LES PLANS DE SONDAGE DU NOUVEAU RECENSEMENT

Philippe BERTRAND, Guillaume CHAUVET,
Barbara CHRISTIAN, Jean-Marie GROSBRAS

Insee, Programme de rénovation du recensement
Maîtrise d'œuvre Méthodologie

Introduction

La loi n° 2002-276 du 27 février 2002 relative à la démocratie de proximité, publiée au Journal officiel n° 50 du 28 février 2002 définit dans son titre V (« Des opérations de recensement », articles 156 à 158) le cadre général des enquêtes de recensement et de la production des chiffres de la population, en particulier au paragraphes VI de l'article 156 :

> *VI. - Les dates des enquêtes de recensement peuvent être différentes selon les communes. Pour les communes dont la population est inférieure à 10 000 habitants, les enquêtes sont exhaustives et ont lieu chaque année par roulement au cours d'une période de cinq ans. Pour les autres communes, une enquête par sondage est effectuée chaque année ; la totalité du territoire de ces communes est prise en compte au terme de la même période de cinq ans. Chaque année, un décret établit la liste des communes concernées par les enquêtes de recensement au titre de l'année suivante.*

L'idée de base repose donc sur l'observation, chaque année, d'une fraction de la population, choisie grâce à des méthodes de sondage plutôt que sur une observation exhaustive tous les 7 à 9 ans. Une première présentation des travaux entrepris en matière de méthodologie du nouveau recensement ont été réalisés, notamment par Jean Dumais [1], [2], [3]. Nous allons présenter ici les choix retenus pour les deux strates spécifiées par la loi, c'est-à-dire les communes de moins de 10 000 habitants et les autres communes. Pour les premières, il s'agit d'opérer une répartition en cinq groupes destinés à être enquêtés exhaustivement dans cycle de cinq ans ; pour les communes de 10 000 habitants ou plus, il s'agit de réaliser des enquêtes annuelles, de sorte que la totalité du territoire soit prise en compte dans une période de cinq ans.

1. Le cas des communes de moins de 10 000 habitants

1.1 Le problème

Les données recueillies au sein des communes enquêtées une année donnée, combinés à celles collectées dans les communes de 10 000 habitants ou plus, servent à produire des données statistiques nationales et régionales, fondées sur les collectes de l'année et valables pour cette même année.

Les résultats de l'ensemble des cinq enquêtes d'un cycle servent à produire des populations légales pour l'ensemble des circonscriptions administratives du territoire et des statistiques « détaillées » aux niveaux communal et infra-communal, à valeur pour l'année médiane du cycle.

Cette double exigence conduit à définir les critères de représentativité auxquels doit souscrire chacun des groupes.

En effet, les estimations nationales et régionales seront d'autant plus fiables que les groupes seront individuellement l'image exacte de l'ensemble. L'argument vaut aussi pour les estimations détaillées qui prennent en compte les enquêtes dans les cinq groupes : les estimations construites par l'agrégation de plusieurs parties sont d'autant plus précises que ces parties sont homogènes.

C'est pourquoi dans toutes les régions, les groupes de communes de moins de 10 000 habitants sont l'image fidèle de leur région en termes de population par âge, par sexe, par type de logement (individuel ou collectif), de nombre de logements par département

1.2 La méthode

La méthode statistique utilisée est celle des échantillons équilibrés. Généralisant la notion de stratification, elle consiste à choisir des structures de référence et à construire des échantillons reproduisant, le plus fidèlement possible, ces structures.

Dans le cas présent, les structures de référence sont à choisir parmi les variables démographiques et les catégories de logement. Les valeurs cibles sont établies à partir du recensement de 1999. En d'autres termes, on fait l'hypothèse, par exemple, qu'un ensemble de communes dont la population, en 1999, a une structure par âge identique à celle de l'ensemble des communes de moins de 10 000 habitants conservera, au moins pendant un certain temps, une bonne qualité de représentativité sur ce critère.

A quel niveau géographique peut-on assurer une bonne représentativité ? Le problème statistique posé s'exprime en termes de « degrés de liberté ». Pour faire image, on peut se figurer un ensemble de cinq balances dont on veut équilibrer les plateaux à la même hauteur. Les poids placés dans les plateaux sont les communes, et elles sont de tailles disparates. Intuitivement, on voit qu'un équilibrage correct suppose que l'on ait suffisamment de poids à répartir, c'est-à-dire suffisamment de degrés de liberté.

La contrainte de degrés de liberté ne peut être satisfaite au niveau départemental. Elle peut l'être correctement dans des départements pourvus d'un grand nombre de communes de moins de 10 000 habitants mais pas dans les autres. Pour appliquer un principe homogène sur le territoire, on a donc retenu un niveau d'équilibrage régional.

1.3 Les variables de référence

Pour équilibrer les groupes de rotation des communes de moins de 10 000 habitants dans chacune des régions, les variables suivantes, issues du recensement de population de 1999, ont été retenues :

- le nombre de logements ;
- le nombre de logements en immeuble collectif ;
- la population des moins de 20 ans ;
- la population des 20-39 ans ;
- la population des 40-59 ans ;
- la population des 60-74 ans ;
- la population des 75 ans et plus ;
- la population des femmes ;
- la population des hommes ;
- pour chacun des départements de la région, la population totale.

Les variables de type "logement" permettent d'obtenir l'équilibre entre groupes de rotation sur la proportion de logements dans le collectif. Cela a une influence sur la répartition des "grandes petites communes" dans les groupes de rotation. Cela permet également d'obtenir des groupes de rotation qui évolueront de façon plus homogène. Les tranches d'âge et le sexe, variables de population essentielles, assurent l'homogénéité des groupes de rotation pour les structures de population. On s'assure également que les départements sont représentés à leur poids dans chaque groupe.

2. Le cas des communes de 10 000 habitants ou plus.

2.1 Le problème

Les données recueillies au sein des communes enquêtées une année donnée, combinés à celles collectées dans les communes de moins de 10 000 habitants, servent à produire des données statistiques nationales et régionales, fondées sur les collectes de l'année et valables pour cette même année.

Les résultats de l'ensemble des cinq enquêtes d'un cycle servent à produire des populations légales pour l'ensemble des circonscriptions administratives du territoire et des statistiques « détaillées » aux niveaux communal et infra-communal, à valeur pour l'année médiane du cycle.

Cette double exigence conduit à définir les critères de représentativité auxquels doit souscrire chacun des échantillons annuels.

En effet, les estimations nationales et régionales seront d'autant plus fiables que les groupes seront individuellement l'image exacte de l'ensemble. L'argument vaut aussi pour les estimations détaillées qui prennent en compte les enquêtes dans les cinq groupes : les estimations construites par l'agrégation de plusieurs parties sont d'autant plus précises que ces parties sont homogènes.

C'est pourquoi les échantillons enquêtés annuellement doivent être une image fidèle de leur commune en termes de population par âge, par sexe, par type de logement (individuel ou collectif), de la répartition infra-communale des logements.

2.2 Unités échantillonnées, bases et taux de sondage

Le plan de sondage est un plan "à l'adresse", toute adresse échantillonnée étant enquêtée de façon exhaustive. Cette contrainte est forte si l'on doit prendre en compte le niveau infracommunal, niveau auquel il faut pouvoir obtenir des estimations détaillées ayant une bonne précision .

Le sondage utilisera comme base de sondage le « répertoire d'immeubles localisés » (RIL). Ce répertoire est une liste d'adresses (résidentielles, institutionnelles ou commerciales) repérées individuellement de façon à créer une cartographie numérisée où l'adresse est géocodée. Le RIL sera d'abord alimenté par les résultats du RP99 permettant ainsi de décrire statistiquement chaque immeuble résidentiel.

Le RIL sera mis à jour en continu à partir des permis de construire, des permis de démolir, taxe d'habitation, La Poste, etc.), des échanges d'information entre les communes concernées et l'Insee et l'observation directe sur le terrain.

Compte tenu de la contraint budgétaire, le taux global de sondage est tel qu'au terme d'une période de cinq ans 40% des logements de la commune sont enquêtés, soit 8% par an. Les données recueillies dans ces cinq ans sont combinées pour élaborer des résultats valides pour l'année médiane du cycle, extrapolés à l'intégralité des logements de la commune de cette année.

Cette base de sondage pose deux principaux problèmes statistiques pour la production des résultats : les effets de grappe inhérents aux adresses et la qualité des informations annuelles actualisant la base de sondage. Cela conduit à considérer trois strates : les adresses «de grande taille », les adresses « nouvelles » et les « autres » adresses.

2.3 Effets de grappe : le traitement des adresses de grande taille

Le problème majeur est la variance de la taille des unités à échantillonner. En effet, la présence d'une adresse contenant parfois jusqu'à plusieurs dizaines de logements pose un problème d'effet de grappe : les estimations communales et infra-communales pour certaines variables peuvent-être très sensibles à la présence ou non de ces adresses dans l'échantillon. Par exemple, dans le cas d'un échantillon aléatoire simple de grappes, on sait que la variance de l'estimateur du total d'une variable *Y* s'écrit :

$$V(\hat{T}(Y)) = M^2(1-t)\frac{S_g^2}{m}$$

M est le nombre total de grappes, *m* le nombre de grappes de l'échantillon, *t* le taux de sondage et S_g^2 la variance inter-grappes, c'est à dire :

$S_g^2 = \frac{1}{M-1}\sum_i (Y_i - \overline{Y})^2$ où Y_i est le total de la variable dans la grappe *i*.

L'on voit sur cette formule que la variance peut être élevée si la taille des grappes est très disparate.

C'est pourquoi il a été décidé de créer une strate particulière constituée de ces adresses. Cette strate sera enquêtée exhaustivement au cours d'un cycle de 5 ans. Il n'y aura donc pas de composante due à l'échantillonnage dans le calcul de la variance au sein de cette strate pour les estimations détaillées. Cette stratégie a pour avantage principal d'améliorer la précision dans les IRIS2000 contenant des adresses de grande taille. A budget global constant, la contrepartie est un taux d'échantillonnage un peu moindre dans le reste de la base.

Les adresses de la strate "adresses de grande taille" sont réparties en 5 groupes de rotation, en cherchant à égaliser le nombre de logements et à optimiser la répartition spatiale dans la commune. Chacun des groupes est enquêté exhaustivement au cours du cycle.

Des travaux de simulation ont été effectués dans le but de déterminer le meilleur compromis pour fixer le seuil de définition des adresses de grande taille. Les simulations semblent indiquer que le seuil raisonnable qui peut être proposé est celui qui consiste à déclarer comme adresses de grande taille dans une commune donnée celles qui représentent 10% du nombre de logements de la commune, avec un plancher de 60 logements.

L'exemple suivant illustre le gain, en matière de réduction d'effet de grappe, apporté par la stratification. On a simulé trois plans de sondage dans l'arrondissement de Lyon 8 et on a estimé, en particulier la population de l'IRIS 0602 de cet arrondissement. Il s'agit d'estimations simples non redressées. Le premier plan est sans stratification, le deuxième avec une strate d'adresses de grande taille représentant 10% des logements de l'arrondissement, le troisième avec une strate dont le seuil est fixé à 20%. Chaque plan est simulé 500 fois. Dans le premier cas le coefficient de variation apparent de la population estimée de l'IRIS est de 12,5%, il baisse à 4,6% dans le deuxième cas et remonte à 5% dans le dernier. La remontée s'explique par le fait que si l'on privilégie trop les adresses de grande taille, le taux de sondage devient trop faible dans les autres. Le compromis à 10% apparaît assez généralement dans les simulations effectuées sur d'autres communes.

2.4. La prise en compte des modifications annuelles : le traitement des adresses nouvelles

Chaque année, un constat est fait, en concertation avec les communes, de l'évolution du parc des logements. Les immeubles détruits sont naturellement enlevés de la base d'adresses, les constructions sont introduites avec leur nombre de logements supposés (figurant dans les permis de construire). Or les méthodes d'estimation utilisent le critère «nombre de logements » comme variable principale d'extrapolation et il importe donc que l'exactitude de ce critère soit avérée, notamment par les vérifications opérées sur le terrain.

C'est pourquoi les adresses nouvelles d'une année seront enquêtées exhaustivement lors du cycle qui suit. Elles seront ensuite ventilées dans les groupes des « autres » adresses de façon à maintenir les équilibrages pour les critères de référence.

En flux annuel, les adresses nouvelles représentent en moyenne 1% des logements des communes.

2.5 La strate des autres adresses

Les adresses de la strate "autres adresses" sont au départ réparties en 5 groupes de rotation équilibrés. Les critères d'équilibrage sont analogues à ceux qui ont prévalu à la constitution des groupes de rotation des communes de moins de 10 000 habitants, à savoir :

- le nombre de logements ;
- le nombre de logements en immeuble collectif ;
- la population des moins de 20 ans ;
- la population des 20-39 ans ;
- la population des 40-59 ans ;
- la population des 60-74 ans ;
- la population des 75 ans et plus ;
- la population des femmes ;
- la population des hommes.

Les valeurs d'initialisation sont celles du recensement de 1999.

2.6 Les échantillons annuels

Avant chaque collecte annuelle, les cinq groupes de la base de sondage ont donc été mis à jour. Ils comprennent trois strates : les adresses de grande taille, les adresses nouvelles et les autres adresses. Toutes les adresses des deux premières strates sont enquêtées exhaustivement et un échantillon aléatoire est prélevé dans la troisième. Pour ce tirage de deuxième phase, on introduit comme critères d'équilibrage le nombre de logements de la strate, le nombre de logements collectifs, le poids des IRIS en nombre de logements. Le taux de sondage d'une année est ajusté de sorte que la proportion de logements enquêtés soit égal à 40% du groupe.

On introduit toutefois une « clause de sauvegarde » pour la strate des autres adresses. En effet, dans quelques communes, le poids des adresses nouvelles alourdi par le stock 1999-2003 peut conduire à un taux de sondage trop faible dans la strate des autres adresses et, donc, détériorer la représentativité globale de certains IRIS. L'analyse menée à partir des tendances observées entre 1990 et 1999 conduit à préconiser que le taux de sondage pratiqué dans le groupe des autres adresses d'une année donnée ne devrait pas être inférieur à 25%. Si cela devait être le cas, on ajusterait le taux en renonçant à l'exhaustivité des adresses nouvelles devant être enquêtées cette année-là. Le nombre de communes concernées par cet ajustement est estimé à moins de trente. Il s'agit de communes dont la population est entre 10 000 et 30 000 habitants, situées dans la couronne de grandes agglomérations.

Bibliographie

[1] Dumais J., Isnard M., « Le sondage de logements dans les grandes communes dans le cadre du Recensement rénové de la population », *Actes des VIIe Journées de Méthodologie Statistique,* Paris, 4 et 5 décembre 2000, Tome 1, pp 37-50, INSEE.

[2] Dumais J., Bertrand Ph., Kauffmann B., « Sondage, estimation et précision dans la rénovation de recensement de la population », *Actes des VIIe Journées de Méthodologie Statistique,* Tome 1, pp 51-75, INSEE

[3] Dumais J., « Quelques aspects méthodologiques du recensement rénové de la population en France », in *Enquêtes, modèles et applications,* pp. 467-479, 2001