k - k_1} \dfrac{N - N^{exh} - N_1^{exh}}{N - N^{exh}} \text{ sinon} \end{cases}$;
- ✓ équilibrage sur le vecteur X / π^1 .

L'EM est alors constitué de s_2 ainsi que des ZAE exhaustives. Cet échantillon-maître possède les mêmes caractéristiques qu'un EM constitué de manière indépendante : mêmes ZAE exhaustives, probabilités d'inclusion identiques et équilibrage sur le même vecteur de variables X.

Une description détaillée de la démarche retenue pour déterminer les différentes probabilités de tirage et les variables d'équilibrages de chaque phase est donnée en annexe 2.

3.2. Impact du processus de tirage en deux phases sur la qualité des estimations issues de l'échantillon maître

On s'intéresse ici à l'impact d'un tirage de l'EM en deux phases sur la qualité des estimations, par rapport à un tirage direct de l'EM. La procédure de tirage en deux phases présentée ci-dessus assurant la construction d'un EM possédant les mêmes caractéristiques que s'il avait été constitué de manière indépendante – probabilités d'inclusion identiques et équilibrage sur les mêmes variables –, elle ne devrait pas avoir de conséquence sur le biais des différentes estimations. En revanche, le fait de procéder à un sondage en deux phases rajoute un niveau d'aléa supplémentaire au processus de tirage de l'EM par rapport à un tirage direct, ce qui peut potentiellement dégrader la précision des estimations.

Afin de quantifier cet impact, nous avons procédé par simulations, en utilisant les données du recensement de 1999 :

- Le cadre théorique des échantillons Octopusse étant celui d'un sondage en plusieurs phases, l'estimateur retenu pour ces simulations est l'estimateur en expansion[8] : pour un échantillon-maître donné, tiré selon un plan de sondage à probabilités inégales π_{ZAE} , le total, pour une année de rotation *g* donnée[9], d'une variable d'intérêt Y à partir de cet échantillon de ZAE est estimé par :

$$\hat{Y}^{g}_{\text{expansion}} = \sum_{\text{ZAE grandes communes} \in \text{EM}} \frac{Y_{ZAE}}{\pi_{ZAE}} + \sum_{\text{ZAE petites communes} \in \text{EM}} \frac{Y^{g}_{ZAE}}{\frac{1}{5}\pi_{ZAE}}$$

  où l'on dispose des totaux de Y par [ZAE ⊗ groupe de rotation] Y^g_{ZAE} dans les petites communes, et des totaux de Y par ZAE dans les grandes communes Y_{ZAE} . Ces totaux sont supposés connus (on ne prend pas en compte ici l'impact du tirage de second degré des logements à l'intérieur des ZAE).

- Mille échantillons-maître indépendants sont tout d'abord tirés selon le plan de sondage détaillé au début du 1er paragraphe, et qui correspond au plan de sondage qui serait utilisé

[7] On notera que $\pi_i^{2/1}$ est une quantité dépendant de i, mais définie préalablement à tout tirage. S'agissant d'un tirage en deux phases, la probabilité conditionnelle de seconde phase vaut 0 si $i \notin s_1$ et $\pi_i^{2/1}$ si $i \in s_1$.

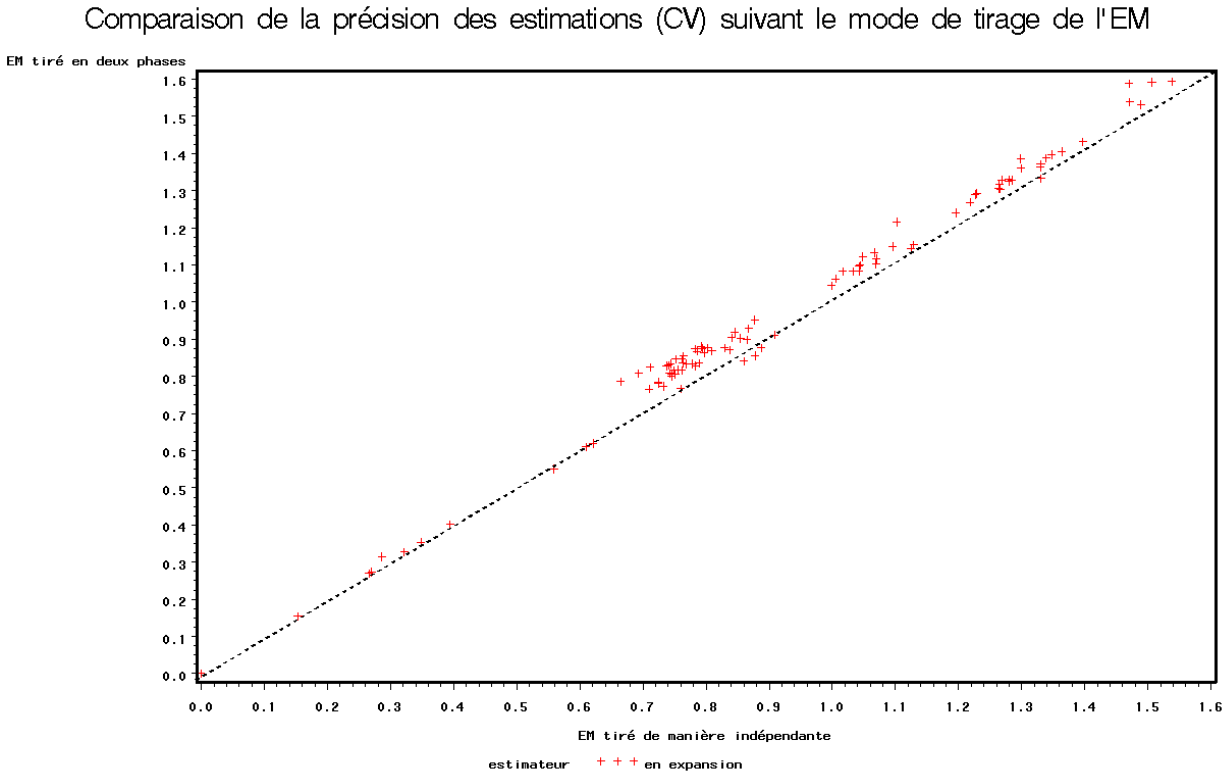
[8] Cf. annexe 4 pour une description plus détaillée du cadre théorique des échantillons d'Octopusse et de l'estimateur en expansion.

[9] Cf. annexe 3 pour une définition précise des concepts de groupe de rotation et de ZAE « petites communes » et « grandes communes ».

pour la constitution de l'échantillon-maître d'Octopusse en l'absence d'EMEX. Pour chacun de ces échantillons, on calcule, pour différentes variables d'intérêt classiques disponibles dans le RP99, et pour chaque année de rotation, les estimateurs $\hat{Y}^{g}_{expansion}$: on obtient ainsi 1000 estimations, à la distribution desquelles on s'intéresse. Sont alors déterminés les erreurs relatives empiriques de ces 1000 estimations, ainsi que leur coefficient de variation. Ces résultats constituent la situation de référence.

- On calcule ensuite, suivant le même protocole, des statistiques similaires mais cette fois-ci fondées sur mille échantillons-maîtres indépendants tirés selon la méthode de sondage en deux phases présentée ci-dessus.

Concernant l'effet du tirage en deux phases sur l'erreur relative, les simulations viennent confirmer la théorie[10] : pour toutes les variables d'intérêt, les écarts d'erreurs relatives sont systématiquement inférieurs à 0,2 %, et ce quelle que ce soit l'année de rotation considérée. La procédure de tirage en deux phases n'a donc aucune conséquence sur le biais des estimations. L'impact sur la précision s'avère quant à lui très mesuré, comme le prouve le graphique suivant :



Graphique 1 : comparaison des coefficients de variation entre la situation de référence et le tirage en deux phases de l'EM

Grille de lecture : chaque croix représente l'estimation d'une variable d'intérêt donnée. Pour les croix situées au-dessus de la bissectrice en pointillé, la précision de l'estimation obtenue à partir d'un EM tiré en deux phases est inférieure à celle obtenue à partir d'un EM tiré de manière indépendante.

Ces résultats valident donc la méthode de tirage conjointe de l'EM et de l'EMEX en deux phases.

[10] On trouvera en annexe 5 les erreurs relatives empiriques et les coefficients de variations pour la situation de référence et le tirage en plusieurs phases.

4. Généralisation : tirage en trois phases de l'EM, de l'EMEX restreint et de l'EMEX élargi

Le procédé précédent permet donc de sélectionner un EMEX de même taille que l'EM, sans détérioration de la qualité de l'EM, et tel que l'ensemble EM+EMEX possède une représentativité régionale suffisante pour assurer la production de statistiques régionales de bonne qualité.

Comme indiqué dans l'introduction de cette partie, le choix d'un EMEX de dimension égale à celle de l'EM se justifie par la pratique en matière d'extensions régionales. En particulier, ceci assure, pour la majorité des régions, une réserve de logements suffisante pour la réalisation d'enquêtes respectant le principe de disjonction des échantillons sur une durée de cinq ans minimum. Cependant, pour certaines régions particulièrement demandeuses d'extensions régionales, cette réserve risque de s'avérer insuffisante. Afin de pallier cet écueil, l'UMS a décidé de mettre en place non pas un unique EMEX, mais deux EMEX, de tailles équivalentes à celle de l'EM :

- un EMEX dit *restreint*, qui sera mobilisé pour la majeure partie des extensions régionales ;
- un EMEX dit *élargi*, qui n'interviendra que dans le cas des grosses extensions régionales pour lesquelles la base de logements EM+EMEX restreint se révèle insuffisante.

Toutefois, la création de l'EMEX élargi ne doit pas se faire au détriment de la qualité des EM et EMEX restreint. L'objectif est donc ici de déterminer une procédure de tirage, permettant de tirer conjointement trois échantillons de ZAE – l'EM, l'EMEX restreint et l'EMEX élargi – disjoints, vérifiant les propriétés suivantes :

- on veut tirer k ZAE pour l'EMEX restreint et k ZAE pour l'EMEX élargi (autant que pour l'EM) ;
- les probabilités d'inclusion pour les ZAE EM doivent être identiques à ce qu'elles seraient si l'on tirait l'EM seul ;
- les probabilités d'inclusion pour les ZAE de l'EM+EMEX restreint doivent être identiques à ce qu'elles seraient si l'on tirait l'ensemble EM+EMEX en deux phases ;
- enfin on veut assurer un équilibrage sur la variable vectorielle X, pour l'EM seul, pour l'ensemble EM+EMEX restreint et pour l'ensemble EM+EMEX restreint+EMEX élargi.

Pour ce faire, on va procéder de façon similaire à la partie précédente, en réalisant un sondage en trois phases :

- préalablement à tout tirage, on commence par retirer de la base de sondage les k_0 ZAE exhaustives définies dans le cadre de l'EM seul. Ces ZAE seront affectées automatiquement à l'EM, ainsi qu'aux ensembles EM+EMEX restreint et EM+EMEX restreint+EMEX élargi. L'univers U est donc constitué de l'ensemble des ZAE de la région privé des k_0 ZAE exhaustives ;
- toujours avant la première phase de tirage, on retire de l'univers U l'ensemble s_0 constitué des k_1 ZAE correspondant aux ZAE pseudo-exhaustives qui apparaîtraient si l'on effectuait un tirage en deux phases comme précédemment. Ces ZAE seront affectées automatiquement à l'EM+EMEX restreint ainsi qu'à l'EM+EMEX restreint+EMEX élargi. Ceci définit un nouvel univers de tirage $\widetilde{U} = U \setminus s_0$;
- première phase : tirage d'un échantillon $\widetilde{s}_1$, de taille 3k-k_1, avec des probabilités d'inclusion π_i^1. L'ensemble $s_1 = s_0 + \widetilde{s}_1$ représentera l'ensemble EM+EMEX restreint+EMEX élargi ;
- deuxième phase : conditionnellement au tirage de $\widetilde{s}_1$, tirage d'un échantillon $\widetilde{s}_2$ au sein de $\widetilde{s}_1$, de taille 2k- k_1, avec des probabilités d'inclusion $\pi_i^{2/1}$. L'ensemble $s_2 = s_0 + \widetilde{s}_2$ représentera l'EM+EMEX restreint.

La probabilité d'inclusion finale d'une unité *i* dans s_2 est alors $\boxed{\pi_i^2 = \begin{cases} 1 \text{ si } i \in s_0 \\ \pi_i^1 \pi_i^{2/1} \text{ si } i \notin s_0 \end{cases}}$

- troisième phase : conditionnellement au tirage de s_2, tirage d'un échantillon s_3 au sein de s_2, de taille k, avec des probabilités d'inclusion $\pi_i^{3/2}$, qui représentera l'EM.

La probabilité d'inclusion finale d'une unité *i* dans s_3 est alors $\boxed{\pi_i^3 = \pi_i^2 \pi_i^{3/2} = \begin{cases} \pi_i^{3/2} \text{ si } i \in s_0 \\ \pi_i^1 \pi_i^{2/1} \pi_i^{3/2} \text{ si } i \notin s_0 \end{cases}}$

(i) Respect des conditions d'équilibrage

a) La condition d'équilibrage sur la variable X, pour l'ensemble EM+EMEX restreint+EMEX élargi, s'écrit $\sum_{i \in s_1} \frac{X_i}{\pi_i^1} = \sum_{i \in U} X_i \Leftrightarrow \boxed{\sum_{i \in \tilde{s}_1} \frac{X_i}{\pi_i^1} = \sum_{i \in \tilde{U}} X_i}$ **(1)**.

b) La condition d'équilibrage sur la variable X, pour l'EM+EMEX restreint, s'écrit $\sum_{i \in s_2} \frac{X_i}{\pi_i^2} = \sum_{i \in U} X_i \Leftrightarrow \sum_{i \in \tilde{s}_2} \frac{X_i}{\pi_i^2} = \sum_{i \in \tilde{U}} X_i$, soit $\boxed{\sum_{i \in \tilde{s}_2} \frac{X_i}{\pi_i^1 \pi_i^{2/1}} = \sum_{i \in \tilde{U}} X_i}$ **(2)**.

Les égalités **(1)** et **(2)** entraînent $\sum_{i \in \tilde{s}_2} \frac{X_i}{\pi_i^1 \pi_i^{2/1}} = \sum_{i \in \tilde{s}_1} \frac{X_i}{\pi_i^1}$ **(3)**. Inversement, l'égalité **(3)** jointe à l'égalité **(1)** entraîne l'égalité **(2).**

Or si l'on récrit l'égalité **(3)** sous la forme $\boxed{\sum_{i \in \tilde{s}_2} \frac{X_i / \pi_i^1}{\pi_i^{2/1}} = \sum_{i \in \tilde{s}_1} (X_i / \pi_i^1)}$ **(3 bis)**, on voit que cette dernière s'interprète comme une condition d'équilibrage sur la variable X / π^1 lors du tirage conditionnel de l'échantillon $\tilde{s}_2$ au sein de $\tilde{s}_1$, $\tilde{s}_1$ jouant alors le rôle d' « univers conditionnel ».

c) La condition d'équilibrage sur la variable X, pour l'EM seul, s'écrit $\sum_{i \in s_3} \frac{X_i}{\pi_i^3} = \sum_{i \in U} X_i$, soit

$\boxed{\sum_{i \in s_3} \frac{X_i}{\pi_i^2 \pi_i^{3/2}} = \sum_{i \in U} X_i}$ **(4)**.

Les égalités **(4)** et **(2)** entraînent $\boxed{\sum_{i \in s_3} \frac{X_i}{\pi_i^2 \pi_i^{3/2}} = \sum_{i \in s_2} \frac{X_i}{\pi_i^2}}$ **(5)**. Inversement, l'égalité **(5)** jointe à l'égalité **(2)** entraîne l'égalité **(4).**

Or si l'on récrit l'égalité **(5)** sous la forme $\boxed{\sum_{i \in s_3} \frac{X_i / \pi_i^2}{\pi_i^{3/2}} = \sum_{i \in s_2} (X_i / \pi_i^2)}$ **(5 bis)**, on voit que cette dernière s'interprète comme une condition d'équilibrage sur la variable X / π^2 lors du tirage conditionnel de l'échantillon s_3 au sein de s_2, s_2 jouant alors le rôle d' « univers conditionnel ».

d) Par suite :

- l'équilibrage de l'ensemble EM+EMEX restreint+EMEX élargi sur la variable X est obtenu en équilibrant l'échantillon $\tilde{s}_1$ sur X lors de la première phase ;
- pour que l'ensemble EM+EMEX restreint soit aussi équilibré sur la variable X, il faut et il suffit, lors du tirage de deuxième phase, d'équilibrer l'échantillon $\tilde{s}_2$ sur la variable X/π^1 ;
- enfin, pour assurer un équilibrage sur la variable X de l'EM seul, il convient de tirer l'échantillon s_3 en équilibrant sur la variable X/π^2.

On notera qu'à ce stade les valeurs des probabilités d'inclusion importent peu – certaines peuvent d'ailleurs être égales à 1 –, de même que les tailles des échantillons considérés.

(ii) Respect des probabilités d'inclusion et des tailles d'échantillon

On souhaite respecter les probabilités d'inclusion :

- <u>pour l'EM seul :</u> on veut que la probabilité finale d'une ZAE dans l'échantillon s_3 soit identique à celle que l'on aurait si l'on tirait l'EM seul, soit $\pi_i^3 = k\frac{N_i}{N-N^{exh}}$. Ceci entraîne $\boxed{\pi_i^2\pi_i^{3/2} = k\frac{N_i}{N-N^{exh}}}$ **(α)** ;
- <u>pour l'ensemble EM+EMEX restreint :</u> on veut que la probabilité finale d'une ZAE dans s_2 soit identique à celle que l'on aurait si l'on tirait l'EMEX en deux phases, soit $\pi_i^2 = \begin{cases} 1 \text{ si } i \in s_0 \\ (2k-k_1)\frac{N_i}{N-N^{exh}-N_1^{exh}} \text{ si } i \notin s_0 \end{cases}$. D'où $\boxed{\pi_i^1\pi_i^{2/1} = (2k-k_1)\frac{N_i}{N-N^{exh}-N_1^{exh}}}$ **(β)**.

Par ailleurs, on veut tirer trois échantillons de tailles fixes :

- $\tilde{s}_1$ de taille 3k-k_1, ce qui impose la condition $\boxed{\sum_{i\in\tilde{U}}\pi_i^1 = 3k-k_1}$ **(γ)** ;
- conditionnellement à $\tilde{s}_1$ et au sein de ce dernier, $\tilde{s}_2$ de taille 2k-k_1, ce qui donne $\boxed{\sum_{i\in\tilde{s}_1}\pi_i^{2/1} = 2k-k_1}$ **(δ)** ;
- enfin, conditionnellement à $s_2 = s_0 + \tilde{s}_2$ et au sein de ce dernier, s_3 de taille k, ce qui implique $\boxed{\sum_{i\in s_2}\pi_i^{3/2} = k}$ **(ε)**.

Il s'agit donc de déterminer des probabilités π_i^1, $\pi_i^{2/1}$ et $\pi_i^{3/2}$ appartenant à [0,1] et vérifiant les conditions **(α)** à **(ε)**.

(α) se réécrit $\begin{cases} \pi_i^{3/2} = k\frac{N_i}{N-N^{exh}} \text{ si } i \in s_0 \quad (\alpha_1) \\ \pi_i^1\pi_i^{2/1}\pi_i^{3/2} = k\frac{N_i}{N-N^{exh}} \text{ si } i \notin s_0 \quad (\alpha_2) \end{cases}$.

Or, pour $i \notin s_0$, **(α_2) + (β)** implique : $\pi_i^{3/2} = \frac{k}{2k-k_1} \frac{N - N^{exh} - N_1^{exh}}{N - N^{exh}}$. Les probabilités $\pi_i^{3/2}$ sont donc entièrement déterminées :

$$\boxed{\pi_i^{3/2} = \begin{cases} k \frac{N_i}{N - N^{exh}} \text{ si } i \in s_0 \\ \frac{k}{2k-k_1} \frac{N - N^{exh} - N_1^{exh}}{N - N^{exh}} \text{ si } i \notin s_0 \end{cases}}$$

On vérifie alors que ces probabilités satisfont bien la condition (ε) :

$$\begin{aligned} \sum_{i \in s_2} \pi_i^{3/2} &= \sum_{i \in s_0} k \frac{N_i}{N - N^{exh}} + \sum_{i \in \tilde{s}_2} \frac{k}{2k-k_1} \frac{N - N^{exh} - N_1^{exh}}{N - N^{exh}} \\ &= k \frac{N_1^{exh}}{N - N^{exh}} + \frac{k}{2k-k_1} \frac{N - N^{exh} - N_1^{exh}}{N - N^{exh}} \underbrace{\sum_{i \in \tilde{s}_2} 1}_{2k-k_1} = k \left[\frac{N_1^{exh} + N - N^{exh} - N_1^{exh}}{N - N^{exh}} \right] = k \end{aligned}$$

Il reste donc à trouver π_i^1 et $\pi_i^{2/1}$ appartenant à [0,1] et vérifiant les conditions **(γ)**, **(δ)** et **(β)**.

De **(β)**, on tire $\pi_i^{2/1} = \frac{2k-k_1}{\pi_i^1} \frac{N_i}{N - N^{exh} - N_1^{exh}}$ **(β bis)**, ce qui injecté dans **(δ)** conduit à $\sum_{i \in \tilde{s}_1} \frac{N_i}{\pi_i^1} = \underbrace{N - N^{exh} - N_1^{exh}}_{\sum_{i \in \tilde{U}} N_i}$. Or cette relation est assurée par l'équilibrage sur la taille effectué lors du tirage de première phase.

Ainsi, les probabilités $\pi_i^{2/1}$ étant déterminées par l'équation **(β Bis)** et la condition **(δ)** étant assurée par l'équilibrage réalisé lors du tirage de première phase, il ne reste plus qu'à déterminer les probabilités π_i^1 satisfaisant l'équation **(γ)** ainsi que les conditions $\pi_i^1 \in [0,1]$ et $\pi_i^{2/1} \in [0,1]$. En reportant la condition $\pi_i^{2/1} \in [0;1]$ dans l'équation **(β Bis)**, on obtient :

$$\boxed{2k - k_1 \frac{N_i}{N - N^{exh} - N_1^{exh}} \leq \pi_i^1 \leq 1} \quad (\eta)$$

Cette équation est bien admissible, puisqu'on a mis de coté les ZAE pseudo-exhaustives s_0, donc les quantités $2k - k_1 \frac{N_i}{N - N^{exh} - N_1^{exh}}$ sont bien inférieures ou égales à 1. En outre, cette équation est compatible avec **(γ)** : en effet, d'après **(β)** on a $\sum_{i \in \tilde{U}} 2k - k_1 \frac{N_i}{N - N^{exh} - N_1^{exh}} \leq \sum_{i \in \tilde{U}} \pi_i^1 = 3k - k_1$, ce qui est bien vérifié puisque $\sum_{i \in \tilde{U}} N_i = N - N^{exh} - N_1^{exh}$ par définition de $\tilde{U}$.

Pour déterminer les π_i^1, on va ici encore effectuer un tirage de $3k\text{-}k_1$ unités avec des probabilités proportionnelles à leur taille. Ceci va conduire à définir de nouvelles unités exhaustives pour ce tirage

de première phase, au nombre de k_2 et de taille cumulée N_2^{exh}. En pratique, on procède donc comme suit :

- ✓ Pour les k_2 unités nouvellement exhaustives, on prend $\pi_i^1 = 1$. Ces unités seront retenues d'office pour la sélection de l'EM+EMEX restreint+EMEX élargi. En revanche, elles n'appartiendront pas nécessairement à l'EM+EMEX restreint, puisqu'elles seront tirées dans $\tilde{s}_2$ avec une probabilité $\pi_i^{2/1} = 2k - k_1 \frac{N_i}{N - N^{exh} - N_1^{exh}}$ d'après l'équation **(β Bis)**.

- ✓ On construit alors un nouvel univers $\tilde{\tilde{U}}$ (de taille $N - N^{exh} - N_1^{exh} - N_2^{exh}$) en retranchant de $\tilde{U}$ ces unités. On tire ensuite dans $\tilde{\tilde{U}}$ un nombre $3k\text{-}k_1\text{-}k_2$ d'unités avec des probabilités proportionnelles à leur taille $\boxed{\pi_i^1 = (3k - k_1 - k_2) \frac{N_i}{N - N^{exh} - N_1^{exh} - N_2^{exh}}}$, ce qui assure la condition **(γ)**. En deuxième phase, ces unités seront tirées, d'après l'équation **(β Bis)**, avec une probabilité $\pi_i^{2/1} = \frac{2k - k_1}{3k - k_1 - k_2} \frac{N - N^{exh} - N_1^{exh} - N_2^{exh}}{N - N^{exh} - N_1^{exh}}$.

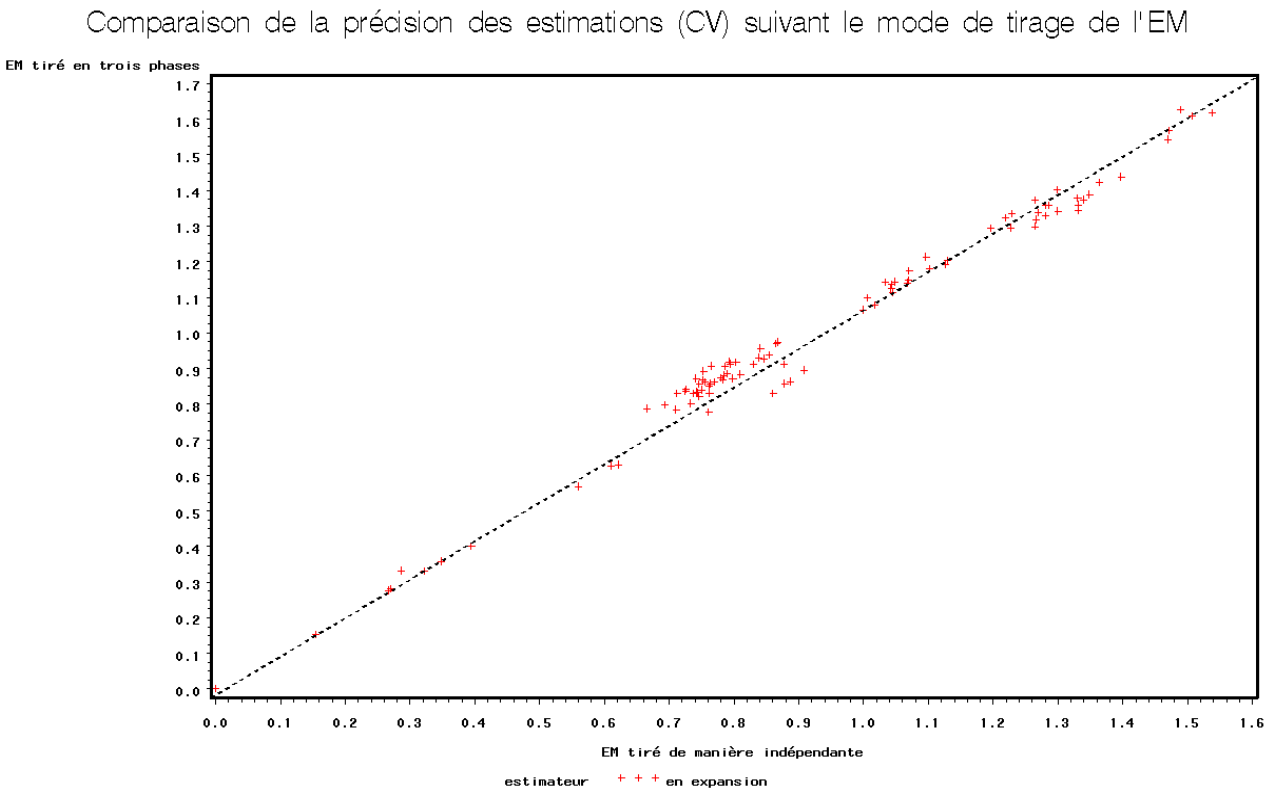
Les caractéristiques du plan de sondage en trois phases ayant été déterminées, il reste à vérifier empiriquement que ce mode de tirage ne dégrade pas la qualité des estimations, que ce soit pour l'EM seul ou pour l'ensemble EM+EMEX restreint. À cette fin, des simulations similaires à celles de la partie précédente ont été menées, et amènent aux conclusions suivantes :

- tout comme dans le cas d'un tirage en deux phases, la procédure de tirage en trois phases n'a aucun impact sur le biais des estimations ;

- l'impact en terme de précision concernant les estimations réalisées à partir de l'EM est sensiblement identique à ce que l'on obtient en effectuant un tirage indépendant, comme le prouve le graphique 2 ;

- on observe des résultats similaires pour ce qui est de la précision des estimations réalisées à partir de l'EM+EMEX restreint par rapport à l'EM+EMEX tiré en deux phases.

Au final, ce procédé de tirage en trois phases permet d'obtenir un échantillon-maître présentant des caractéristiques statistiques – en matière de probabilités d'inclusion, d'équilibrage et de qualité des estimations – sensiblement identiques à celles que l'on obtiendrait en le constituant de manière indépendante, ainsi que des échantillons-maîtres pour les extensions régionales restreint et élargi vérifiant les propriétés voulues. **En conséquence, cette méthode a été retenue pour le tirage effectif des EM, EMEX restreint et EMEX élargi définitifs qui a eu lieu fin septembre 2007.**



Graphique 2 : comparaison des coefficients de variation entre tirage en 2 phases tirage en 3 phases de l'EM

Grille de lecture : chaque croix représente l'estimation d'une variable d'intérêt donnée. Pour les croix situées au-dessus de la bissectrice en pointillé, la précision de l'estimation obtenue à partir d'un EM tiré en trois phases est inférieure à celle obtenue à partir d'un EM tiré de manière indépendante.

Conclusion

Les méthodes décrites dans ce papier d'échantillonnage équilibré en plusieurs phases permettent de résoudre de façon appropriée la question du tirage simultané des ZAE de l'EM et de l'EMEX, tout en imposant des probabilités d'inclusion finales fixées, ainsi que des conditions d'équilibrage à la fois au niveau de l'ensemble EM + EMEX et de l'EM seul. Elles permettent notamment de prendre en compte la difficulté liée au fait que certaines unités doivent être retenues d'office aux différentes phases du tirage, compte tenu des valeurs supérieures à 1 que prendraient des probabilités proportionnelles à la taille. Elles ont permis également de s'affranchir des difficultés théoriques et pratiques rencontrées en 2002, où le tirage de l'EMEX avait été réalisé postérieurement et conditionnellement à celui de l'EM.

Ces méthodes ont été effectivement mises en œuvre lors du tirage de l'ensemble des ZAE fin 2007.

Comme c'est le cas pour les ZAE de l'EM[11], les ZAE retenues pour l'EMEX en cas d'extension régionale seront assujetties à un processus de calage après tirage, visant à substituer à leurs poids de sondage de nouveaux poids calés utilisant différentes variables socio-démographiques, dans l'optique d'assurer une meilleure « représentativité » des échantillons ainsi constitués et de diminuer la variance des estimateurs. Le cadre dans lequel se déroulera cette opération sera toutefois plus compliqué que celui du seul EM. En effet, d'une part, on ne sait pas à l'avance quelles régions demanderont des extensions et de quelle ampleur elle seront (devant mobiliser l'EMEX restreint ou élargi) ; d'autre part, ces régions peuvent différer d'une enquête à l'autre.

Ainsi, pour chaque enquête bénéficiant d'extensions régionales, on mobilisera un EMEX pour les régions considérées et l'EM pour les autres. Le calage sera effectué séparément pour chaque région à extension à partir de l'ensemble des ZAE tirées pour l'EMEX retenu, en utilisant des totaux de

[11] Sébastien Faivre. *Le projet Octopusse de nouvel échantillon-maître de l'Insee.* JMS 2009

calage régionaux. Pour les autres régions, l'ensemble des ZAE de l'EM seront calées globalement sur des totaux relatifs à l'univers considéré (France entière moins les régions à extensions).

Un chantier ultérieur à développer sera celui du calcul de précision associé à ces différentes procédures, tant au niveau du 1er degré de tirage (ZAE) qu'à celui du tirage final des logements.

Bibliographie

- Jean-Claude Deville, Yves Tillé. *Efficient balanced sampling : the Cube method*. Octobre 2003.
- Guillaume Chauvet, Yves Tillé. *De nouvelles macros SAS d'échantillonnage équilibré*. Insee, Actes des Journées Méthodologie Statistique 2005.
- Sébastien Faivre. *Le projet Octopusse de nouvel échantillon-maître de l'Insee.* JMS 2009.
- Marc Christine, Laurent Wilms. *Problèmes théoriques et pratiques de la construction de l'EMEX*. Insee, Actes des Journées Méthodologie Statistique 2002.
- Fabien Guggemos. *Simulations de tirages de zones d'action enquêteurs pour les enquêtes ménages de l'Insee*. JMS 2009.

Annexes

Annexe 1
Méthode de tirage de deux échantillons disjoints, équilibrés sur la même variable, avec les mêmes probabilités d'inclusion, mise en œuvre pour la constitution des groupes de rotation du recensement.

Annexe 2
Méthode de tirage en deux phases de l'EM et de l'EMEX.

Annexe 3
Présentation succincte du recensement rénové de la population et des Zones d'Action Enquêteurs.

Annexe 4
Le cadre théorique des échantillons Octopusse et l'estimateur en expansion.

Annexe 5
Erreur relative et précision des estimations pour des échantillons maîtres tirés seuls, en deux phases et en trois phases.

ANNEXE 1 : Méthode de tirage de deux échantillons disjoints équilibrés sur la même variable, avec les mêmes probabilités d'inclusion, mise en œuvre pour la constitution des groupes de rotation du recensement.

Tiré de l'article « De nouvelles macros SAS d'échantillonnage équilibré » de Guillaume CHAUVET et Yves TILLE.

On tire tout d'abord un premier échantillon s_1 avec les probabilités $\pi_i^1 = \pi_i$, équilibré sur les variables X et $\frac{X}{1-\pi}$: la première variable assure l'équilibrage de s_1 sur X, tandis que la seconde variable permet d'obtenir un équilibrage du complémentaire de s_1 sur X.

En effet, soit $\bar{s}_1$ le complémentaire de s_1 dans la population U. Les probabilités d'inclusion dans cet échantillon sont alors $\bar{\pi}_i = P(i \notin s_1) = 1 - \pi_i$, aussi l'échantillon $\bar{s}_1$ est dit équilibré sur la variable X si $\sum_{i \in \bar{s}_1} \frac{X_i}{1-\pi_i} = \sum_{i \in U} X_i$.

Or, on a pour s_1 par définition :

$$\begin{cases} \sum_{i \in s_1} \frac{X_i}{\pi_i} = \sum_{i \in U} X_i \quad (\alpha) \\ \sum_{i \in s_1} \frac{1}{\pi_i} \frac{X_i}{1-\pi_i} = \sum_{i \in U} \frac{X_i}{1-\pi_i} \quad (\beta) \end{cases}$$

(α) traduit l'équilibrage de l'échantillon s_1 sur la variable X voulue, et (β) sur la variable $\frac{X}{1-\pi}$.

Pour ce qui est du complémentaire, on a :

$$\sum_{i \in \bar{s}_1} \frac{X_i}{1-\pi_i} = \sum_{i \in U} \frac{X_i}{1-\pi_i} - \sum_{i \in s_1} \frac{X_i}{1-\pi_i} \underbrace{=}_{\beta} \sum_{i \in s_1} \frac{1}{\pi_i} \frac{X_i}{1-\pi_i} - \sum_{i \in s_1} \frac{X_i}{1-\pi_i} = \sum_{i \in s_1} \frac{X_i}{\pi_i} \underbrace{=}_{\alpha} \sum_{i \in U} X_i$$

et $\bar{s}_1$ est donc bien équilibré sur X lui aussi.

On sélectionne ensuite le second échantillon s_2 dans le complémentaire de s_1, avec les probabilités d'inclusion conditionnelles à s_1 $\pi_i^{2/s_1} = \frac{\pi_i}{1-\pi_i} 1_{i \in \bar{s}_1}$ et équilibré sur la variable $\frac{X}{1-\pi}$. L'échantillon ainsi obtenu est bien équilibré sur la variable X avec des probabilités d'inclusion π_i :

✓ Pour les probabilités d'inclusion, on a :

$$P(i \in s_2) = P(i \in s_2 \cap (i \in s_1 \cup i \in \bar{s}_1)) = \underbrace{P(i \in s_2 \cap i \in s_1)}_{0} + P(i \in s_2 \cap i \in \bar{s}_1)$$

$$= P(i \in s_2 \cap i \in \bar{s}_1) = E(1_{i \in \bar{s}_1} 1_{i \in s_2})$$

$$\text{Or } E(1_{i \in \bar{s}_1} 1_{i \in s_2}) = E(E(1_{i \in \bar{s}_1} 1_{i \in s_2} | s_1)) = E(1_{i \in \bar{s}_1} \underbrace{E(1_{i \in s_2} | s_1)}_{\pi_i^{2/s_1}}) = \frac{\pi_i}{1-\pi_i} \underbrace{E(1_{i \in \bar{s}_1})}_{1-\pi_i} = \frac{\pi_i}{1-\pi_i}(1-\pi_i)$$

$$\text{On a donc bien } P(i \in s_2) = \frac{\pi_i}{1-\pi_i}(1-\pi_i) = \pi_i.$$

✓ En ce qui concerne l'équilibrage, on a par définition $\sum_{i\in s_2}\frac{\frac{X_i}{1-\pi_i}}{\frac{\pi_i}{1-\pi_i}}=\sum_{i\in \bar{s}_1}\frac{X_i}{1-\pi_i}=\sum_{i\in U}X_i$, la dernière égalité venant du fait que $\bar{s}_1$ est équilibré sur X. Or, $\sum_{i\in s_2}\frac{\frac{X_i}{1-\pi_i}}{\frac{\pi_i}{1-\pi_i}}=\sum_{i\in s_2}\frac{X_i}{\pi_i}$. On a donc $\sum_{i\in s_2}\frac{X_i}{\pi_i}=\sum_{i\in U}X_i$, et l'échantillon s_2 est donc bien équilibré sur X.

Au final, on obtient bien deux échantillons disjoints, équilibrés sur la même variable vectorielle X et avec des probabilités d'inclusion π_i.

ANNEXE 2 : Méthode de tirage en deux phases de l'EM et de l'EMEX.

On se place dans une région donnée. On a déterminé antérieurement, sur la base de paramètres standards, le nombre de ZAE de l'EM à tirer dans la région ainsi que les ZAE exhaustives.

On veut tirer simultanément les ZAE de l'EM et de l'EMEX sous les contraintes suivantes :

- on veut tirer k ZAE-EMEX (autant que de ZAE EM) ;
- on veut des probabilités d'inclusion pour les ZAE EM identiques à ce qu'elles seraient si l'on tirait l'EM seul ;
- on veut assurer un équilibrage sur une variable (vectorielle) X, pour l'EM seul et pour l'ensemble EM+EMEX.

Pour cela, on va réaliser un **échantillonnage en deux phases**.

- préalablement à tout tirage, on commence par retirer de la base de sondage les k_0 ZAE exhaustives définies dans le cadre de l'EM seul. Ces ZAE seront affectées automatiquement à l'EM, ainsi qu'à l'ensemble EM+EMEX ;

- première phase : tirage d'un échantillon s_1, de taille 2 k, avec des probabilités d'inclusion π_i^1, qui représentera l'ensemble EM+EMEX ;

- seconde phase : conditionnellement au tirage de s_1, tirage d'un échantillon s_2 au sein de s_1, de taille k, avec des probabilités d'inclusion $\pi_i^{2/1}$, qui représentera l'EM seul.

La probabilité d'inclusion finale d'une unité *i* dans s_2 est alors $\boxed{\pi_i = \pi_i^1 \pi_i^{2/1}}$:

$$\pi_i = P(i \in s_2) = E(1_{i\in s_2}) = E_1(\underbrace{E_{2/1}(1_{i\in s_2})}_{\pi_i^{2/1} 1_{i\in s_1}}) = E_1(\pi_i^{2/1} 1_{i\in s_1}) = \pi_i^{2/1} \underbrace{E_1(1_{i\in s_1})}_{\pi_i^1} = \pi_i^{2/1}\pi_i^1$$

(i) Respect des conditions d'équilibrage

La condition d'équilibrage sur la variable X, pour l'ensemble EM+EMEX, s'écrit $\boxed{\sum_{i\in s_1} \frac{X_i}{\pi_i^1} = \sum_{i\in U} X_i}$ **(1)** où U est l'univers (ensemble des ZAE de la région privé des k_0 ZAE exhaustives).

La condition d'équilibrage sur la variable X, pour l'EM seul, s'écrit $\sum_{i\in s_2} \frac{X_i}{\pi_i} = \sum_{i\in U} X_i$, soit

$$\boxed{\sum_{i\in s_2} \frac{X_i}{\pi_i^1 \pi_i^{2/1}} = \sum_{i\in U} X_i} \quad \textbf{(2)}.$$

Les égalités **(1)** et **(2)** entraînent $\sum_{i\in s_2} \frac{X_i}{\pi_i^1 \pi_i^{2/1}} = \sum_{i\in s_1} \frac{X_i}{\pi_i^1}$ **(3)**. Inversement, l'égalité **(3)** jointe à l'égalité **(1)** entraîne l'égalité **(2).**

Or si l'on récrit l'égalité **(3)** sous la forme $\boxed{\sum_{i\in s_2} \frac{X_i/\pi_i^1}{\pi_i^{2/1}} = \sum_{i\in s_1} (X_i/\pi_i^1)}$ **(3 bis)**, on voit que cette dernière s'interprète comme une condition d'équilibrage sur la variable X/π^1 lors du tirage conditionnel de l'échantillon s_2 au sein de s_1, s_1 jouant alors le rôle d'«univers conditionnel».

Par suite :

Pour assurer un équilibrage sur une variable X, pour l'EM seul et pour l'ensemble EM+EMEX, il suffit de tirer un premier échantillon s_1 équilibré sur X, avec des probabilités d'inclusion π_i^1, qui représentera l'ensemble EM+EMEX, puis, au sein de s_1 et conditionnellement au tirage de ce dernier, un échantillon s_2, **équilibré sur la variable** X/π^1, avec des probabilités d'inclusion $\pi_i^{2/1}$, qui représentera l'EM seul.

On notera qu'à ce stade les valeurs des probabilités d'inclusion importent peu – certaines peuvent d'ailleurs être égales à 1 –, de même que les tailles des échantillons considérés.

(ii) Respect des probabilités d'inclusion et des tailles d'échantillon

On veut que la probabilité finale d'une ZAE dans l'échantillon s_2 soit identique à celle que l'on aurait si l'on tirait l'EM seul, soit $\pi_i = k\frac{N_i}{N-N^{exh}}$. Ceci entraîne $\boxed{\pi_i^1\pi_i^{2/1} = k\frac{N_i}{N-N^{exh}}}$ **(4)**.

Par ailleurs, on veut tirer deux échantillons de tailles fixes :

- ✓ s_1 de taille 2k, ce qui impose la condition $\boxed{\sum_{i\in U}\pi_i^1 = 2k}$ **(5)**
- ✓ conditionnellement à s_1 et au sein de ce dernier, s_2 de taille k, ce qui donne $\boxed{\sum_{i\in s_1}\pi_i^{2/1} = k}$ **(6)**.

Il s'agit donc de déterminer des probabilités π_i^1 et $\pi_i^{2/1}$ appartenant à [0,1] et vérifiant les conditions **(4)** à **(6)**.

Or l'équation **(4)** permet de tirer $\boxed{\pi_i^{2/1} = k\frac{1}{\pi_i^1}\frac{N_i}{N-N^{exh}}}$ **(4 bis)**. En injectant cette valeur dans l'équation **(6)**, on obtient $\sum_{i\in s_1} k\frac{1}{\pi_i^1}\frac{N_i}{N-N^{exh}} = k$, soit :

$$\boxed{\sum_{i\in s_1}\frac{N_i}{\pi_i^1} = N-N^{exh}} \quad \textbf{(7)}$$

Cette équation s'interprète comme une **condition d'équilibrage sur la taille des unités lors du tirage de l'échantillon s_1**, le terme de droite étant précisément la taille de l'univers dans lequel on tire s_1. Or dans le problème qui nous intéresse, le nombre de résidences principales des ZAE N_i fait justement partie du vecteur des variables d'équilibrage X. La condition **(7)** est donc automatiquement assurée par le respect de la condition **(1)**.

Ainsi, les probabilités $\pi_i^{2/1}$ étant déterminées par l'équation **(4 Bis)** et la condition **(6)** étant assurée par l'équation d'équilibrage **(1)**, il ne reste plus qu'à déterminer les probabilités π_i^1 satisfaisant l'équation **(5)** ainsi que les conditions $\pi_i^1 \in [0,1]$ et $\pi_i^{2/1} \in [0,1]$. En reportant la condition $\pi_i^{2/1} \in [0,1]$ dans l'équation **(4 Bis)**, on obtient :

$$\boxed{k\frac{N_i}{N-N^{exh}} \le \pi_i^1 \le 1} \quad \textbf{(8)}$$

Notons tout d'abord que cette équation **(8)** est bien admissible. En effet, puisqu'on a mis de côté au départ les ZAE exhaustives, on est assuré que les quantités $k\frac{N_i}{N-N^{exh}}$, qui représentent les probabilités d'inclusion des ZAE non exhaustives, si on ne tirait que le seul EM, sont bien inférieures ou égales à 1.

Par ailleurs, cette équation **(8)** est compatible avec la condition **(5)**. En effet, d'après **(8)** on a $\sum_{i\in U} k\frac{N_i}{N-N^{exh}} \le \sum_{i\in U}\pi_i^1 = 2k$, ce qui est bien vérifié puisque $\sum_{i\in U} N_i = N - N^{exh}$ par définition de U.

Pour déterminer les π_i^1, une solution naturelle consisterait à prendre $\pi_i^1 = 2k\frac{N_i}{N-N^{exh}}$: cela correspondrait au tirage de 2k unités avec des probabilités proportionnelles à leur taille. Toutefois, il se peut que, pour certaines unités, on ait : $2k\frac{N_i}{N-N^{exh}} > 1$, puisque, par construction, on a seulement $k\frac{N_i}{N-N^{exh}} \le 1$.

Ceci va conduire à définir de nouvelles unités exhaustives pour ce tirage de première phase, dites **unités pseudo–exhaustives** dans la suite. En notant k_1 le nombre de ces unités, et N_1^{exh} la taille cumulée de ces unités, on procède donc comme suit :

- ✓ Pour les k_1 unités pseudo exhaustives, on prend $\pi_i^1 = 1$. Ces unités seront retenues d'office pour la sélection de l'EM+EMEX. En revanche, elles n'appartiendront pas nécessairement à l'EM, puisqu'elles seront tirées dans s_2 avec une probabilité $\pi_i^{2/1} = k\frac{N_i}{N-N^{exh}}$ d'après l'équation **(4 Bis)**.
- ✓ On construit alors un nouvel univers U_1 (de taille $N - N^{exh} - N_1^{exh}$) en retranchant de U ces unités pseudo-exhaustives. On tire ensuite dans U_1 un nombre $2k\text{-}k_1$ d'unités avec des probabilités proportionnelles à leur taille $\boxed{\pi_i^1 = (2k-k_1)\frac{N_i}{N-N^{exh}-N_1^{exh}}}$, ce qui assure la condition **(5)**.

En seconde phase, ces unités seront tirées, d'après l'équation **(4 Bis)**, avec une probabilité $\pi_i^{2/1} = \frac{k}{2k-k_1}\frac{N-N^{exh}-N_1^{exh}}{N-N^{exh}}$.

ANNEXE 3 : Présentation succincte du recensement rénové de la population et des Zones d'Action Enquêteurs.

(i) Le recensement rénové de la population

Dans le courant de la précédente décennie, l'Insee a décidé de passer du principe de recensement exhaustif de la population française effectué à intervalles de temps quasi-réguliers (7 à 9 ans) à un nouveau mode de recensement par sondage annuel. Celui-ci a définitivement été mis en place en janvier 2004.

Désormais, les communes de moins de 10000 habitants, ou « petites communes », sont recensées exhaustivement tous les cinq ans par roulement : pour cela, cinq groupes de rotation ont été définis aléatoirement, dans lesquels ont été réparties ces petites communes.

Pour ce qui est des communes comprenant 10000 habitants ou plus, ou « grandes communes », elles font l'objet d'une enquête de recensement plus complexe, par sondage chaque année au taux moyen de 8%. Les adresses de ces communes sont réparties aléatoirement entre cinq groupes de rotation disjoints et successivement enquêtées par le recensement pendant cinq ans au taux moyen de 40%. Plus précisément, le traitement des logements diffère selon que le logement appartient à une « grande adresse[12] », une « adresse neuve » ou une « autre adresse » :

- lors de la première phase du RP de construction des groupes de rotation, les grandes adresses et les adresses neuves ont été affectées, pour la majorité des grandes communes, de manière déterministe à un des cinq groupes tandis que les autres adresses étaient réparties de manière aléatoire entre ces groupes ;

- une fois les groupes de rotation constitués, la seconde phase du tirage sélectionne, pour un groupe de rotation donné, les adresses qui seront enquêtées par le recensement. Les grandes adresses et les adresses neuves sont enquêtées exhaustivement, puis les autres adresses sont échantillonnées de telle sorte que l'échantillon total de logements (y compris les grandes adresses et les adresses neuves enquêtées exhaustivement) représente 40% des logements du groupe de rotation enquêté.

Ce changement de méthodologie, s'il induit certes la perte du caractère exhaustif du recensement, offre en contrepartie de nombreux avantages, dont le principal réside dans la fraîcheur des données recueillies : avec cette méthode de collecte, il y aura en effet chaque année un recensement exhaustif dans environ 7000 petites communes et une enquête de recensement par sondage dans environ 900 grandes communes. Au bout de cinq ans, c'est-à-dire à l'issue de l'année 2008, l'ensemble des petites communes seront recensées ainsi que 40% des logements des grandes communes, ce qui permettra notamment d'établir chaque année, à partir de fin 2008, des chiffres de population légale basés sur les cinq dernières collectes annuelles.

(ii) Des Zones d'Action Enquêteurs intégrant l'aspect rotatif du nouveau recensement

Afin d'assurer un tirage systématique des échantillons d'enquêtes auprès des ménages dans la partie la plus récente de la base de sondage, la construction des ZAE a reposé sur les principes suivants :

- les ZAE sont des zones fixes pour une durée prévue de 10 ans[13], et ce afin de pouvoir leur affecter un enquêteur stable dans le temps et localisé à proximité ;
- toute grande commune constitue à elle seule une ZAE, appelée ZAE grande commune ou ZAEGC ;
- pour permettre de tirer chaque année les échantillons d'enquête dans la fraction recensée l'année précédente, les ZAE petites communes (ZAEPC) comportent des communes de chaque groupe de rotation ;
- les ZAE respectent les frontières régionales, d'une part, afin de les rattacher sans ambiguïté à l'une des directions régionales (DR) de l'Insee, et d'autre part, pour pouvoir tirer des échantillons régionaux sans problème via l'EMEX ;

[12] Est considérée comme grande adresse toute adresse dont le nombre de logements est au moins égal à 60 et qui est telle que l'ensemble des grandes adresses ne représente pas plus de 10% des logements de la commune.
[13] Avec cependant la possibilité d'effectuer un renouvellement au bout de cinq ans si nécessaire.

- enfin, pour assurer une réserve suffisante de logements chaque année, chaque groupe de rotation d'une ZAEPC contient un nombre minimal de logements principaux susceptibles d'être enquêtés, fixé à 300, et ce afin d'assurer, d'une part, une charge annuelle de travail suffisante pour l'enquêteur affecté à la zone en question, et d'autre part, de satisfaire le principe de disjonction qui impose de pouvoir effectuer plusieurs enquêtes la même année sans réinterroger les mêmes logements.

Ce travail de constitution des ZAE a été mené de l'automne à l'hiver 2006 par la division Échantillonnage et Traitement Statistique des Données de l'UMS et a conduit à partitionner le territoire national en 3785 Zones d'Action Enquêteurs, dont 892 ZAEGC et 2893 ZAEPC. Signalons enfin le statut particulier des villes de Paris, Lyon et Marseille pour lesquelles chacun des arrondissements a été considéré comme constitutif d'une ZAE.

ANNEXE 4 : Le cadre théorique des échantillons Octopusse et l'estimateur en expansion.

Le cadre théorique des échantillons Octopusse est celui d'un sondage en plusieurs phases :

On s'intéresse à l'ensemble U des logements. Ces logements sont partitionnés en grappes, les grappes étant les communes. Le symbole i désigne les communes et l les logements.

- La première phase se compose de la construction des groupes de rotation du recensement. Il convient ici de distinguer le cas des petites communes de celui des grandes communes :
  - ✓ pour les petites communes, on a partitionné l'ensemble de ces communes en cinq groupes de rotations GR_g, $g \in \{1,...,5\}$, de manière aléatoire et avec des probabilités $\pi_i^g = P(i \in g) = 1/5$. Chaque année, un de ces groupes de rotation est recensé de manière exhaustive. Les logements appartenant par nature à une seule et unique commune, la probabilité qu'un logement l appartenant à une petite commune *i* du groupe de rotation g soit recensé est donc $\pi_l^g = \pi_i^g$;
  - ✓ pour les grandes communes, cette opération se traduit, pour chaque grande commune, par une répartition des logements dans un des cinq groupes de rotation GR_g, $g \in \{1,...,5\}$, de manière aléatoire et avec des probabilités $\pi_{1,l}^g = P(l \in g)$. Chaque année, un de ces groupes de rotation est recensé par sondage à probabilités inégales $\pi_{2,l}^g$. Au final, la probabilité qu'un logement d'une grande commune appartenant au groupe de rotation g soit recensé est de $\pi_l^g = \pi_{1,l}^g * \pi_{2,l}^g$

  Conditionnellement aux cinq groupes de rotations GR_g ainsi définis, on effectue alors une seconde partition de l'ensemble des communes en un ensemble de K unités primaires selon un algorithme déterministe : il s'agit de l'étape de construction des ZAE.

- La deuxième phase est celle du tirage de l'échantillon-maître : un échantillon de ZAE est ainsi selectionné selon la méthode exposée en détail dans la présentation aux JMS 2009 du système Octopusse, avec des probabilités $\pi_{ZAE} = P(ZAE \in EM)$.

- La troisième et dernière phase est celle du tirage de l'échantillon d'une enquête Ménages à proprement parler : on interroge, par tirage systématique à probabilités égales, des logements de l'échantillon-maître recensés l'année précédant celle de l'enquête, c'est-à-dire des logements appartenant à l'intersection EM $\cap$ GR_g[14]. Pour une ZAE donnée, on sélectionne donc un échantillon ech_{ZAE}^g de n_{ZAE}^g logements parmi N_{ZAE}^g.

  Au final, l'échantillon ech^g au niveau national est constitué de la réunion des échantillons ech_{ZAE}^g.

On distingue donc trois niveaux d'aléas emboîtés :

- l'aléa lié au recensement rénové, qui conditionne la construction des groupes de rotation ainsi que celle des ZAE ;
- l'aléa de sondage associé à la constitution de l'échantillon-maître ;
- enfin l'aléa de sondage concernant le tirage des logements à enquêter pour une ZAE et un groupe de rotation donné.

Dans ce contexte, l'estimateur classique du total d'une variable Y est l'**estimateur en expansion** :

[14] L'indice g désigne ici le groupe de rotation recensé l'année précédant l'enquête.

$$\hat{T}_y^g = \sum_{ZAE \in EM} \sum_{l \in ZAE \cap GR_g \cap ech^g} \frac{Y_l}{\pi_l^g . \pi_{ZAE} . \frac{n_{ZAE}^g}{N_{ZAE}^g}}$$

Dans les simulations menées dans les parties 3.2 et 4 de cet article, on se place dans le cadre d'un recensement pour la troisième phase de tirage : en effet, on dispose, à partir des données du RP99, des totaux de Y par [ZAE ⊗ groupe de rotation] Y_{ZAE}^g dans les petites communes, et des totaux de Y par ZAE dans les grandes communes Y_{ZAE}. Dans un tel contexte, la formule de l'estimateur en expansion se simplifie grandement, et on obtient en distinguant ZAE petites communes et ZAE grandes communes :

$$\hat{Y}_{\text{expansion}}^g = \sum_{\text{ZAE grandes communes} \in EM} \frac{Y_{ZAE}}{\pi_{ZAE}} + \sum_{\text{ZAE petites communes} \in EM} \frac{Y_{ZAE}^g}{\frac{1}{5}\pi_{ZAE}}$$

ANNEXE 5 : Erreur relative et précision des estimations pour des échantillons-maîtres tirés seuls, en deux phases et en trois phases.

Groupe de rotation	EM tiré de manière indépendante		EM tiré en deux phases		EM tiré en trois phases	
	Erreur relative (en %)	CV (en %)	Erreur relative (en %)	CV (en %)	Erreur relative (en %)	CV (en %)
	Population sans double compte au RP99					
1	0,09	0,75	0,06	0,82	0,108	0,840
2	-0,06	0,79	-0,01	0,84	-0,027	0,885
3	-0,19	0,74	-0,16	0,83	-0,202	0,832
4	0,09	0,79	0,05	0,88	0,038	0,912
5	0,03	0,77	0,09	0,83	0,031	0,862
Moyenne	-0,01	0,27	0,01	0,27	-0,010	0,276
	Nombre de résidences principales au RP99					
1	-0,06	0,72	0,06	0,77	0,124	0,784
2	0,11	0,71	0,07	0,79	0,066	0,835
3	0,02	0,72	-0,10	0,79	-0,131	0,788
4	-0,10	0,67	0,01	0,83	0,002	0,872
5	0,04	0,74	-0,04	0,78	-0,061	0,841
Moyenne	0,00	0,00	0,00	0,00	0,000	0,000
	Population des résidences principales au RP99					
1	0,08	0,74	0,05	0,81	0,101	0,834
2	-0,01	0,78	0,03	0,83	0,019	0,880
3	-0,15	0,74	-0,13	0,83	-0,173	0,829
4	0,04	0,79	0,01	0,87	-0,008	0,906
5	0,01	0,75	0,05	0,82	0,009	0,861
Moyenne	-0,01	0,27	0,00	0,27	-0,011	0,281
	Nombre de décès entre le RP90 et le RP99					
1	-0,23	1,49	-0,21	1,53	-0,169	1,626
2	-0,39	1,54	-0,31	1,59	-0,368	1,617
3	0,04	1,47	0,10	1,54	0,059	1,567
4	0,16	1,47	0,14	1,59	0,157	1,544
5	0,41	1,51	0,44	1,59	0,362	1,609
Moyenne	0,00	0,76	0,03	0,77	0,008	0,776
	Nombre de naissances entre le RP90 et le RP99					
1	0,17	1,05	0,10	1,12	0,178	1,143
2	-0,05	1,07	-0,01	1,10	-0,014	1,148
3	-0,20	1,04	-0,18	1,10	-0,193	1,113
4	-0,06	1,07	-0,08	1,12	-0,068	1,174
5	0,08	1,04	0,10	1,10	0,070	1,124
Moyenne	-0,01	0,62	-0,01	0,62	-0,006	0,630
	Nombre d'individus de 0 à 19 ans au RP99					
1	0,16	1,01	0,12	1,06	0,178	1,098
2	-0,03	1,04	0,00	1,08	-0,010	1,136
3	-0,29	1,02	-0,27	1,08	-0,294	1,077
4	0,03	1,03	0,01	1,08	-0,027	1,142
5	0,05	1,00	0,12	1,04	0,071	1,064
Moyenne	-0,01	0,61	0,00	0,61	-0,016	0,626
	Nombre d'individus de 20 à 59 ans au RP99					
1	0,09	0,80	0,07	0,86	0,116	0,872
2	-0,06	0,84	0,00	0,87	-0,021	0,930
3	-0,23	0,78	-0,19	0,87	-0,242	0,868
4	0,12	0,84	0,08	0,90	0,061	0,956
5	0,02	0,81	0,09	0,87	0,021	0,884
Moyenne	-0,01	0,35	0,01	0,35	-0,013	0,358
	Nombre d'individus de plus de 60 ans au RP99					
1	0,03	1,13	-0,01	1,15	0,018	1,191
2	-0,11	1,13	-0,06	1,15	-0,069	1,203
3	0,05	1,07	0,03	1,13	0,009	1,140
4	0,08	1,10	0,04	1,22	0,052	1,180
5	0,03	1,10	0,02	1,15	0,001	1,213
Moyenne	0,02	0,56	0,00	0,55	0,002	0,569
	Revenu net imposable de l'année 1996					
1	-0,10	0,88	-0,10	0,95	-0,081	0,911
2	0,05	0,86	0,10	0,90	0,089	0,969
3	-0,11	0,85	-0,09	0,92	-0,139	0,927
4	0,22	0,87	0,15	0,93	0,161	0,973
5	-0,07	0,85	-0,04	0,90	-0,057	0,938
Moyenne	0,00	0,39	0,00	0,40	-0,005	0,400
	Revenu fiscal 2004					
1	0,07	0,75	0,03	0,81	0,056	0,820
2	0,27	0,80	0,29	0,88	0,274	0,916
3	-0,18	0,71	-0,16	0,82	-0,210	0,830
4	-0,02	0,76	-0,07	0,86	-0,095	0,907
5	-0,10	0,75	-0,06	0,81	-0,097	0,867
Moyenne	0,01	0,29	0,01	0,32	-0,014	0,331

	Nombre de ménages fiscaux en 2004					
1	0,13	0,73	0,08	0,77	0,129	0,802
2	0,16	0,76	0,22	0,84	0,194	0,858
3	-0,10	0,69	-0,09	0,81	-0,138	0,798
4	-0,09	0,75	-0,11	0,85	-0,125	0,892
5	-0,10	0,75	-0,06	0,80	-0,094	0,855
Moyenne	0,00	0,15	0,01	0,15	-0,007	0,153
	Nombre de personnes dans les ménages fiscaux en 2004					
1	0,12	0,76	0,09	0,82	0,132	0,850
2	0,17	0,83	0,23	0,88	0,197	0,912
3	-0,13	0,76	-0,10	0,85	-0,157	0,829
4	-0,14	0,79	-0,16	0,88	-0,182	0,919
5	-0,07	0,78	-0,01	0,83	-0,056	0,874
Moyenne	-0,01	0,32	0,01	0,33	-0,013	0,331
	Nombre de chômeurs au sens de l'ANPE au T1 2004					
1	0,20	1,20	0,14	1,24	0,201	1,293
2	0,23	1,33	0,26	1,33	0,287	1,343
3	-0,08	1,28	-0,06	1,33	-0,096	1,357
4	-0,15	1,26	-0,18	1,31	-0,186	1,299
5	-0,15	1,23	-0,11	1,29	-0,141	1,295
Moyenne	0,01	0,86	0,01	0,84	0,013	0,830
	Nombre de chômeurs au sens de l'ANPE au T1 2005					
1	0,28	1,22	0,23	1,27	0,284	1,322
2	0,29	1,33	0,35	1,37	0,358	1,359
3	-0,18	1,27	-0,17	1,33	-0,216	1,337
4	-0,21	1,28	-0,22	1,33	-0,222	1,329
5	-0,18	1,27	-0,17	1,30	-0,190	1,318
Moyenne	0,00	0,88	0,00	0,86	0,003	0,858
	Nombre de chômeurs au sens de l'ANPE au T1 2006					
1	0,27	1,23	0,21	1,29	0,252	1,335
2	0,40	1,35	0,44	1,40	0,437	1,389
3	-0,08	1,33	-0,07	1,36	-0,113	1,378
4	-0,16	1,30	-0,18	1,36	-0,192	1,342
5	-0,38	1,29	-0,36	1,33	-0,390	1,358
Moyenne	0,01	0,89	0,01	0,88	-0,001	0,864
	Nombre de chômeurs au sens de l'ANPE au T1 2005					
1	0,18	1,26	0,15	1,32	0,163	1,373
2	0,45	1,40	0,51	1,43	0,480	1,438
3	-0,11	1,36	-0,12	1,41	-0,164	1,422
4	-0,06	1,34	-0,05	1,39	-0,080	1,373
5	-0,44	1,30	-0,41	1,39	-0,427	1,401
Moyenne	0,00	0,91	0,02	0,91	-0,006	0,896

Analyse en moyenne et variance de la distribution empirique des estimateurs, fondée sur 1000 tirages indépendants d'échantillons-maîtres tirés respectivement en une phase, deux phases et trois phases