La rénovation du recensement français : principes et méthode

Jean-Michel DURR
Direction générale de l'Institut national de la statistique et des études économiques
18 boulevard Adolphe Pinard
75675 PARIS Cédex 14
jean-michel.durr@insee.fr

Jean-Marie GROSBRAS
Direction générale de l'Institut national de la statistique et des études économiques
18 boulevard Adolphe Pinard
75675 PARIS Cédex 14
jean-marie.grosbras@insee.fr

1. Les raisons de la rénovation

Le recensement général de la population est conduit en France de manière régulière depuis 1801 avec pour objectifs de déterminer la population légale de chacune de ses circonscriptions administratives et de décrire les caractéristiques démographiques et sociales du territoire à tous les niveaux géographiques, des quartiers des communes au pays dans son ensemble. Ainsi, le 32e recensement s'est déroulé en 1999 selon le schéma habituel. Toutefois, certains éléments ont conduit la France à revoir ce dispositif. Tout d'abord, l'intervalle intercensitaire a eu tendance à s'allonger et on est passé de recensements quinquennaux avant la guerre à des écarts entre les recensements de 7, puis 8 ans. Le dernier recensement, prévu initialement en 1997, a été repoussé à 1999, soit 9 ans après le précédent, pour des raisons budgétaires. De plus, le mouvement de décentralisation que connaît la France depuis plus de vingt ans a généré de nombreux besoins de données statistiques afin d'éclairer les politiques locales. Le recensement, source d'information locale par excellence, doit donc s'adapter et fournir des données plus fraîches et toujours finement localisées.

C'est pourquoi un programme de rénovation du recensement de la population a été engagé à l'Insee dès la fin des années 90, sous la contrainte budgétaire de ne pas accroître le coût le l'opération. La France ne disposant pas de registre de population et le contexte national ne s'y prêtant pas, il a été décidé d'emprunter une voie originale basée sur l'idée de « recensement tournant » proposée par Leslie Kish dans les années 80 [Kish, L., 1981 et 1990]. Le principe en est le suivant : pour les communes dont la population est inférieure au seuil de 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 étant prise en compte au terme de la même période de cinq ans. Chaque année à partir de 2004, les enquêtes auront lieu en janvier et février.

Pour mener à bien cette rénovation, un cadre juridique nouveau s'est avéré nécessaire en raison des implications légales du recensement. Ainsi, la loi « démocratie de proximité » du 27 février 2002 décrit dans son titre V le principe et les nouvelles modalités de réalisation des opérations de recensement.

2. Un double plan de sondage

Considérons d'abord le domaine des communes de moins de 10 000 habitants, qui accueillent la moitié de la population française. Dans chaque région, cinq groupes de rotation de communes seront formés. La méthode statistique utilisée est celle des échantillons équilibrés [Deville, Tillé,

(2000)]. 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.

Pour équilibrer les groupes de rotation des communes de moins de 10 000 habitants dans chacune des régions, l'équilibrage est effectué sur la distribution par âge, par sexe, par type de logement (individuel ou collectif), et sur le nombre de logements par département.

La prise en compte de la répartition par type de logement (individuel ou collectif) a une influence sur la répartition des plus grandes de ces communes dans les groupes de rotation. Elle 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.

Le problème est différent pour les communes de 10 000 habitants ou plus, dans lesquelles un échantillon doit être enquêté tous les ans. Il s'agit d'un plan de sondage "à l'adresse", tous les logements d'une même adresse échantillonnée étant enquêtés pour des raisons de robustesse du mode opératoire.

La base de sondage est 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 a été initialisé par les adresses du recensement de 1999, permettant ainsi de décrire statistiquement chaque immeuble résidentiel. Il est mis à jour en continu à partir de sources administratives, des échanges d'information entre les communes concernées et l'Insee et de l'observation directe sur le terrain.

Compte tenu de la contrainte 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, puis 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.

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 peuvent être très sensibles pour certaines variables à la présence ou non de ces adresses dans l'échantillon. C'est pourquoi il a été décidé de créer une strate particulière constituée de ces adresses. Après simulations, elle a été déterminée comme comprenant les adresses les plus grandes représentant 10% des logements de la commune. Cette strate sera enquêtée exhaustivement au cours d'un cycle de cinq 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 quartiers 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.

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 supposé (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 ventilées année après année dans les groupes des « autres » adresses de façon à maintenir les équilibrages pour les critères de référence.

Les adresses de la strate "autres adresses" sont au départ réparties en cinq 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.

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 quartiers 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.

3. Les mécanismes d'estimation

Dès la fin du premier cycle quinquennal des enquêtes de recensement et en régime de croisière, l'Insee publiera chaque année des résultats statistiques détaillés aux niveaux communal et infra-communal. Le principe est que les données publiées l'année A sont issues d'estimation à valeur pour l'année A-2. Par exemple, pour une commune de 10 000 habitants ou plus, les données publiées en A incorporent les résultats de cinq collectes annuelles successives, de A-4 à A, agrégées pour produire des estimations à l'année médiane de la période de cinq ans.

Les méthodes mises en œuvre diffèrent selon les strates auxquelles appartiennent les communes, c'est-à-dire selon qu'elles ont moins de 10 000 habitants ou plus.

Pour les communes de moins de 10 000 habitants, l'idée la plus simple est d'utiliser la tendance observée pour une commune aux recensements les plus proches la concernant. Ainsi, la population A-2 publiée l'année A sera établie selon le groupe auquel appartient la commune, c'est-à-dire selon qu'elle a été recensée en A-4, A-3, A-2, A-1 ou A. La règle est alors la suivante :

Tableau 1.

Dates de recensement	**Action pour estimer l'année A-2**
A-9, **A-4**	On extrapole à A-2 la droite (A-9 – A-4)
A-8, **A-3**	On extrapole à A-2 la droite (A-8 – A-3)
A-7, **A-2**	On garde le recensement A-2
A-6, **A-1**	On interpole A-2 sur la droite (A-6 – A-1)
A-5, **A**	On interpole A-2 sur la droite (A-5 – A)

On voit ainsi que l' « horizon » des extra-interpolations est au maximum de deux ans. Pour le démarrage des estimations, c'est-à-dire en 2008 pour des estimations en 2006, le point de départ des extra-interpolations sera le recensement de 1999. Cette méthode de base présente l'avantage de fonctionner en l'absence de toute autre information que celle fournie par les enquêtes de recensement.

Une approche complémentaire est d'utiliser, pour améliorer les extrapolations, les évolutions du nombre de logement constatées dans les fichiers de la taxe locale d'habitation.

Il reste à constituer le fichier de micro-données relatif à l'année A-2 destiné aux tabulations portant sur une partie quelconque du territoire. Il sera constitué des unités statistiques enquêtées au cours des années A-5 à A, affectées d'une pondération destinée à les amener à leur poids en A-2. Prenons, par exemple, le cas d'une commune recensée en A, avec une population de 105, alors qu'elle avait une population de 100 en A-5. Par interpolation, la population en A-2 est donc estimée à 103. Les unités seront pondérées par le coefficient 103/105 = 0,98.

Pour illustrer la méthode et analyser ses risques, on a procédé à des simulations basées sur des jeux d'hypothèses d'évolution entre les deux derniers recensements. Dans la grande majorité des cas, les procédures d'estimation donneront des résultats «raisonnables ». Les cas les plus fragiles sont ceux où des événements importants se produisent peu après les recensements. En ce cas les estimations de base mettent deux ans à les incorporer, ce qui est néanmoins plus satisfaisant que dans le système de recensements traditionnels où le délai de prise en compte était beaucoup plus long. C'est aussi dans ces cas que l'amélioration des estimations à l'aide d'évolutions constatées sur des sources administrative (taxe d'habitation par exemple) pourra être bénéfique.

Pour les communes de 10 000 habitants ou plus, il s'agit de consolider cinq enquêtes de recensement successives, de A-4 à A pour produire des résultats millésimés à l'année médiane A-2. La difficulté vient de ce que, comme on l'a vu ci-dessus, l'enquête d'une année donnée s'exécute avec une base de sondage actualisée par rapport à l'année précédente, c'est-à-dire intégrant la démographie des logements. Une méthode simple est de rassembler, pour A-2, les données annuelles récoltées pendant la période, en les affectant de leur poids de sondage. Dans la réalité, les coefficients dépendront du poids respectifs des strates. En effet, la strate des adresses de grande taille ne représente pas, en général, exactement 10% des logements. Ils dépendront aussi des calages éventuellement pratiqués, comme par exemple sur le nombre de logements présents au moment de l'enquête A-2. Pour avoir des statistiques plus fiables à l'infra-communal, l'ajustement pourra se faire pour chaque quartier.

Pour avoir une approche de la précision des résultats des sondages, on a procédé aux simulations suivantes sur quelques communes, à partir du recensement de 1999. On constitue les strates d'adresses puis les groupes de rotation, puis on tire cinq échantillons successifs. L'opération est répétée 500 fois et on analyse, pour un certain nombre de variables-témoin, la répartition des résultats. On obtient une précision moyenne de l'ordre de 6% pour un effectif de l'ordre de 500, de 4% à partir d'un effectif de 1 000, de 3% à partir d'un effectif de 2 000 et de l'ordre de 2% à partir de 5 000.

En conclusion, même si le nouveau recensement ne produira pas des données d'une précision égale à celle qu'apportaient les recensements généraux une fois tous les huit ans, il permettra de suivre, avec une bonne précision à tous les niveaux géographiques, les évolutions de la population et de ses caractéristiques démographiques et sociales. Les utilisateurs devront cependant s'habituer à travailler sur des données représentant des situations moyennes et non plus strictement datées. Une réflexion avec des utilisateurs est en cours pour mesurer les nouvelles utilisations permises par le nouveau recensement.

REFERENCES

Deville, J.C., Tillé, Y. (2000), « Echantillonnage équilibré par la méthode du cube et estimation de variance » , Journées de méthodologie, décembre 2000, INSEE, Paris.

Dumais, J, Durr, J.-M. (2001), « La rénovation du recensement français », Actes du symposium 2001 de Statistique Canada.

Kish, L. (1990), « Recensement par étapes et échantillons avec renouvellement complet », Techniques d'enquêtes, Vol 16, N° 1, pp. 67-86, Statistique Canada, Ottawa, juin 1990.

RÉSUMÉ

In order to answer the increasing demand for fresher data and spread the burden of conducting the census over a longer period, the French National Institute of Statistics and Economic Studies has conducted a large program to redesign the census, based on the "rolling census concept" proposed by L. KISH. In this rolling census, every commune under the threshold of 10 000 inhabitants will be surveyed once within a five year period ; larger communes will be divided into five rotation groups of addresses, each rotation group being surveyed in one of the five years. This paper presents the principles and the methodology of this operation.