

SELECTION, PONDERATIONS, ALLOCATIONS..., DANS LES NOUVEAUX ECHANTILLONS DES ENQUETES MENAGES DE L'INSEE.

Marc Christine¹, Sébastien Faivre²

RESUME

Le nouveau recensement rotatif mis en place en France conduit à revoir la constitution des échantillons des enquêtes ménages de l'Insee. D'une part, pour impacter chaque année la fraction recensée la plus récente au sein de zones géographiques fixes, il a fallu construire des unités primaires spécifiques. Ensuite, le tirage de ces unités, articulant l'échantillon-maître national et le complément destiné aux extensions régionales, a nécessité l'élaboration d'une méthode en plusieurs phases, respectant des conditions d'équilibrage et des probabilités d'inclusion fixées. Enfin, les choix d'estimateurs conduisent à revisiter la problématique des pondérations des unités finales (logements) et à vérifier leur qualité.

MOTS CLES : Echantillon-Maître, Recensement rotatif, Unités Primaires

ABSTRACT

The new rotative Census settled in France leads to review the building of samples for household surveys conducted by Insee. First, to draw samples from the last part of territory covered by Census, in fixed areas, it has been necessary to build specific primary units. Then, drawing those units - some of them devoted to the national Master sample, the others to regional extensions - needs a two-phase method, using both balancing conditions and fixed probabilities of selection. At last, different kinds of estimators are studied, with impact on weightings and quality of the final samples of dwellings resulting from this process.

KEY WORDS: Master Sample, Rotative Census, Primary Units

1. LE CONTEXTE DU NOUVEAU RECENSEMENT

1.1 Le système actuel (Echantillon-Maître 1999).

Depuis la décennie 60, les échantillons des enquêtes nationales auprès des ménages réalisées par l'Insee sont sélectionnés dans des listes de logements constituées à partir de chaque recensement de la population. Ces listes sont complétées par des sources annexes (fichiers de permis de construire) permettant la couverture des logements construits postérieurement au dernier recensement disponible, dits « neufs ».

1.2 La nécessité de prendre en compte la nouvelle méthodologie de recensement.

Depuis janvier 2004, *une nouvelle méthodologie de recensement « rotatif annuel »* a été mise en place. Elle distingue les Petites communes (moins de 10 000 habitants) : constitution aléatoire de 5 groupes de rotation de *communes (à probabilités égales) et* recensement exhaustif annuel d'un des groupes de rotation ; et les Grandes communes : constitution aléatoire de 5 groupes de rotation *d'adresses* ; tirage *chaque année* d'un échantillon de logements au sein d'un groupe donné (environ 8% des logements de la commune) et recensement de ces logements.

Le « nouveau recensement » va donc fournir des listes d'unités « échantillonnables » renouvelées chaque année, permettant l'apport d'« information fraîche » mais en contrepartie de la perte d'exhaustivité sur l'ensemble du territoire. Une refonte globale du système d'échantillonnage des enquêtes ménages est donc rendue nécessaire. Le projet « *nouveaux systèmes d'échantillonnage* » (OCTOPUSSE³) a été lancé fin 2003 et doit permettre le tirage des premières enquêtes pour mai 2009.

¹ Marc CHRISTINE, Institut National de la Statistique et des Etudes Economiques, 18 bd Adolphe Pinard, 75675 Paris Cedex 14, France, marc.christine@insee.fr

² Sébastien FAIVRE, INSEE, sebastien.faivre@insee.fr

³ Organisation Coordonnée de Tirages Optimisés Pour une Utilisation Statistique des Echantillons.

2. LES NOUVELLES ORIENTATIONS RETENUES POUR LES FUTURS ECHANTILLONS

Les constantes liées à l'organisation de la *collecte en face à face*, et la nécessité de ne pas disperser les lieux d'enquêtes et de limiter les déplacements des enquêteurs, subsistent. Elles rendent toujours utile un système de type « Echantillon-maître » : celui-ci comprendra *des ZAE (zones d'action enquêteurs) constituant des unités primaires, construites une fois pour toutes*, dont un échantillon sera tiré aléatoirement pour pouvoir leur associer un enquêteur stable dans le temps et localisé à proximité, et au sein desquelles seront tirés les échantillons finaux de logements.

L'innovation principale consiste à sélectionner les échantillons des enquêtes dans la fraction des unités primaires tirées recensée l'année précédente. Les avantages de ce principe sont nombreux : minimiser le nombre de logements transformés, détruits, hors champ ; améliorer le **ciblage de certaines catégories de population, grâce à la fraîcheur de l'information** disponible sur les logements de la base de sondage ; **s'affranchir d'un système complémentaire pour l'échantillonnage des logements « neufs ».**

3. LA CONSTRUCTION DES UNITES PRIMAIRES (ZAE, Zone d'Action Enquêteur).

3.1 Contraintes et objectifs.

Il s'agissait donc de construire des ZAE au sein de chaque région :

- réalisant une partition du territoire
- comportant des logements des 5 groupes de rotation du RP
- avec un nombre minimal de logements « échantillonnables » par groupe de rotation pour disposer d'une réserve suffisante pour tirer plusieurs échantillons distincts d'enquête la même année sans réinterroger les mêmes logements.

3.2 La construction effective des ZAE.

Les règles suivantes de constitution des ZAE ont été adoptées :

- **constitution des ZAE en respectant les frontières régionales.**
- **séparation ZAE grandes communes (ZAEGC) et ZAE petites communes (ZAEPCC)**
- **une grande commune constitue une ZAE à elle seule car elle contient des logements de tous les groupes de rotation.**
- **au moins 300 résidences principales par groupe de rotation dans chaque ZAE.**
- **avec l'objectif de minimiser leur étendue géographique.**

Les ZAEPCC sont des agrégats de communes. Leur construction (c'est-à-dire la réalisation automatique d'une partition « optimale » du territoire métropolitain sous les contraintes précitées) a constitué un problème nouveau qui a nécessité la conception et la mise en œuvre d'algorithmes spécifiques, pour être assurée de manière automatique.

3.3 Le résultat de la construction des ZAE.

Au final, les 36 613 communes françaises au 1^{er} janvier 2006 ont été affectées à une ZAE, soit 2 893 ZAEPCC et 850 ZAEGC.

On notera que l'algorithme de construction des ZAE est déterministe mais l'affectation initiale des communes en groupes de rotation est aléatoire : il en résulte que *les ZAE sont des « objets aléatoires » (contrairement à la situation habituelle en termes d'unités primaires).*

3.4 Qualité de la construction des ZAE.

Pour juger de l'optimalité en termes d'étendue, on a vérifié que les distances intra-ZAE étaient comparables avec celles des unités primaires constituées pour l'Echantillon-Maître actuel et demeuraient acceptables.

La moyenne des distances par la route sur les 2893 ZAEPIC constituées est estimée à 10 km, à comparer à celle des 3202 UP 1999 constituées, soit 8 km. La distance annuelle maximale moyenne pour les ZAEPIC est de 18 km.

4. ALLOCATION ET TIRAGE DES ZAE

4.1 Calcul de l'allocation.

Les ZAE sont tirées proportionnellement à leur taille (nombre de logements principaux), certaines étant retenues d'office (« exhaustives »).

Le nombre de ZAE-EM à tirer a été fixé en prenant l'hypothèse conventionnelle suivante (analogue à celle prise pour l'EM 99) : **pour une enquête nationale au taux moyen de 1/2000** (un peu moins de 12.000 logements principaux), **on affecte 20 Fiches-adresses par enquêteur**⁴.

On montre que le seuil d'exhaustivité vaut : $S = \frac{e}{\tau}$, pour un échantillon d'enquête au taux moyen τ , chaque enquêteur ayant une charge moyenne de e fiches-adresses. **Ce seuil ne dépend pas de la région considérée. On obtient aussi le nombre de ZAE à tirer dans la sous-strate non exhaustive :**

$$k = \frac{\tau(N - N^{exh})}{e}, \text{ où : } N^{exh} = \sum_{i / N_i \geq \frac{e}{\tau}} N_i,$$

avec : N_i la taille de la ZAE i et N la taille de la région (nombre de logements principaux).

Résultats :

- le seuil d'exhaustivité résultant est à 40.000 logements principaux
- 37 grandes communes exhaustives (qui seront affectées à plusieurs enquêteurs)
- 488 ZAE non exhaustives tirées dont :
 - o 286 ZAE-PC
 - o 202 ZAE-GC non exhaustives.

4.2 Tirage des ZAE.

Le tirage est stratifié *par région* (cas particulier de l'Île de France : on sépare la « petite⁵ » et la « grande » couronne). Il est également équilibré sur des *taux régionaux*. Il est nécessaire d'équilibrer non seulement au niveau de l'ensemble de la ZAE *mais aussi de chacun des 5 groupes de rotation*, de manière à avoir chaque année une base de sondage « représentative ». Cela multiplie le nombre de contraintes d'équilibrage (cinq contraintes pour une variable) et réduit d'autant le nombre de variables indépendantes à introduire. On notera que la base de sondage annuelle est équilibrée sur le total des communes du groupe de rotation impacté mais pas sur le total France métropolitaine et que les groupes de rotation ne sont pas rigoureusement équivalents.

Le choix des variables d'équilibrage est issu de nombreuses simulations visant à déterminer les variables d'équilibrages « optimales » vis-à-vis de la qualité de l'équilibrage. On a finalement retenu : le nombre de résidences principales des ZAE par groupe de rotation ; le revenu fiscal 2004 ventilé par groupe de rotation ; enfin, le nombre de résidences principales dans les différents types d'espace (rural, périurbain et urbain).

4.3 Le problème du tirage simultané EM-EMEX

⁴ Le principe général étant qu'une ZAE est affectée à un seul enquêteur, hormis celles formées de grandes communes exhaustives.

⁵ Communes limitrophes de PARIS ou à proximité immédiate.

En fait, dans chaque région ont été tirés deux jeux de ZAE :

- les ZAE de la partie « Echantillon-Maître » d'OCTOPUSSE dans lesquelles seront tirées les enquêtes « nationales » en l'absence d'extensions régionales.
- les ZAE de la partie « extensions régionales » d'OCTOPUSSE qui constitueront un complément de zones à enquêter en cas de tirage d'extensions régionales, afin de mieux couvrir le territoire régional (EMEX).

Pour la strate « non exhaustive » (cf. supra, § 4.1), le nombre de ZAE tirées au titre des extensions régionales est égal au nombre de ZAE tirées pour la partie Echantillon-Maître national.

Il a été retenu un tirage simultané de l'ensemble des ZAE au sein de chaque région (ZAE Echantillon-Maître et ZAE Extensions régionales) : on tire d'abord l'ensemble de toutes les ZAE qui serviront pour l'EM ou les EMEX, puis au sein de cet ensemble, celles qui formeront spécifiquement chacune des composantes (EM, EMEX). On adopte une méthode de tirage basée sur un tirage équilibré au niveau régional, avec des probabilités proportionnelles à la taille des ZAE.

On montre que, pour assurer un équilibrage sur le total d'une variable X, **à la fois 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.

Des ZAE exhaustives spécifiques EMEX apparaîtront (retenues d'office, si l'on mobilise l'EMEX, mais seulement avec une certaine probabilité si l'on n'utilise que l'EM).

5. TIRAGE DES LOGEMENTS DANS LES ZAE : ALLOCATIONS ET PONDERATIONS

Au sein de chaque ZAE tirée, les unités secondaires (logements) sont tirées par sondage aléatoire simple, dans le groupe de rotation à impacter.

Toutefois, dans les grandes communes, une phase intermédiaire est nécessaire. En effet :

– la 1^{ère} phase RP conduit à une affectation inégale et « semi-déterministe » (et non à probabilités égales 1/5) des adresses dans les différents groupes de rotation (notamment les grandes adresses) : il est donc nécessaire de reconstituer une « pseudo-probabilité » d'affectation.

– lors de la 2^{ème} phase RP (tirage de logements à recenser dans le groupe de rotation annuel en grande commune), il y a une surreprésentation des « adresses neuves » et des « grandes adresses » (recensées d'office dans chaque groupe de rotation d'adresses).

Cela nécessite donc un **rééchantillonnage** des logements au sein de la base de sondage pour disposer d'une base effective de logements à poids identiques.

5.1 Types d'estimateurs envisagés

Deux formules d'estimation sont possibles (cas des petites communes) :

- « **en expansion** » (analogue à HORWITZ-THOMSON) :
$$\hat{T}_L(Y) = 5 \sum_{k \in S_{ZAE}} \frac{N_{k,t}}{\pi_k} \bar{y}_{k,t}$$

- « **calé par ZAE** » (ou **par ratio**), avec calage a priori sur le nombre de résidences principales de la ZAE :

$$\hat{T}(Y) = \sum_{k \in S_{ZAE}} \frac{N_k}{\pi_k} \bar{y}_{k,t}$$

$N_{k,t}$: taille de la **fraction de la ZAE k tombant dans le groupe de rotation t**

N_k : taille totale de cette ZAE

$n_{k,t}$: taille de l'échantillon aléatoire simple de logements puisé dans le groupe de rotation t

$\bar{y}_{k,t}$: moyenne empirique des Y sur l'échantillon de logements tiré au second degré dans la ZAE k et le groupe t.

π_k : probabilité d'inclusion de la ZAE k.

A chacun de ces estimateurs correspondent des jeux de poids différents pour les logements tirés au 2nd degré.

On peut résumer les propriétés comparatives des deux estimateurs dans le tableau suivant :

Tableau 1 - Propriétés comparatives des deux estimateurs

	Equi-pondération	Equi-allocation (pour ZAE non exhaustives)	Biais	
			Conditionnel aux groupes de rotation	Non conditionnel
Estimateur calé	OUI	OUI, quasiment	OUI	Probable, incalculable
			NON, sous hypothèse d'homogénéité des comportements entre communes d'une même ZAE	
Estimateur en expansion	← ANTINOMYQUE →		OUI	NON
		« Raboutage » nécessaire si équi pondération		

5.2 Impact sur le calcul des allocations.

L'équiallocation est un objectif (charges de travail uniformes par enquêteur). Si elle ne peut être satisfaite rigoureusement, les calculs d'allocation cherchent à **minimiser la dispersion des pondérations finales logements**, sous des contraintes :

- de taille totale d'échantillon fixée
- de charge minimale et maximale par enquêteur (c'est-à-dire par ZAE non exhaustive) : par exemple, fourchette 20-40 pour une enquête de 20 000 logements.

Les poids des logements pris en compte dépendent de la forme de l'estimateur retenu. La minimisation peut se faire nationalement ou par région.

6. RETOUR SUR LE TIRAGE DES ZAE : QUALITE, CALAGE ET PONDERATIONS

On a cherché à analyser la qualité du tirage des ZAE en comparant :

- l'estimation (à partir de l'échantillon de ZAE) du total « France entière » de différentes variables auxiliaires (obtenue à partir des totaux observés sur les ZAE tirées, supposés connus, i.e. avant tirage du second degré)
- avec le vrai total France entière (connu par des sources exhaustives, RP 99).

On constate des erreurs relatives, plus ou moins importantes, variables d'un groupe de rotation à l'autre, quel que soit le type d'estimateur retenu, notamment pour la segmentation par type d'espace (rural / périurbain / urbain) et la répartition du nombre de personnes employées par secteur (variables imparfaitement ou pas prises en compte dans l'équilibrage lors du tirage).

Ces erreurs peuvent être pénalisantes, en particulier pour les enquêtes annuelles, *mesurant des évolutions, alors que, par construction, les communes impactées d'une année sur l'autre seraient distinctes (pour les ZAEPC).*

Une solution pour remédier à ce problème : le calage des ZAE.

L'objectif est d'effectuer *chaque année* un (re)calage des ZAE pour obtenir une base de sondage annuelle « représentative ». On part donc des poids initiaux donnés par l'un ou l'autre des deux estimateurs proposés ci-dessus (en expansion ou calé par ZAE). Le calage est effectué nationalement (sauf en cas d'extension régionale) et pour chaque groupe de rotation séparément. Il en résulte *plusieurs jeux de poids calés*.

Les variables de calage utilisées ont été : les variables d'équilibrage lors du tirage des ZAE (nombre de résidences principales, revenu fiscal total, âge en trois tranches, nombre de résidences principales dans les espaces urbain/périurbain/rural) ; les variables d'emploi par secteur ; la répartition par tranches de taille d'unité urbaine.

Le calage a ainsi permis d'obtenir une erreur relative égale à zéro sur les variables de calage, sans la voir augmenter pour les autres variables d'intérêt.

Conséquences du calage.

Les nouveaux poids calés des ZAE (*recalculés chaque année*) sont réinjectés dans le calcul des allocations de logements par ZAE, lesquelles *dépendent donc de ces nouveaux poids aléatoires* et non plus directement des poids initiaux des ZAE.

On a donc mis en place une méthodologie innovante de calage des unités primaires. Une validation empirique de cette méthode a été opérée sur la base de l'échantillon de ZAE tiré. De surcroît, cette technique permet de résoudre le dilemme du choix d'estimateurs et conduit à choisir l'estimateur en expansion calé (au lieu de l'estimateur par le ratio calé, pour lequel un « double calage » devrait alors être effectué).

La méthode permettra également d'incorporer une *information récente* en calant sur les données du nouveau RP, disponibles début 2009 (alors que la construction et le tirage des ZAE ont utilisé des données du RP 1999).

7. CONCLUSIONS ET PERSPECTIVES

Le recours à un RP annuel permet un gain substantiel en qualité de la base de sondage (possibilité de cibler sur des caractéristiques fraîches) mais a entraîné une complexification importante du processus d'échantillonnage. Des solutions innovantes ont dû être imaginées dans un contexte inhabituel : construction d'unités primaires aléatoires, vérification et amélioration de la qualité d'un tirage conditionnel de celles-ci via une procédure de calage des UP.

De nombreuses autres questions seront à étudier et feront l'objet de travaux ultérieurs :

- impact du choix de l'estimateur sur les allocations et les estimations (EQM)
- mise en œuvre d'une procédure de calage pour les grandes communes
- changement éventuel de statut des communes (grande / petite) à partir de 2011
- calcul de précision...

REFERENCES

Caron N. et Christine M. (2006). *Les échantillons des enquêtes « ménages » fondés sur le nouveau recensement français*. Communication présentée au 4^{ème} Colloque Francophone sur les Sondages (Québec, mai 2005).

Faivre S. et Christine M. (2007). *Les nouvelles orientations pour le projet Nouveau Echantillons des enquêtes ménages de l'INSEE*, Communication au 5^{ème} Colloque Francophone sur les Sondages (Marseille, Novembre 2007).

Faivre S. et Loonis V. (2007). *Construction des Zones Action Enquêteurs dans le projet Nouveaux Echantillons des enquêtes ménages de l'INSEE*, Communication au 5^{ème} Colloque Francophone sur les Sondages.