JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

BENOÎT RIANDEY

La rénovation du recensement de la population. La précision des données collectées sur une année

Journal de la société française de statistique, tome 140, n° 4 (1999), p. 41-47

<http://www.numdam.org/item?id=JSFS_1999__140_4_41_0>

© Société française de statistique, 1999, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (http://publications-sfds.math.cnrs.fr/index.php/J-SFdS) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

LA RÉNOVATION DU RECENSEMENT DE LA POPULATION

La précision des données collectées sur une année

Benoît RIANDEY *

UN MICRO-RECENSEMENT FRANÇAIS PUISSANT ET PARTICULIER

Jean Dumais et Anne-Marie Dussaix se sont exprimés très clairement sur le sondage qui se substituera au recensement dans les communes de plus de 10 000 habitants : ce sondage stratifié par région et équilibré sur la pyramide des âges de 1999 constitue une sorte de *micro-recensement* très optimisé au taux de 8 %. Il est beaucoup plus volumineux que le micro-recensement allemand réalisé au taux de 1 %. L'efficacité du sondage équilibré [1] est spectaculaire, en particulier pour une variable apparemment aussi peu liée à l'âge que le nombre de véhicules du ménage.

Cette opération est complétée dans chaque région par un sondage au $1/5^{ème}$ des communes de moins de 10 000 habitants, également équilibré sur la pyramide des âges de 1999, en définitive un *recensement tournant* sur cinq ans.

Associés, ces deux sondages annuels constituent chaque année un micro-recensement puissant et très particulier car les fortes grappes communales contrastent avec le maillage très fin des unités primaires des grandes communes.

Les synthèses par subtile agrégation des données sur cinq ans constituent le *recensement rénové.* Sa précision en est fortement consolidée, point que je n'évoquerai guère car la question qui m'est posée est relative à la qualité de la brique annuelle.

L'intérêt de cette question ne tient pas seulement à son effet sur la précision finale des données agrégées. Les données annuelles fourniront isolément des estimations statistiques potentiellement sans biais. Elles permettront de calculer des évolutions entre deux années, quoique moins efficacement que les modèles de séries chronologiques sur série plus longue. Elles pourront servir

* INED, 133 boulevard Davout, 75980 Paris cedex 20 ; riandey@ined.fr

1. Voir travaux de Deville et Tillé (1999, 2000).

de première phase d'enquête sur des thèmes spécialisés : sous-populations, migrations, transport urbain.

La variance très petite des estimations pour la France entière ou pour les régions doit être saluée sans nécessiter d'autres développements. Par contre, l'ambition de l'INSEE de fournir des estimations pour des domaines d'un million d'habitants appelle quelques commentaires. On aurait pu prendre pour exemples les agglomérations de Lyon, Marseille et Lille qui dans l'échantillon-maître de l'INSEE constituent des *strates* mais seulement des domaines de dimension souhaitée dans le *recensement rénové.* Le sondage parmi ces énormes grappes communales, pouvant approcher 10 000 habitants, n'y est pas équilibré sur la pyramide. Mais cet effet de grappe disparaît dans la synthèse sur cinq ans puisqu'elle englobe l'ensemble du recensement tournant. Avec seulement 391 000 habitants, le département de la Drôme illustre ces effets : les estimations démographiques annuelles sont entachées d'une erreur relative de sondage de 15 % et celles de la synthèse d'à peine 0,2 % [2]. Cette amélioration d'un facteur 75 tient pour un facteur $\sqrt{5}$ à l'accroissement de l'échantillon sur cinq vagues et à un facteur 33 à la neutralisation de cet effet de grappe. Ainsi les données annuelles ne constituent probablement pas une source excellente pour les zones infra-régionales. En particulier, une base de sondage moins riche comprenant les autres petites communes devra être utilisée, hormis pour les enquêtes nationales.

LA QUESTION DU BIAIS

L'avantage des grandes communes en matière de variance s'inverse dès qu'on s'intéresse au risque de biais : les communes de moins de 10 000 habitants connaîtront pour seul changement l'atténuation du rendez-vous citoyen tandis que l'étalement de la collecte sur 5 ans facilitera son contrôle. Toutefois, l'orientation actuelle en faveur d'une unique période de collecte annuelle, certes lourde pour l'INSEE, facilitera la campagne de communication et l'indispensable obtention de taux de réponse très élevés, quoiqu'elle rende improbable la constitution d'équipes d'agents recenseurs professionnels. Elle résout les questions qu'en l'absence d'un échantillonnage temporel susciterait l'étalement de la collecte au cours de l'année quant à la signification d'un état au premier janvier.

De ce fait, mon propos concernera maintenant essentiellement la qualité des données annuelles dans les communes de plus de 10 000 habitants.

Cette question est essentielle car, contrairement à la variance, les biais ne s'amenuisent pas par agrégation jusqu'au niveau national. L'excellence du mode de sondage en terme de variance n'exonère aucunement les responsables du recensement de leur recherche du niveau zéro des biais. Cette question appelle des réponses multiples par catégorie de variables en distinguant plusieurs échelles géographiques : effectif de la population légale, pyramide

2. Non compris l'erreur liée aux opérations d'imputation dans la synthèse.

des âges, taille des ménages, catégories sociales, nationalité, navettes quotidiennes, migrations intérieures et extérieures, solde migratoire. Ces variables sont inégalement affectées par la rénovation. J'évoquerai aussi les données administratives susceptibles de compenser la disparition de données censitaires au niveau le plus fin.

De fait, l'imprévisibilité actuelle de la qualité du recensement futur tient à celle du comportement du citoyen sondé, à l'efficacité de l'appareil de collecte et d'analyse, à la mobilisation et à la qualité des fichiers administratifs qui étaieront les estimations. A long terme, la précision des estimations s'appuie aussi sur la robustesse d'un système quand même assez complexe.

Certainement, les variables et les zones fragiles au recensement le demeureront largement dans le futur, même si le projet s'efforce d'y remédier dans les grandes communes. Pour cette raison, je partirai des fragilités repérées dans la dernière enquête de contrôle du recensement, celle de 1990 [3], ou dans l'enquête Emploi de l'INSEE. Une tabulation de l'enquête de contrôle sur les communes de plus de 10 000 habitants fournirait une information précieuse sur les taux d'omission élevés dans les grandes communes. Toutefois, le marbre de la population légale ne dissuade-t-il pas un peu de lever le capot comme dans toute bonne enquête par sondage ?

Attachons-nous d'abord à **la qualité différentielle par variable.**

Le Comité Scientifique du *recensement rénové* avait estimé que la **population légale** des communes et autres unités administratives aurait une précision suffisante si le mode de sondage et la communication suscitaient un taux de réponse satisfaisant.

Les dénombrements de ménages fournis par le fichier de la taxe d'habitation et ceux des individus issus du futur fichier de l'assurance maladie (universelle) consolideront les estimations communales. Dans leur état actuel, les fichiers des caisses primaires d'assurance maladie comprennent une proportion élevée d'adresses d'assurés et d'ayant droits ayant migré vers une autre caisse primaire [4] (un peu moins de 30 % selon l'enquête CNAMTS-CREDES sur la protection sociale). L'instauration du Répertoire Inter-régimes de l'Assurance Maladie (RNIAM) conduira, à long terme, à un apurement de ces fichiers décentralisés d'individus. La généralisation de l'enquête CNAMTS-CREDES à tous les régimes sur la base des individus (et non des assurés) est appelée par les démographes comme outil de suivi du recensement. Toutefois l'élimination progressive des doublons n'est pas favorable à l'usage de ces fichiers pour un lissage temporel des estimations.

En l'absence de tout biais de collecte, les données annuelles du *recensement rénové* produiront une **pyramide des âges** parfaite. Mais, paradoxalement, le

3. Coeffic N. « L'enquête post-censitaire de 1990. Une mesure de l'exhaustivité du recensement ». Population 6 1993, pp. 1655-1682.
Coeffic N. « L'enquête de mesure du degré d'exhaustivité du recensement de 1990 » in « Le recensement de population de 1990 : Innovation méthodologiques », INSEE Méthodes n°52-53, 1995, pp. 60-148.

4. Car sans incidence sur la gestion des prestations. Ces dénombrements localisés sont bien sûr anonymes.

futur recensement pourrait fournir une excellente pyramide des âges nationale presque indépendamment de la collecte, grâce au dispositif universel de l'assurance maladie : le sexe et l'âge de tout résident figureront au futur répertoire sans doubles comptes de l'assurance maladie. Le *recensement rénové* constitue un sondage de très grande ampleur pour corriger l'effet de l'émigration définitive et la présence « d'immortels » dans ces fichiers. Par contre, les migrations intérieures perturberont la ventilation de la pyramide nationale vers les niveaux territoriaux fins (d'une taille inférieure au seuil communal de 10 000 habitants). Cette information sur la répartition par âge sera probablement défaillante au niveau infra-communal où elle est si utile pour élaborer la carte scolaire.

Les 20-24 ans sont des **âges** difficiles à recenser ou enquêter, comme en témoignent les données brutes de l'excellente enquête Emploi annuelle. Si nécessaire, les statistiques de l'assurance maladie permettront d'en réévaluer le nombre, mais avec des risques de biais pénalisant les jeunes vivant seuls, en collectivité ou en sous-location.

Les personnes âgées sont souvent sous-représentées dans les échantillons à cause de leur crainte d'ouvrir leur porte, et éventuellement pour raison de santé ou de multi-résidence. La collecte aréolaire créerait une synergie favorable à la résorption du refus des personnes âgées et au contact avec les personnes seules peu présentes au domicile, du moins si l'enquête est attrayante. Optimisation du sondage et qualité de la collecte sont antinomiques sur ce point.

Aucune autre source ne permet de redresser la **distribution de la taille des ménages**. Si sur ce point le recensement est fragile, toutes les enquêtes par sondage françaises seront biaisées à cet égard, mais personne ne sera plus en mesure de le constater. En pur artefact, la proportion de jeunes vivant seuls serait déflatée au profit de celles des jeunes non émancipés ou vivant en couple.

Le recensement étalé dans le temps abandonne la technique du double enregistrement des étudiants introduite au recensement de 1999 car elle repose sur l'exhaustivité. Cette innovation était la meilleure réponse au vieux problème de la population comptée à part. On voit mal quels fichiers administratifs viendraient compenser la fragilité de leur estimation directe. En particulier le dénombrement administratif des étudiants est largement sujet à caution car assez partiel et truffé de doubles comptes (6 % selon le CNIS), faute d'un bon identifiant.

La connaissance des multi-résidences ne peut reposer seulement sur un bon questionnaire. Les biais de sondage sont redoutables à cet égard. L'enquête de contrôle du recensement de 1990[2] contient à ce sujet une information qui serait très utile à la rénovation du recensement, mais qui – à ma connaissance – n'a pas été exploitée. Ce serait d'ailleurs un excellent sujet de thèse. La définition conventionnelle de la *résidence principale* est source de difficultés nouvelles que le sondage traite plus difficilement que la collecte exhaustive.

Le dénombrement annuel des collectivités est une excellente intention. C'est une opération difficile. Effet de grappe, erreur de double compte et procédure téléphonique feront l'objet d'une évaluation en une autre occasion.

Le dénombrement des résidents **étrangers** ou immigrés (selon qu'on considère le critère de la nationalité actuelle ou celui du pays de naissance) constitue un point faible des recensements [5] et des enquêtes. La situation actuelle est confuse car la sous-population de référence est ambiguë : l'ensemble des résidents étrangers irréguliers échappent-ils au recensement ? Le recensement rénové ne se traduira-t-il pas par une baisse du taux de réponse des étrangers ? L'amélioration des sources administratives relatives aux entrées d'étrangers (en situation régulière) et aux titres de séjour en vigueur nous renseignera mieux sur les flux d'entrée que sur le stock encore présent. A moyen terme, il y a peu d'espoir de statistique de calage pour cette sous-population.

Les **catégories indépendantes** non agricoles (artisans, commerçants) répondent moins bien à beaucoup d'enquêtes qu'au recensement. Si leur taux de réponse au recensement rénové est médiocre, leur sous-représentation devrait être sensible, tandis que la mobilisation de dénombrements d'indépendants semble peu réaliste.

A un niveau géographique fin, les géographes perdront leur outil d'observation de **la structure sociale des quartiers** au moment où la codification automatique des CSP pourrait rendre plus accessible une caractérisation sociale exhaustive des ménages. Aucun fichier exhaustif ne couvrira l'ensemble du champ social. Plus généralement, Anne-Marie Dussaix a certainement raison de penser que l'efficacité remarquable de l'équilibrage du sondage sur la pyramide devrait avoir peu d'effet sur le calage de la structure sociale. Ce semble être un point plus faible du projet tant au niveau des petites unités que des grappes de communes.

Le recensement ne sera bientôt plus la source d'estimations localisées sur le **chômage**, puisque les statistiques de l'ANPE seront prochainement disponibles à des niveaux assez fins. Sans davantage approcher le concept du chômage au sens du BIT, ce sera un progrès car, au recensement de 1990, les chômeurs présentaient un taux d'omission double[3] de celui observé pour l'ensemble de la population adulte (erreur de couverture), tandis que le chômage était sur-déclaré de 280 000 personnes [6] par rapport à l'enquête Emploi (erreur de mesure).

Les recensements fournissent une information unique, mais imprécise sur les **soldes migratoires** et les **migrations intérieures** [7]. Brigitte Baccaïni [8] annonce d'emblée que tout cela est à revoir : le changement de résidence

5. Rouault D., Thave S, 1997, « L'estimation du nombre d'immigrés et d'enfants d'immigrés », INSEE Méthodes n°66.

6. Rouault-Galdo D., « Les écarts d'estimation de la population active française au recensement et à l'enquête annuelle sur l'emploi. D'où viennent les divergences ? », Population 6, 1993, pp. 1683-1704.

7. Voir plus précisément : Courgeau D., « Migrants et migrations », Population 1, 1973, pp. 95-129.

8. Baccaïni B. « Analyse des migrations internes et estimation du solde migratoire externe au niveau local à l'aide des données censitaires », Population 4-5, 1999, pp. 801-815.

depuis le recensement antérieur n'a plus de sens pour une opération de collecte étalée. En revanche, les deux futures questions mesurant la mobilité depuis un an et depuis 5 ans satisfont les démographes. En 1999, la description des champs migratoires bénéficie du codage automatique exhaustif des communes et non plus du seul sondage au quart, mais cette embellie n'aura été qu'un flash puisque le *recensement rénové* reposera sur un sondage à 8 ou 20 %. La synthèse des vagues successives ouvre des possibilités complexes à élucider.

Les données annuelles en A et $A + 5$ permettront de calculer un solde migratoire externe sur 5 ans ; mais, les deux stocks sont estimés par sondage, à un taux proche de 8 % (celui des grandes communes). Cette estimation par la différence subira donc la fluctuation d'échantillonnage et le biais de non-réponse probablement accru des étrangers. C'est une bonne raison pour abandonner l'usage de soldes migratoires externes, d'une précision passée contestable, et investir sur l'amélioration des sources de sécurité sociale. Cette critique vaut pour les soldes migratoires internes. De multiples sources administratives décentralisées sont perturbées par les migrations, faute d'un outil de suivi, au lieu d'en fournir une mesure !

La mesure des **navettes quotidiennes** pâtira de l'étalement de la collecte entre communes car les navettes concernent très souvent des ensembles inter-communaux (agglomérations, communautés urbaines, bassins d'emploi) recensés au cours d'années différentes. C'est généralement le cas des navettes observées au lieu de travail. Le recensement rénové ne fournira donc pas de données annuelles à ce niveau, tout au plus des agrégations un peu insolites de vagues successives. Cette lacune est partiellement compensée par la richesse du fichier administratif des Déclarations annuelles des données sociales : en plus du salaire, de la catégorie socioprofessionnelle de chaque salarié etc., il fournit ses adresses de travail et résidence, potentiellement géocodées à l'îlot. L'exemple des navettes illustre un constat : l'**inter-communalité** est la grande perdante du *recensement rénové.*

CONCLUSION

Comme tout sondage, le *recensement rénové* fournira sur les grandes communes des estimations annuelles de précision décroissante avec la taille de l'aire concernée et excellemment présentée par le texte de Dumais, Bertrand et Kauffmann. Au niveau infra-communal, on regrettera vite l'exhaustivité du recensement pour décrire les micro-quartiers et on explorera les fichiers administratifs disponibles. Souvent, ils ne traiteront que d'une fraction de la population : les salariés, les allocataires du régime général de sécurité sociale. La première qualité du recensement est d'être général !

La pertinence de ces appréciations qualitatives dépendra des taux de réponse obtenus par le futur recensement. Différents scénarios se présentent à l'esprit :

S0 : recensement de 1990 : omissions estimées 1,8 % (et 0,7 % de doubles comptes) ;

S1 : micro-recensement allemand : taux d'échec 3 % ;

S2 : premier passage de l'enquête Emploi (aréolaire) : taux d'échec 10 % ;

S3 : enquête ménage obligatoire de l'INSEE bien accueillie : taux d'échec environ 15 % ;

S4 : enquête ménage obligatoire de l'INSEE moins bien accueillie : taux d'échec bien supérieur à 15 %.

QUEL SCÉNARIO SE RÉALISERA ? QUEL SCÉNARIO EST ACCEPTABLE ?

Quel effet aura-t-il, non seulement sur la population légale et la pyramide des âges, indicateurs assis sur des fichiers administratifs, mais aussi sur l'effectif des ménages d'une seule personne, celui d'une catégorie sociale, celui des étudiants ou des étrangers ?

Il est difficile à un statisticien seul de définir à partir de quel seuil de non réponse (5 %, 7 % ?) on devrait tirer la sonnette d'alarme. Ce seuil différerait selon les objectifs, et le Comité Scientifique ne s'est pas exprimé sur le sujet. D'ailleurs, si on l'avait fait dans les années 50, on n'oserait plus réaliser d'enquêtes par sondage, du fait de la dégradation des taux de réponse.

Plus exact dans sa dénomination, plus clair dans sa présentation au public, ce micro-recensement annuel obligatoire devra faire ses preuves sur le terrain pour que s'impose le dispositif novateur complexe brillamment imaginé par nos collègues et le dépérissement de notre recensement bicentenaire. L'option d'une collecte concentrée sur le début de l'année est un élément très rassurant en ce sens. L'apurement des adresses du fichier de l'assurance maladie demeure l'autre clé du dispositif. Ces incertitudes empêchent d'apporter aujourd'hui une réponse autre que théorique à la question posée sur la précision des données annuelles.