VERS UNE UTILISATION APPROFONDIE DES SOURCES AUXILIAIRES AU RECENSEMENT DE LA POPULATION ?

Sébastien Hallépée[1]

[1] *Institut National de la Statistique et des Études Économiques (Insee), 18, boulevard Adolphe Pinard – 75675 Paris Cedex 14, sebastien.hallepee@insee.fr*

Résumé.

Le recensement rénové a introduit une collecte en continu, distincte pour les communes de moins de 10 000 habitants et de plus de 10 000 habitants. Pour ces premières, la collecte reste exhaustive. L'innovation majeure réside en l'utilisation de fichiers fiscaux pour extrapoler les populations dont le recensement est le plus ancien. Pour les grandes communes, le recours à un échantillon pour établir les résultats du recensement est la nouveauté la plus marquante du nouveau processus. Nous allons explorer comment l'utilisation de sources auxiliaires permettra d'améliorer la qualité des estimations pour ces grandes communes. Le repérage des hébergements touristiques permettant d'isoler ces structures atypiques dans le processus d'estimation est un premier pas mis en place pour le RP 2014. Il pourra être prolongé par l'utilisation plus poussée de sources auxiliaires dans les traitements conduisant aux estimations du RP.

Mots-clés. Recensement, Calage, Point atypique, Strate d'estimation, Population légales.

1. La méthode actuelle du recensement

Le recensement de la population français s'appuie sur une nouvelle méthode qui a été initiée à partir d'un mode de collecte dont la première enquête s'est déroulée en 2004. Les opérations du recensement sont réalisées sur la base d'un partenariat entre les communes et l'Insee. L'ensemble du dispositif est ainsi construit autour de la commune, premier partenaire et utilisatrice des estimations qui en découlent. Une des missions principales du recensement est de fournir des populations légales pour les circonscriptions administratives et collectivités territoriales. L'enjeu de la détermination de cette population est important, car de nombreux textes de lois s'y réfèrent et son niveau entre notamment en compte dans le calcul de la dotation versée par l'État aux communes.

Le mode de collecte et les méthodes de calculs du recensement différent selon la taille de la commune. Pour les communes de moins de 10 000 habitants, un recensement exhaustif a été conservé. Il se déroule une fois tous les cinq ans, les communes étant réparties en cinq groupes déterminant leur date de collecte. Pour les communes de 10 000 habitants et plus (appelées par la suite grandes commune ou GC), on enquête tous les ans 8 % des adresses de chaque commune. Les résultats du RP d'un millésime N donné seront construits à partir du cumul de cinq enquêtes annuelles de recensement (EAR) rentrant dans la fenêtre d'estimation ($N-2$ à $N+2$). L'échantillon est tiré du répertoire d'immeuble localisé (RIL) mis à jour conjointement par l'Insee et chacune des grandes communes.

Les utilisateurs du recensement, au premier rang desquels les élus locaux, attendent du recensement une grande précision. Le recours à un échantillon pour fournir ces résultats n'allait donc pas de soi et reste toujours une source d'interrogation des utilisateurs du recensement.

Le plan de sondage mis en place et détaillé dans l'article de Bertrand, Chauvet, Christian et Grosbras (2002) répond en partie à cet enjeu. Chaque année 8 % des logements de chaque grande commune sont recensés. L'unité d'échantillonnage est l'adresse, ce qui réduit les ambiguïtés sur le

terrain pour que les agents recenseurs enquêtent effectivement les unités attendues. En contrepartie, la taille de ces unités peut être très conséquente. L'introduction d'une strate exhaustive pour les grandes adresses limite l'ampleur de l'effet de grappes. L'interrogation exhaustive des adresses nouvelles, permet que leur qualité temporairement moindre dans le RIL, notamment pour le nombre de logements, n'ait qu'un effet limité sur les estimations. Enfin, le recours à un sondage équilibré sur les petites adresses connues permet de tirer parti d'informations auxiliaires.

La qualité des estimations est aussi obtenue grâce à la méthode de pondération explicitée dans le document de Godinot (2005). Un calage du cumul de cinq EAR successives, représentant environ 40 % des logements de la commune, est réalisé sur le nombre de logements par Iris connu exhaustivement dans le RIL.

L'innovation la plus marquante introduite en 2004 est l'introduction de l'échantillonnage pour les communes de plus de 10 000 habitants, rendant possible le passage à un recensement en continu. Cependant, le recours au fichier de la taxe d'habitation pour extrapoler les estimations de populations de 40 % des petites communes dont la dernière collecte de recensement est la plus ancienne est une autre innovation majeure introduite dans le nouveau processus. Passée plus inaperçue, Lermechin (2016) montre qu'elle permet pourtant d'assurer une égalité de traitement entre toutes les communes en fournissant les populations légales pour chaque commune relative au même millésime.

Pour les grandes communes, les méthodes d'estimation mises en place n'incluent pas l'utilisation de sources externes. Les estimations communales sont relativement robustes en se basant uniquement sur des informations du seul processus du recensement. Néanmoins, la taille relativement réduite de la commune, l'insuffisance de la normalisation de l'adressage, un profil touristique, l'existence d'adresses dont le nombre moyen de personnes par logement est atypique sont autant de facteurs qui peuvent fragiliser les estimations.

Un traitement par winsorisation, mis en place dès les premières estimations du RP permet de limiter l'impact de certaines adresses rendant les estimations instables au niveau communal mais ne fait pas l'objet du présent article. En revanche nous allons développer comment un traitement spécifique des hébergements touristiques annule leur influence et comment une approche plus globale intégrant plus systématiquement une information auxiliaire pourrait réduire l'impact d'adresses dont les caractéristiques s'écartent de la moyenne de son quartier.

2. Réduire l'influence des hébergements touristiques

En plus d'apporter un éclairage sur la population, le recensement doit fournir un dénombrement des logements de chaque commune. À ce titre, les hébergements touristiques ont un profil particulier puisqu'ils peuvent compter de nombreux logements sans qu'aucune population ne lui soit associée. Le développement récent de résidences hôtelières comme les appart hôtels rend ce problème plus prégnant. Ces structures peuvent avoir une influence importante sur les évolutions des grandeurs estimées dans la commune, entraînant un choc à la hausse du nombre de logements accompagné d'un choc à la baisse du nombre de personnes par logement qui peuvent survenir de manière asynchrone et qu'on ne pouvait corriger qu'a posteriori.

Depuis 2014, un repérage systématique des hôtels, campings et résidences hôtelières a été entrepris dans l'ensemble des RIL des grandes communes. Elle s'est finalisée en 2015 pour préparer l'enquête de 2016. Cette mise à jour de grande ampleur permet la mise en place de traitement particulier sur ces structures.

En matière d'échantillonnage, elles ont été isolées dans une strate exhaustive. L'occupation de ces structures est très différente des adresses d'habitation classique. L'exclusion de ces structures de la procédure d'échantillonnage rend les adresses restantes plus homogènes et réduit l'aléa d'échantillonnage. Cette opération ne sera effective de manière complète qu'après un cycle

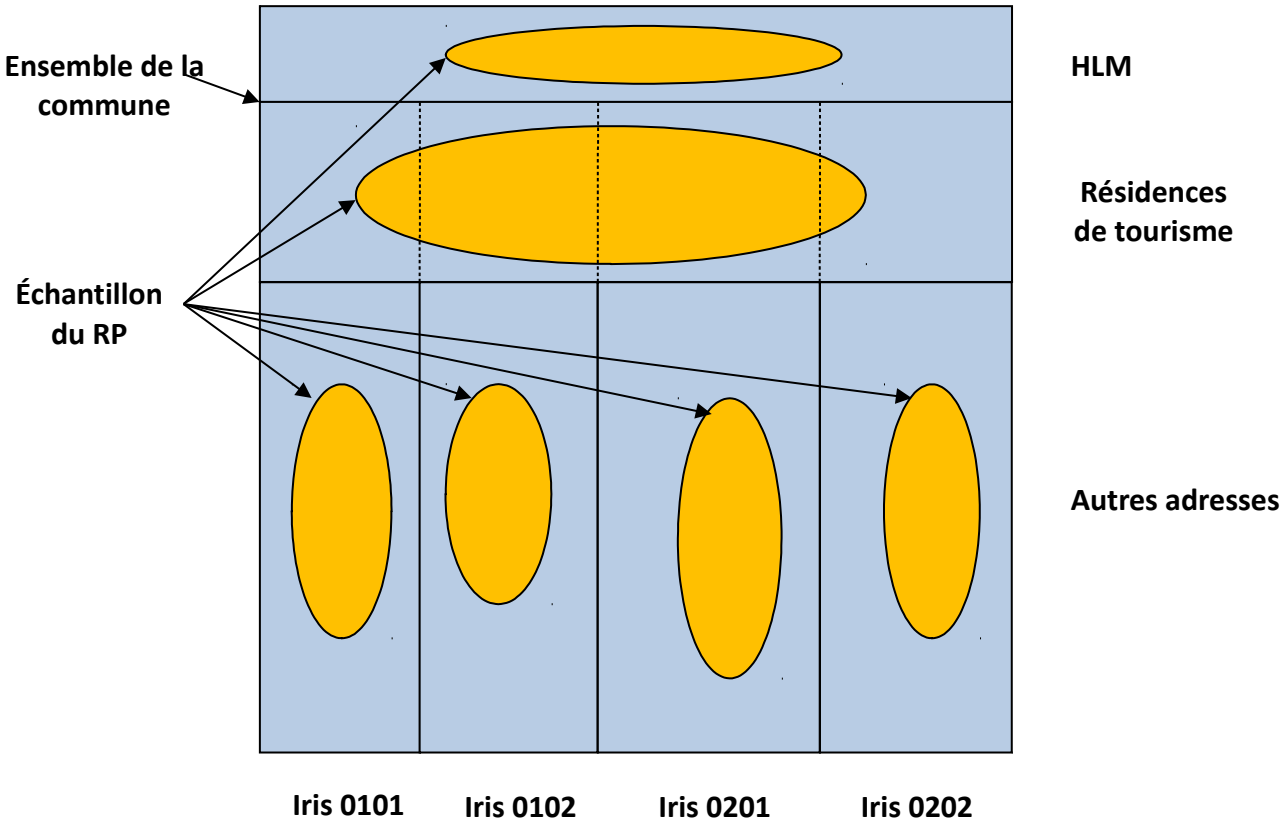
complet, soit l'enquête annuelle de recensement de 2020.

Ces structures sont peu nombreuses au niveau national. En juillet 2015, cette strate comportait 11 500 structures regroupant près de 100 000 logements, surtout présents dans les résidences hôtelières. Même si elles sont présentes dans plus de la moitié des grandes communes, ces structures sont très concentrées. En effet, 10 communes regroupent plus du quart des logements des hébergements touristiques. Ce sont dans ces communes où le gain de qualité devrait être le plus déterminant.

Pour les estimations, ces structures ont aussi été exclues des traitements antérieurs. Depuis le RP 2014, le calage par Iris sur le nombre de logement du Ril se fait sur les structures touristiques d'une part et par Iris sur les autres adresses d'autre part. Cette nouvelle post-stratification permettra d'éviter d'attendre un cycle complet pour améliorer la qualité des estimations. En régime courant, elle permettra également d'éviter qu'une telle structure soit estimée, à tort, comme une habitation classique tant qu'elle n'a pas été recensée pour la première fois.

Cette méthode s'inspire des estimations spécifiques mises en place sur la commune du sud de la France dont les estimations de population étaient particulièrement délicates. Cette commune dont la population dépasse à peine les 10 000 habitants est très touristique et compte des adresses de très grande taille (immeubles d'une part, adresses non normalisées d'autre part) soumis à la procédure d'échantillonnage[1] et de profils très hétérogènes. Environ 15 % des logements de la commune sont des structures d'hébergement touristique, quasiment inoccupés. En parallèle une adresse de HLM compte près de 4 personnes par logements.

Schéma 1 : Post-stratification utilisée



L'application d'une post-stratification a permis, grâce à un repérage complet des structures qui posaient potentiellement problème dans le RIL, de proposer une estimation de la population de la commune plus juste et de réduire sa volatilité. La différence entre les deux méthodes conduit à des écarts de près de 10 % de la population certaines années. Néanmoins, cette méthode mise en place a

1 En 2004, dans cette commune, seules les adresses de plus de 241 logements étaient considérées comme grandes et retenues exhaustivement dans l'échantillon.

posteriori ne corrige pas la volatilité de la composition de l'échantillon sur ces adresses atypiques. La nouvelle méthode permet d'élargir le périmètre géographique de cette amélioration et de proposer une solution appliquée en amont du dispositif.

La présentation visera à faire le bilan de l'apport de cette nouvelle méthode en termes d'instabilité des estimations pour l'ensemble des 930 grandes communes et plus particulièrement pour les plus touristiques d'entre-elles.

3. Un élargissement du principe

Cette première étape permet de corriger un problème bien identifiable et concentré sur un nombre limité d'adresses. Néanmoins, l'ajout d'une variable auxiliaire qu'est le repérage de l'intégralité des structures touristiques dans la base de sondage est très coûteux. Par ailleurs, l'exploitation accrue des sources administratives associée à leur localisation fine permet d'envisager l'intégration d'informations auxiliaires plus diverses et plus directement corrélées à la population et aux caractéristiques socio-démographiques que l'on souhaite estimer par l'intermédiaire du recensement. Le répertoire statistique des logements semble prometteur de ce point de vue. Il fournit un éclairage complet sur les sources fiscales que l'on peut rapprocher du recensement pour compléter d'un niveau de population fiscale, l'information auxiliaire actuelle qui ne fait intervenir qu'un nombre de logements.

La principale difficulté du recours à ces sources réside dans l'appariement entre les deux sources dont les niveaux d'identification et les concepts diffèrent.

L'intégration d'une population fiscale pour améliorer les estimations de population du recensement peut sembler contre-nature. Néanmoins, la source fiscale offre plusieurs garanties intéressantes. Tout d'abord, les concepts fiscaux (cadastre, taxe d'habitation) sur lesquels reposent l'appariement sont très inertes. Le risque de rupture de série liée à une modification fiscale est donc relativement réduit. Par ailleurs, contrairement au recensement, la source fiscale est disponible tous les ans de manière exhaustive. Elle apporte donc une vision plus complète de chaque territoire d'estimation.

Enfin, elle permettrait de limiter l'impact d'adresses dont le profil est atypique, de manière exhaustive et sans l'intervention très coûteuse a priori de nombreux gestionnaires. Dans un premier temps, l'amélioration des estimations de population pourrait être l'objectif poursuivi par la mise en place d'une méthode incluant des informations auxiliaires plus riches. Dans cette perspective, deux innovations pourraient être mises en place, l'une sur l'échantillonnage, l'autre sur la pondération.

D'une part, cet appariement permettra le repérage exhaustif de toutes les adresses atypiques au sens du calcul de la population. Il s'agit d'adresses dont le nombre moyen de personnes par logement s'écarte de la moyenne du quartier auquel elles appartiennent. De ce fait, la définition de la strate exhaustive pourrait être revue. Actuellement, elle est composée des adresses dont le nombre de logements dépasse un certain seuil, propre à chaque commune. La construction de la strate pourrait mêler ce concept simple à une mesure de la différence d'estimation de la population communale liée à chaque adresse selon son appartenance ou non à l'échantillon, si on soumet l'adresse à la procédure d'échantillonnage. L'impact de la procédure d'échantillonnage liée à chaque adresse k d'un Iris i est alors calculée de la manière suivante :

$$I_k = NbLog_i^{Ril} * \left(\frac{NbInd_{k,-i}^{TH}}{NbLog_{k,-i}^{TH}} - \frac{NbInd_{k,-i}^{TH} * \left(NbLog_k^{Ril} - d_k * NbLog_i^{Ril} \right) + d_k * NbInd_i^{TH}}{NbLog_{k,-i}^{TH} * \left(NbLog_k^{Ril} - d_k * NbLog_i^{Ril} \right) + d_k * NbLog_i^{TH}} \right)$$

Les adresses dont l'impact est le plus fort seraient alors intégrées à la strate exhaustive.

D'autre part, l'utilisation des variables fiscales dans la procédure de calage des poids pourra être utilisé en complément du nombre de logements issus du Ril. Pour améliorer l'estimation de la population de la commune, le nombre de personnes par logement pourrait être contrôlé au même titre que le nombre de logements qui occupe une position centrale actuellement. Intégrer la

population fiscale de chaque adresse dans la procédure de calage améliorerait la précision de l'estimation de la population communale. Une comparaison de cette nouvelle méthode avec la méthode actuelle permettra d'évaluer le gain de qualité qu'elle apporte sur la période allant de 2006 à 2013. Un indicateur de volatilité du nombre de personnes par logement permettra de comparer les deux méthodes.

4. Conclusion

Utiliser des informations auxiliaires de manière ciblée ou de manière plus large permet d'améliorer la qualité des estimations des populations légales.

Un objectif plus large pourrait être retenu si l'on souhaite améliorer la qualité des résultats statistiques issus du RP. Pour ce faire, il faudrait adapter les méthodes présentées ci-dessus aux thématiques que l'on souhaite améliorer. Par exemple l'amélioration de la structure par âge des résultats du RP pourrait être obtenue en intégrant ce critère au niveau de la phase d'échantillonnage et celle de repondération. Dans la définition de la strate exhaustive, ce critère pourrait ainsi être pris en compte en proposant un indicateur qui cible les adresses dont l'occupation par âge est atypique. En parallèle, la structure par âge observée au sein des sources fiscales pourrait être ajoutée aux variables de calage. En contrepartie, il faut rester vigilent à ne pas multiplier les objectifs au risque que cela ne se fasse au détriment de l'objectif principal qui reste l'amélioration de la qualité des estimations de population.

Bibliographie

[1] Ardilly, P. (2006), *Les Techniques de Sondages*, Technip.

[2] Bertrand, P., Chauvet, G., Christian, B., Grosbras, J.M. (2002), *Les plans de sondages du nouveau recensement*, Journées de Méthodologie Statistiques de l'Insee.

[3] Deroyon, T. (2015), Traitement des observations atypiques d'une enquête par winsorisation : Application aux Enquêtes Sectorielles Annuelles, Journées de Méthodologie Statistiques de l'Insee.

[4] Deville, J.C. and Särndal, C.E. (1992), *Calibration estimators in survey sampling*, Journal of the American Statistics Association 87, 376-82.

[5] Godinot, A. (2005), *Pour comprendre le recensement de la population*, Insee Méthodes

[6] Lermechin, H (2016), *Quel est l'apport de la taxe d'habitationà l'extrapolation du nombre de pésidences principales au Recensement de la Population*, Économie et Statistique n°483-484-485.

[7] Sautory, O. (1993), Document de Travail N°F9310 de L'Insee (1993), *la Macro CALMAR, Redressement d'un échantillon par calage sur marges*