Le projet Octopusse de nouvel échantillon-maître de l'Insee

Sébastien FAIVRE[1] et alii

Les échantillons des enquêtes ménages réalisées par l'Insee sont, depuis les années 60, tirés dans les fichiers de logements constitués à l'issue de chaque recensement « général », complétés par des listes de logements « neufs » (construits après le dernier recensement) issues des fichiers de permis de construire. Afin de limiter les déplacements des enquêteurs, un ensemble d'unités primaires (zones géographiques connexes) étaient tirées à l'initialisation de l'échantillon-maître et restaient en vigueur pour la durée de vie du système (période intercensitaire).

Ce système qui a prévalu jusqu'à l'aube du XXIème siècle n'est plus compatible avec la révolution opérée par le nouveau recensement et, notamment, son caractère désormais rotatif et partiel.

En effet, depuis janvier 2004, le recensement de la population s'effectue de manière annuelle. Les petites communes (moins de 10.000 habitants) ont été réparties aléatoirement en 5 groupes de rotation, qui sont chacun recensés exhaustivement une année donnée. Pour les grandes communes (10.000 habitants ou plus), on interroge chaque année par sondage un échantillon de logements représentant environ 8% des logements de la commune.

Il était donc nécessaire de repenser complètement le système d'échantillonnage des enquêtes ménages. Dans ce cadre, un projet baptisé Octopusse[2] a été lancé en 2003. Il sera opérationnel à la mi-2009 (premiers tirages d'enquêtes possibles à partir de mai 2009).

L'innovation principale du futur système consiste à bénéficier de la « fraîcheur » du nouveau recensement, c'est-à-dire à ***utiliser comme base de sondage, pour les enquêtes réalisées au cours de l'année n+1, les listes de logements recensés au début de l'année n.***

De ce fait, le concept d'unités primaires a dû être repensé, pour tenir compte du principe de fraîcheur mais aussi de la nécessité de permettre à chaque enquêteur de rester affecté à une zone donnée et d'y réaliser des enquêtes chacune des cinq années du cycle. Ainsi, seule la fraction de l'unité primaire appartenant au dernier groupe recensé est mobilisée pour le tirage d'une enquête donnée.

La construction des nouvelles unités primaires (ou ZAE : Zones d'Action Enquêteurs) a suscité un travail méthodologique innovant et relativement complexe : il s'agissait en effet de construire des zones fixes, comportant des logements appartenant à chacun des cinq groupes de rotation. Chaque grande commune constitue une ZAE à elle toute seule car une fraction est recensée chaque année. Pour les petites communes, il s'agissait de construire des agrégats

[1] DSDS, Insee

[2] ***O***rganisation ***C***oordonnée de ***T***irages ***O***ptimisés ***P***our une ***U***tilisation ***S***tati***S***tique des ***E***chantillons.

composés de communes appartenant aux 5 groupes de rotation, avec un nombre minimal de 300 logements principaux dans chacun de ces groupes (afin de disposer chaque année d'une réserve suffisante de logements), tout en visant à minimiser l'étendue géographique de ces zones. Une solution automatisée performante a pu être mise en œuvre : on a ainsi constitué 2893 ZAE sur l'ensemble du territoire (auxquelles s'ajoutent 850 grandes communes).

Un échantillon de ZAE a ensuite été tiré au sein de chaque région avec des probabilités proportionnelles à leur taille, et sous des conditions d'équilibrage impliquant différentes données socio-démographiques : population, revenu, type d'espace (rural/urbain), nature de la ZAE (petites ou grandes communes). Les 37 grandes communes ayant plus de 40 000 résidences principales au RP 1999 ont été retenues d'office (cf. communication de Fabien GUGGEMOS).

Cependant, l'introduction des contraintes d'équilibrage ne suffit pas pour assurer directement la « représentativité » des cinq bases de sondage annuelles, compte-tenu des limites à l'équilibrage liées au faible nombre de ZAE tirées par région et à la nécessité de décliner une variable d'équilibrage en cinq variables au niveau groupe de rotation pour équilibrer les cinq bases annuelles sur cette variable. Par ailleurs, les variables d'équilibrage sont issues pour la plupart du recensement de 1999. Or c'est plutôt sur la base des données des premières populations légales que sera appréciée la représentativité des bases de sondage annuelles Octopusse et ces données ne seront disponibles qu'à l'initialisation du système en mai 2009.

Une solution de calage des ZAE sera alors mise en œuvre pour assurer la représentativité des bases de sondage annuelles, sur la base des informations issues des dernières données détaillées des populations légales.

L'objectif fondamental lors du tirage de l'échantillon d'une enquête donnée est d'**assurer l'équiprobabilité finale des logements** (sauf dans le cas où le concepteur d'enquête souhaite assurer une sur-représentation de certaines catégories de logements ou de ménages[3]). Cependant, l'autre objectif **de tirer, au sein de chaque ZAE sélectionnée aléatoirement, un nombre égal de fiches-adresses** (pour répartir au mieux la charge de collecte entre les enquêteurs et éviter un effet de grappe important dans certaines ZAE fortement impactées), conduit à la recherche d'une solution de compromis.

L'obtention d'un échantillon de logements à probabilités égales se fait alors en deux étapes.

- on cherche tout d'abord à corriger les inégalités entre les probabilités de tirage des logements au sein d'une même ZAE (dues aux taux de sondage différents des logements dans le RP, au sein des grandes communes, selon le type d'adresse à laquelle ils appartiennent). On procède pour cela à **un rééchantillonnage des logements, qui constitue une 3ème phase de tirage.** On obtient à l'issue de cette phase une « base utile » dans laquelle chaque logement chargé a une probabilité identique[1].

- pour une enquête donnée, les logements seront tirés au moyen d'un tirage systématique à probabilités égales au sein de la base utile annuelle de chaque ZAE sélectionnée.

[3] Cependant, même en cas de surreprésentation de certaines catégories de population, la première étape reste un tirage à probabilités égales (avant un tirage de seconde phase pour surreprésenter certains groupes).

[1] Aux arrondis près dans le calcul du nombre de logements à rééchantillonner.

Les allocations de logements à tirer dans chaque ZAE sont calculées en cherchant à rendre les poids finaux des logements les plus voisins possibles, sous la contrainte de la taille totale d'échantillon et de contraintes pratiques à respecter au niveau de la réalisation des enquêtes (nombre minimum et maximum de fiches-adresses à tirer par ZAE…).

L'article présente les différents aspects méthodologiques du fonctionnement de l'échantillon-maître Octopusse évoqués ci-dessus.

Il aborde aussi la problématique de la construction de l'EMEX (Échantillon-Maître pour les Extensions Régionales), composé de ZAE supplémentaires mobilisées pour les extensions régionales[2] de façon que l'ensemble des ZAE EM et EMEX soit « représentatif » au niveau régional. En effet, la construction de l'EMEX a conduit à des travaux méthodologiques originaux pour la mise au point d'une méthode de tirage d'échantillons équilibrés emboîtés, restant représentatifs à chaque niveau de tirage de la population de départ (France métropolitaine) pour les variables d'équilibrage introduites (cf. communication de Marc CHRISTINE et Emmanuel GROS).

Enfin, on présente le cas du tirage d'échantillons ZUS où la méthode retenue est de tirer cinq sous-échantillons dans cinq bases annuelles ZUS correspondant à chacun des cinq groupes de rotation, les bases ZUS rentrant en service avec un décalage d'un an par rapport à la base EM/EMEX pour éviter de « vider » de ses logements ZUS la base de sondage active EM/EMEX.

[2] Échantillon complémentaire à l'échantillon national financé par une administration régionale.