Le passage à une collecte par sondage : quel impact sur la précision du recensement ?

Gwennaëlle Brilhault et Nathalie Caron *

Le recensement rénové de la population mis en place en 2004 est une opération tournante qui a permis d'annualiser la collecte, les dépenses et la publication des résultats, tout en assurant des gains d'efficacité des différents acteurs. Il se caractérise également par le fait que seule une fraction de la population des ménages des communes de 10 000 habitants ou plus est recensée : le sondage s'ajoute donc aux facteurs influençant la qualité d'un recensement traditionnel exhaustif. Cette décision a suscité de nombreuses interrogations autour de la variabilité supplémentaire ainsi introduite pour les estimations basées sur ces nouvelles données du recensement. En particulier, les élus locaux s'inquiétaient de l'impact de cette nouveauté sur l'effectif officiel de la population, lequel détermine le montant des ressources financières allouées aux communes par l'État. L'objectif de cet article est de proposer un historique des travaux chiffrant l'impact du sondage sur la précision des résultats du recensement, et d'en donner un état des lieux actuel. Ces travaux se sont déroulés en deux phases. D'abord, des calculs de précision ont été menés *via* des simulations basées sur les données du recensement exhaustif de 1999. Ensuite, ces premiers résultats ont été affinés en procédant à des calculs à partir des premières données d'un cycle complet du recensement rénové. Ces travaux méthodologiques démontrent que l'ordre de grandeur de cette imprécision est négligeable et par conséquent que le passage à la collecte par sondage n'a pas dégradé la qualité des résultats du recensement.

Codes JEL : C13, C83 et J11.

Mots clés : recensement de la population, variance d'échantillonnage, enquête par sondage, précision résultats recensement.

* *Insee, Direction des statistiques démographiques et sociales, Division méthodes et traitements des recensements.*

Les auteurs remercient Anne Rhodes, Sylvie Rousseau et les deux référés anonymes pour leurs commentaires constructifs sur la première version de cet article.

Le recensement a d'abord pour objet de fournir, pour chacune des communes, l'effectif officiel de la population appelé « populations légales ». Cet effectif est primordial puisqu'il détermine notamment le montant de la dotation globale de fonctionnement[1] ainsi que de nombreuses dispositions du code électoral. Le recensement fournit également un certain nombre de caractéristiques sur l'ensemble de la population française et sur ses logements, à un niveau national mais aussi à un niveau géographique fin (communal ou infra communal pour les plus grandes communes). À ce titre, il constitue une source d'informations incontournable pour de nombreux acteurs (pouvoirs publics, acteurs locaux, particuliers). Il est ainsi utilisé pour établir la répartition des services de santé, certaines politiques de prévention et gestion des risques, etc. : on dénombre près de 350 articles de lois ou de codes qui se réfèrent au recensement. Face à de tels enjeux, il est donc essentiel de savoir quelle confiance peut être accordée aux résultats d'un recensement, surtout avec la mise en place en 2004 d'une procédure innovante de recensement basé sur des échantillons tournants. En effet, si pour les recensements des années 2010, un nombre croissant de pays s'éloignent de la pratique classique du recensement en préférant des procédures s'appuyant sur des registres de population, exclusivement ou en combinaison avec d'autres sources (Valente, 2010), la France se distingue parmi eux par son modèle de recensement qui lui est propre.

La qualité des résultats d'un recensement tout comme celle des résultats d'une enquête dépend de multiples facteurs, et en tout premier lieu de la qualité de la collecte (Henry, 1949). Elle dépend aussi de l'exhaustivité des fichiers et répertoires utilisés pour préparer la collecte, ainsi que de la qualité des différents traitements mis en œuvre : saisie, contrôles, détection puis redressement des anomalies et codification des variables. Dans les standards internationaux, on considère qu'un recensement est de très bonne qualité dès lors qu'il atteint une précision de l'ordre de 1 % pour le chiffre de l'ensemble de la population d'un pays. La multiplicité des facteurs influençant la qualité d'un recensement traditionnel rend difficile voire impossible d'en mesurer précisément l'impact global. Des opérations ponctuelles appelées enquêtes post-censitaires ont été réalisées pour estimer certains facteurs, en particulier les double-comptes et le sous-dénombrement par omission. Ce fut le cas pour les recensements de 1962 et de 1990 (voir encadré 1). En effet, il est illusoire de penser qu'un recensement traditionnel, par nature exhaustif, fournit une réponse parfaite au dénombrement de la population : il existe toujours des individus non recensés (environ 1 % en 1990), en particulier des jeunes adultes ou des personnes vivant seules dans les grandes villes. Notons cependant que les enquêtes post-censitaires sont délicates à réaliser (Coeffic, 1993) et ne permettent pas de mesurer toutes les erreurs de dénombrement, en dépit de leur intérêt certain[2].

Au regard de ces facteurs traditionnels de qualité, la nouvelle méthode française de recensement est de nature à assurer un bon niveau de qualité, précisément parce que l'opération est devenue annuelle et par sondage (voir encadré 2). Ces deux originalités ont permis de répartir la charge de collecte dans le temps et d'augmenter l'efficacité des différents acteurs. En outre, la méthode suppose la tenue d'un répertoire exhaustif d'immeubles dans les communes de 10 000 habitants ou plus, dont la mise à jour annuelle bénéficie de l'expertise de chacune des communes. Cette base d'adresses, en principe complète et à jour[3], permet de garantir au maximum la qualité de la collecte car elle facilite la vérification de l'exhaustivité de l'information recueillie.

Cependant, le sondage qui singularise la méthode française introduit un tout nouveau facteur influençant la qualité des résultats d'un recensement. En effet, depuis 2004, une enquête par sondage s'est substituée à une collecte exhaustive pour estimer près de la moitié de la population française, la population des ménages des communes de 10 000 habitants ou plus. Cette décision a suscité de nombreuses interrogations autour de la variabilité supplémentaire liée au sondage pour les estimations basées sur ces nouvelles données du recensement. En particulier, les élus locaux s'inquiétaient de l'impact de cette nouveauté sur l'effectif officiel de la population et ont souhaité être clairement informés de la nature de l'impact d'un tel choix sur la précision des résultats.

---

1. La dotation globale de fonctionnement (DGF), instituée par la loi du 3 janvier 1979, est un prélèvement opéré sur le budget de l'État et distribué aux collectivités locales. Son montant est établi selon un mode de prélèvement et de répartition fixé chaque année par la loi de finances. Elle est versée aux régions depuis 2004. Cette dotation est constituée d'une dotation forfaitaire et d'une dotation de péréquation.

2. Sur le recensement rénové français, aucune opération d'enquête post-censitaire n'a été mise en place jusqu'ici.

3. L'exhaustivité de la base d'adresses dépend de l'implication de la commune dans son expertise. Les opérations sur la qualité de la base d'adresses réalisées au cours des dernières années indiquent que l'exhaustivité de cette base est globalement atteinte à 1 % près (voir encadré 2 sur les principales opérations de mesure de la qualité dans le recensement rénové).

Ainsi, depuis le début des années 2000 (c'est-à-dire durant la phase de construction de la nouvelle méthodologie du nouveau recensement), plusieurs études méthodologiques ont été réalisées dans le but de chiffrer précisément l'incertitude apportée par le sondage. Des premiers calculs d'intervalles de confiance ont été réalisés à partir du recensement de 1999 (dernier recensement général de population). Ces calculs ont ensuite été affinés à partir de 2009 avec la disponibilité du premier cycle complet de données du recensement rénové, formé des collectes de 2004 à 2008.

Après un rappel des grands principes du plan de sondage du recensement rénové, l'article présente un historique des travaux qui ont eu pour but de chiffrer l'impact de l'introduction du sondage sur les résultats du recensement (partie « Les estimations de précision du recensement rénové avant 2009 »), et donne un état des lieux de l'évaluation de cet impact dont on dispose aujourd'hui (partie « Le contexte plus riche à partir de 2009 permet d'afiner les calculs de précision »). Tous les travaux méthodologiques menés jusqu'à présent illustrent le fait que l'ordre de grandeur de cette imprécision est négligeable et que le passage à la collecte par sondage n'a pas dégradé la qualité des résultats du recensement.

Les grands principes du plan de sondage du recensement rénové

La loi n° 2002-276 du 27 février 2002, relative à la démocratie de proximité, définit en son titre V les principes d'organisation du recensement rénové de la population (Godinot, 2005). En particulier, le paragraphe VI de l'article 156 apporte les précisions suivantes : « *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* ».

Dans la mesure où le sondage ne s'applique qu'aux ménages, seul le principe de recensement de la population des ménages est détaillé ci-dessous. Cette catégorie constitue de loin la principale composante de la population

Encadré 1

LES ENQUÊTES POST-CENSITAIRES DE 1962 ET DE 1990

Du fait de l'importance que revêt à des titres divers un recencement pour ses utilisateurs, particuliers ou pouvoirs publics, des opérations générales de mesure de l'exhaustivité des recensements ont été réalisées dans le passé, en particulier sur le recensement de 1962 et sur celui de 1990 (Coeffic, 1993). Ces deux enquêtes post-censitaires réalisées à 28 ans d'intervalle reposent sur la même méthodologie : l'enquête se déroulait à une date la plus proche possible de celle de fin du recensement afin de pouvoir négliger les mouvements (déménagements, décès et naissances) dans les comparaisons réalisées entre le recensement et l'enquête. L'échantillon de l'enquête était sélectionné de façon à être représentatif de la France métropolitaine et constitué d'aires. Les enquêteurs devaient dénombrer de façon exhaustive tous les logements situés sur les aires sélectionnées. Tous les individus de ces logements devaient ensuite remplir un questionnaire approfondi dans lequel ils précisaient en particulier tous les lieux où ils pouvaient avoir été recensés (résidence secondaire, collectivité, précédent logement en cas de déménagement, etc.). En cas d'incohérences difficilement explicables entre le recensement et l'enquête, l'enquêteur retournait une nouvelle fois sur le terrain afin de compléter et/ou corriger l'enquête si nécessaire. En rapprochant les résultats du recensement et ceux de l'enquête post-censitaire, on pouvait alors déterminer, pour chaque personne, si elle avait été enquêtée une fois (à l'adresse du logement indiqué ou à une autre), deux fois ou aucune fois. Cette méthodologie permettait ainsi d'estimer les omissions et les doubles comptes.

L'échantillon de l'enquête post-censitaire de 1990 comportait 650 aires d'une quarantaine de logements, soit environ 30 000 logements. À partir de cette enquête, on estime que le taux d'omission du recensement de 1990 se situe entre 1,5 % et 2 % et que le taux de double-compte est de l'ordre de 1 %. Ces erreurs de dénombrement touchent particulièrement certaines catégories de population, parmi les plus mobiles : les jeunes adultes, les étrangers, les étudiants et les chômeurs. Ces résultats sont similaires à ceux obtenus pour l'enquête post-censitaire de 1962, malgré des difficultés de collecte plus importantes en 1990 : étudiants plus nombreux, nombre de résidences secondaires plus important, accès aux immeubles collectifs plus difficile avec la mise en place de digicodes.

française (près de 98 %). La population vivant hors ménage (c'est-à-dire en communauté (prisons, maisons de retraites, internats, etc.) ou dans des habitations mobiles, terrestres ou fluviales) est recensée exhaustivement en cinq ans. Celle-ci représente environ un million de personnes.

Un recensement exhaustif tous les cinq ans pour les communes de moins de 10 000 habitants de métropole

Les communes de moins de 10 000 habitants (appelées aussi « petites » communes par abus de langage) ont été réparties, une fois pour toutes, pour chaque région en cinq sous-ensembles appelés « groupes de rotation » par un tirage équilibré[4] (voir Deville et Tillé, 2005 et Tillé, 2011, pour la description d'un tirage équilibré ; Durr et Dumais, 2002 et Godinot, 2003, pour la méthode du recensement) basé sur les variables suivantes établies à partir du recensement de 1999 : nombre de logements dans chacun des départements de la région considérée, nombre de logements individuels, nombre de logements collectifs, population par tranche d'âge (de 0 à 19 ans ; de 20 à 39 ans ; de 40 à 59 ans ; de 60 à 74 ans et 75 ans et plus), population par sexe. Chaque année, la totalité des habitants d'un groupe de rotation donné est recensée, qu'ils vivent dans des logements ordinaires, en communautés ou dans des habitations mobiles, terrestres ou fluviales. En cinq ans, toutes les « petites » communes sont par conséquent recensées (Bertrand, Chauvet, Christian, Grosbras, 2002).

4. Cette technique d'échantillonnage permet d'obtenir des échantillons aléatoires reproduisant le plus fidèlement possible les structures de référence dont on dispose sur l'ensemble de la population de la région. Elle permet ainsi d'améliorer l'efficacité du sondage en diminuant la variance d'échantillonnage.

Encadré 2

LES PRINCIPALES OPÉRATIONS DE MESURE DE LA QUALITÉ DANS LE RECENSEMENT RÉNOVÉ

Avec le système rénové du recensement, qui remplace une collecte exhaustive opérée simultanément auprès de l'ensemble de la population par une opération tournante et par sondage, plusieurs facteurs jouent favorablement sur la qualité. Pour le mesurer et toujours améliorer les processus, différentes opérations sont menées :

- Le caractère plus régulier du recensement rénové par opposition à une opération massive et ponctuelle permet d'obtenir une qualité plus assurée. Chaque année, il est nécessaire de concentrer les efforts sur la partie du territoire concernée par le recensement : une commune de moins de 10 000 habitants sur cinq et une fraction d'adresses dans les communes de 10 000 habitants ou plus. Le recensement rénové en continu produit donc un gain difficilement mesurable, car il permet une plus grande efficacité des acteurs de la collecte. Ceux-ci cumulent leurs expériences de collecte d'une année sur l'autre, alors que pour un recensement traditionnel, les collectes sont trop espacées l'une de l'autre pour permettre une telle capitalisation.

- Des opérations nationales d'enquêtes de mesure de la qualité du RIL ont été organisées à plusieurs reprises depuis 2004 (en moyenne tous les deux ans) afin de mesurer l'adéquation du RIL (Répertoire d'immeubles localisés) avec le terrain : elles ont consisté à procéder au « ratissage » exhaustif d'un certain nombre d'Iris (voir note 5) afin de vérifier la qualité de couverture de la base de sondage d'adresses. Ces opérations indiquent que l'exhaustivité du RIL est globalement atteinte à 1 % près.

- Une opération sur la qualité de la saisie des données est menée chaque année : un échantillon de questionnaires (entre 5 000 et 6 000) fait l'objet d'une double saisie, par le prestataire qui est chargé de la saisie de tous les questionnaires de l'enquête annuelle de recensement et par un autre prestataire, afin d'évaluer la qualité de cette saisie. Les divergences entre les deux saisies sont analysées à l'Insee, de façon à établir des taux d'erreur imputables au prestataire principal sur le codage de chacune des variables. La double saisie est réalisée au fil de la campagne de saisie, ce qui permet, éventuellement, d'en corriger les protocoles pour la suite de la saisie.

- On mène aussi régulièrement des opérations (cinq opérations depuis 2004) concernant la qualité de la codification des champs « activité de l'employeur » et « profession » du questionnaire, en comparant des codages concurrents pour un sous-échantillon de 100 000 questionnaires individuels. Une expertise de tous les codages divergents et de 10 % des codages convergents est ensuite réalisée par un « arbitre ». Celui-ci connaît les deux premiers codages attribués et peut choisir entre ces deux derniers, ou en proposer un troisième différent. La proposition de l'arbitre est considérée comme correcte. Les taux d'erreurs de codage sont de niveaux acceptables (10 % d'erreurs sur la profession codée en 42 postes et 8,7 % d'erreurs sur l'activité codée au niveau division) et en baisse régulière. Les résultats permettent de faire évoluer les outils de codage automatique et les consignes de reprise des échecs de codage dans les directions régionales de l'Insee (Martal, Jacquin, 2012).

Mise à jour de la base de sondage et tirage de l'échantillon annuel pour les communes de 10 000 habitants ou plus de métropole

Pour les communes de 10 000 habitants ou plus de métropole (appelées aussi « grandes » communes), la base de sondage pour la population des ménages est constituée à partir d'un répertoire exhaustif d'adresses, appelé Répertoire d'immeubles localisés (RIL). Ce répertoire est mis à jour, chaque année, par l'Insee avec l'appui des communes, à l'aide d'informations d'origine administrative (permis de construire et de démolir, données de La Poste, données fiscales extraites du fichier de la taxe d'habitation) ; il bénéficie d'une expertise par les communes chaque année en juin.

Les adresses de chaque commune de 10 000 habitants ou plus sont réparties en cinq groupes de rotation. Les cinq groupes d'adresses ont été initialisés pour la 1re enquête annuelle de 2004 par un tirage équilibré donnant l'image la plus fidèle possible de chaque Iris[5] de la commune en termes de population par âge, sexe, type de logement (individuel ou collectif) et nombre de logements. Par définition, ces cinq groupes de rotation représentent bien la population au regard de l'information auxiliaire choisie : ils en assurent des estimations exactes. Ils sont depuis mis à jour chaque année pour tenir compte notamment des créations et des disparitions d'adresses.

Afin de tenir compte de la variabilité du nombre de logements selon les adresses et ainsi éviter des effets de « grappe »[6], les adresses sont réparties en trois catégories : les adresses de grande taille, les adresses nouvelles et les autres adresses :

- Les adresses de grande taille (ou « grandes adresses ») sont celles dont le nombre de logements dépasse un certain seuil, propre à chaque commune : ce sont les adresses d'au moins 60 logements chacune et cumulant au maximum 10 % des logements de la commune. Elles sont réparties dans les cinq groupes de rotation et sont enquêtées exhaustivement en cinq ans ; la répartition se fait par tirage équilibré pour les communes comptant 50 grandes adresses ou plus, de façon déterministe sinon ;

- Les adresses neuves (ou « adresses nouvelles ») qui apparaissent chaque année doivent toutes être recensées en cinq ans ; la répartition se fait par tirage équilibré pour les communes comptant 50 nouvelles adresses ou plus, de façon déterministe sinon ;

- Les autres adresses (ou « petites adresses connues ») sont réparties en cinq groupes équivalents en nombre de logements individuels, en nombre de logements collectifs, et en population par sexe et âge.

La catégorie d'une adresse n'est toutefois pas une caractéristique immuable. Plusieurs types d'évènements peuvent entraîner un changement de catégorie. Ainsi, par exemple, dès qu'une adresse récemment construite a été enquêtée, elle cesse d'être « nouvelle » : en fonction de sa taille, soit elle se voit qualifier de « grande » adresse, soit elle devient une « petite adresse connue » pour les années ultérieures. D'autre part, le seuil des grandes adresses propre à chaque commune n'est pas figé : l'apparition d'une très grande adresse à la suite d'un programme de construction peut, sous l'effet de la contrainte des 25 % maximum[7] de logements en grandes adresses, faire augmenter le seuil et ainsi basculer une « grande » adresse en « petite » adresse. La disparition de grandes adresses (cas d'une restructuration d'adresses par exemple) peut avoir l'effet inverse. Le seuil n'est cependant pas révisé à la baisse chaque année, afin de ne pas perturber l'estimation de la population par des modifications répétées et substantielles des poids des adresses ; ainsi, une révision du seuil à la baisse a été organisée en 2007 suite aux premières collectes sur l'ensemble des grandes communes : le seuil de grandes adresses a été révisé pour 167 des presque 900 communes de 10 000 habitants ou plus. Il s'agit de communes pour lesquelles les RIL avaient pu être grandement améliorés depuis le premier tirage d'échantillon en 2003 (cette amélioration ayant souvent consisté à découper des grandes adresses en plusieurs petites adresses distinguables sur le terrain, afin de gagner en granularité de l'échantillon). Depuis cette révision, très peu de communes ont connu une modification de leur seuil des grandes adresses à la hausse (deux seulement)

---

5. Pour la diffusion du recensement de la population de 1999, l'Insee avait développé un découpage du territoire en mailles de taille homogène (d'environ 2 000 habitants) appelées Iris. Pour le nouveau recensement, l'Iris constitue la brique de base en matière de diffusion de données infra-communales. Toutes les communes de 10 000 habitants et plus et une forte proportion des communes de 5 000 à 10 000 habitants sont découpées en Iris. On dénombre environ 16 100 Iris dont 650 dans les DOM.

6. L'effet de « grappe » est lié au fait que l'on sélectionne des adresses et non des individus ; il traduit la perte de précision due à l'existence d'une similarité entre les individus habitant une même adresse.

7. Le seuil des 10 % de logements dans la strate des grandes adresses a été fixé pour l'initialisation des seuils de grandes adresses en 2003. Chaque année, lors de l'expertise des RIL des grandes communes, la part des logements en grandes adresses est re-calculée : si celle-ci est supérieure à 25 %, alors le seuil de grandes adresses est révisé à la hausse de manière automatique.

et une nouvelle révision des seuils à la baisse concernant toutes les communes a été réalisée lors du tirage de l'échantillon de l'enquête annuelle de recensement de 2015.

Une fois la base de sondage mise à jour, la méthode de sondage consiste à sélectionner, chaque année dans le groupe de rotation de l'année, un échantillon d'adresses « représentatif » de la commune. Les grandes adresses et les adresses neuves du groupe de l'année sont toutes retenues : elles seront enquêtées exhaustivement. Les petites adresses connues sont échantillonnées de sorte que l'échantillon total (y compris les adresses nouvelles et celles de grande taille) représente environ 40 % des logements du groupe de rotation. Les critères d'équilibrage sont le nombre de logements, le nombre de logements en collectif et le nombre de logements à l'Iris. Par la suite, tous les logements d'une adresse échantillonnée seront enquêtés.

Ainsi, en cinq ans, la totalité du territoire de chaque « grande » commune est prise en compte. Le taux global de sondage est tel qu'au terme d'une période de cinq ans, 40 % des logements de la commune auront été enquêtés.

De façon schématique, la population d'une « grande » commune est obtenue en procédant de la façon suivante : on cumule les données recueillies pendant cinq ans sur les 40 % de logements enquêtés puis on extrapole à l'ensemble des logements présents dans la commune à la date de référence souhaitée, en l'occurrence le milieu de la période de cinq ans (voir partie « Du sondage aux résultats »).

L'application des plans de sondage aux départements d'outre-mer (DOM)

Dans les départements d'outre-mer (La Réunion, Martinique, Guadeloupe, Guyane[8]), le faible nombre de communes de moins de 10 000 habitants (67 parmi 112 au recensement de 1999) n'a pas permis de les répartir en cinq groupes aussi strictement équilibrés qu'en métropole. Ces petites communes ont donc été réparties par DOM dans les cinq groupes par choix raisonné de façon à respecter au mieux un même nombre d'habitants par groupe.

Dans les grandes communes des DOM, la qualité de l'adressage n'a pas été jugée suffisante pour constituer un RIL sur des principes similaires à ceux de métropole. La méthode de tirage a été adaptée en conséquence. Ainsi, le territoire de chacune des grandes communes a été découpé en petites zones (appelées îlots[9]) à partir du recensement de 1999 ; les îlots ont ensuite été répartis dans cinq groupes de rotation selon un tirage équilibré reposant sur les mêmes variables que celles utilisées en métropole pour la répartition des adresses[10]. Chaque année, on réalise une enquête cartographique dans les îlots du groupe de rotation de l'année afin de mettre à jour les adresses et le nombre de logements pour chacune d'entre elles. On sélectionne ensuite un échantillon de 40 % d'adresses dans le groupe de rotation de l'année actualisé avec l'enquête cartographique. Précisons que pour les DOM, faute d'information suffisante, il n'y a aucune stratification en grandes adresses, adresses neuves et autres adresses, pas plus qu'il n'y a d'ajustement sur le parc de logements à la date médiane de la période de cinq ans considérée.

Du sondage aux résultats

Les résultats labellisés « Recensement de la population » sont issus du cumul de cinq enquêtes annuelles consécutives. Ils distinguent les populations légales et les résultats statistiques.

Le calcul des populations légales

Pour le calcul des populations légales, l'Insee doit respecter deux grandes contraintes : publier tous les ans avant le 31 décembre la population de toutes les communes relative à une même année. Ceci garantit l'égalité de traitement des communes entre elles et la qualité de la population des ensembles de communes (par exemple des établissements publics de coopération intercommunale − EPCI). On ne peut donc pas retenir comme population légale d'une année donnée pour une petite commune le chiffre de sa collecte de 2009 et pour une autre celui de sa collecte de 2013.

La méthode mise au point pour respecter cette seconde contrainte dans le cadre du calcul de la population des ménages est différente selon la taille des communes et leur appartenance à la métropole ou aux DOM. Plaçons-nous dans le

8. Mayotte (DOM depuis 2011) procède, selon l'article 157 du titre V de la loi n° 2002-276 du 27 février 2002, à des recensements généraux de la population tous les cinq ans.
9. Ils correspondent aux îlots utilisés pour le RP *1999 qui ont été redécoupés à la marge pour tenir compte de l'évolution démographique.*
10. Nombre de logements individuels, nombre de logements collectifs, population par tranche d'âge (0-19 ans/ 20-39 ans/ 40-59 ans/ 60-74 ans/ 75 ans et plus), population par sexe.

cadre du recensement de la population en référence au 1er janvier *n*, appelé *RP n*, qui s'appuie sur les cinq enquêtes annuelles de recensement réalisées entre *n*-2 et *n*+2 :

A - Cas des communes de métropole de moins de 10 000 habitants

Dans ce cas, il convient de se ramener à la date médiane du cycle. Sur cinq années, elles connaissent alternativement différentes méthodes de calcul de leur population légale, en fonction de la position de leur année de collecte par rapport à l'année de référence des populations légales :

- Pour les communes recensées en *n*, on retient le résultat de l'enquête de recensement menée sur le territoire de la commune pour la population des ménages.

- Pour les communes recensées en *n*+1 et en *n*+2, la population est calculée en ramenant les résultats de la collecte en *n*. Pour ce faire, on utilise la tendance observée sur la commune entre la dernière population légale au 01/01/*n*-1 et les résultats de l'enquête de recensement. Cela se traduit par le calcul suivant pour les communes recensées en *n*+1 :

pop men *RP n* = pop men *RP n*-1 +1/2
(pop men *EAR n*+1 - pop men *RP n*-1)

avec :

pop men *RP n* = population des ménages au *RP n* qui est en référence au 1er janvier *n* ;

pop men *EAR n*+1 = population des ménages de l'enquête annuelle de recensement de l'année *n*+1.

De manière similaire, on obtient pour les communes recensées en *n*+2,

pop men *RP n* = pop men *RP n*-1 +1/3
(pop men *EAR n*+2 - pop men *RP n*-1)

On ajoute ensuite la population recensée dans les hôtels.

- Pour les communes recensées en *n*-2 et en *n*-1, on calcule la population des ménages en *n* à partir de la population recensée et de l'évolution du parc de logements connue grâce au fichier de la taxe d'habitation. Comme le nombre d'habitants et le nombre de logements n'évoluent pas forcément de la même façon, on tient compte également de l'évolution du nombre moyen de personnes par ménage.

Cela se traduit par le calcul suivant pour les communes recensées en *n*-1 :

pop men *RP n* = pop men *RP n*-1 *
évol nb log TH *n*/*n*-1 *
différentiel de taille des ménages

avec :

pop men *RP n* = population des ménages au *RP n* qui est en référence au 1er janvier *n* ;

évol nb log TH *n*/*n*-1 = évolution du nombre de logements selon la taxe d'habitation entre les années *n*-1 et *n* ;

différentiel de taille des ménages = prise en compte des évolutions annuelles différentes entre la population et le nombre de logements entre deux enquêtes annuelles de recensement (*EAR*). Ce terme, calculé à partir des données recensées en *n*-1 et de celles recueillies cinq ans plus tôt[11], est défini par :

[(pop men *EAR n*-1/pop men *EAR n*-6)/
(nb log *EAR n*-1/nb log *EAR n*-6)]$^{(1/5)}$

où nb log *EAR n-k* est le nombre de logements de l'enquête annuelle de recensement de l'année *n-k*.

- Pour les communes recensées en *n*-2, le calcul est rigoureusement identique. La seule différence provient de la nature du terme pop men *RP n* : pour les communes recensées en *n*-1, il correspond exactement au résultat de l'enquête de recensement et pour les communes recensées en *n*-2, il est obtenu par une formule similaire utilisant la taxe d'habitation entre les années *n*-2 et *n*-1 et le résultat du recensement de l'année *n*-2. Ainsi, pour les communes recensées en *n*-2, on peut aussi réécrire la formule sous la forme :

pop men *RP n* = pop men *RP n*-2 *
évol nb log TH *n*/*n*-1 *
évol nb log TH *n*-1/*n*-2 *
(différentiel de taille des ménages)^2

11. Pour le calcul des populations des ménages en référence du 1er janvier 2006 au 1er janvier 2009, ce terme est calculé à partir des données recensées en n-*1 et de celles recueillies au* RP *1999.*

où le différentiel de taille des ménages est calculé à partir des données recensées en *n*-2 et de celles recueillies cinq ans plus tôt[12] :

$$[(\text{pop men } EAR\ n\text{-}2/\text{pop men } EAR\ n\text{-}7)/(\text{nb log } EAR\ n\text{-}2/\text{nb log } EAR\ n\text{-}7)]^{(1/5)}$$

nb log *EAR n-k* étant le nombre de logements de l'enquête annuelle de recensement de l'année *n-k*.

On ajoute ensuite la population recensée dans les hôtels.

B - Cas des communes de métropole de 10 000 habitants ou plus

Dans ce cas, la population des ménages est estimée en multipliant le nombre de logements au 1er janvier de l'année de référence par le nombre moyen de personnes par logement estimé lors des cinq dernières collectes par sondage (ce calcul étant décliné au niveau de chacun des Iris de la commune concernée):

$$\text{pop men } RP\ n = \text{nb log au } 01/01/n * \text{nb pers par log GC}$$

avec :

pop men *RP n* = population des ménages (lors logements de fonction des communautés) au *RP n* qui est en référence au 1er janvier *n* ;

nb log au 01/01/*n* = nombre de logements au 1er janvier *n* calculé via la moyenne des nombres de logements des deux RIL de juillet encadrant ce 1er janvier *n* ;

nb pers par log GC = nombre moyen de personnes par logement estimé au moyen des cinq dernières collectes par sondage (collectes *n*-2 à *n*+2) de la grande commune considérée.

On ajoute ensuite la population recensée dans les hôtels.

C - Cas des départements d'outre-mer ne possédant pas de répertoire d'adresses pour les grandes communes

Dans ce cas, la méthode est différente de celle de métropole pour les grandes communes :

- Pour les communes des DOM de 10 000 habitants ou plus, la population des ménages est estimée sur chacun des groupes de rotation à partir des résultats de l'année d'enquête. L'estimateur obtenu est amélioré par ratio en utilisant comme variable auxiliaire le nombre total de logements du groupe de rotation à l'enquête cartographique. La population des ménages au *RP n* correspond alors à la somme des estimations obtenues sur les cinq groupes de rotation.

- Pour les communes des DOM de moins de 10 000 habitants, le principe général est similaire à celui de métropole. Ainsi, pour les populations légales publiées en fin de l'année *n*+2 et en référence au 1er janvier *n* :

  Pour les communes recensées en *n*, on retient le résultat de l'enquête de recensement menée sur le territoire de la commune pour la population des ménages.

  Pour les communes recensées en *n*+1 et en *n*+2, la population est calculée en ramenant les résultats de la collecte en *n*. Pour ce faire, on utilise la tendance observée sur la commune entre la dernière population légale au 01/01/*n*-1 et les résultats de l'enquête de recensement.

  Pour les communes recensées en *n*-2 et en *n*-1, le calcul est différent pour La Réunion et pour les autres DOM. Pour La Réunion, on calcule la population des ménages en *n* à partir de la population recensée et de l'évolution du parc de logements connue grâce au fichier de la taxe d'habitation. Comme le nombre d'habitants et le nombre de logements n'évoluent pas forcément de la même façon, on tient compte également de l'évolution du nombre moyen de personnes par ménage. Pour les autres DOM, on calcule la population des ménages en *n* en prolongeant l'évolution moyenne observée entre les chiffres de la précédente collecte (1999 pour les premières années du *RP* rénové) et les résultats de l'enquête de recensement.

On ajoute ensuite la population recensée dans les hôtels.

Au total, la population municipale s'obtient en sommant la population des ménages[13] (dont le principe de calcul a été donné ci-dessus), la population des communautés (qui sont

12. Pour le calcul des populations des ménages en référence du 1er janvier 2006 au 1er janvier 2010, ce terme est calculé à partir des données recensées en n-2 et celles recueillies au RP 1999.
13. À noter que la population des ménages comprend également la population recensée dans les hôtels.

recensées exhaustivement en cinq ans) et la population composée des personnes sans abri ou vivant dans des habitations mobiles (terrestres ou fluviales). Conformément au décret n° 2003-485 de 2003, on calcule également pour chaque commune la population dite « comptée à part ». Elle comprend des personnes recensées sur une autre commune mais qui ont conservé un lien avec la commune d'intérêt (par exemple des étudiants qui rentrent le week-end chez leurs parents) ainsi que des personnes qui lui sont rattachées administrativement[14].

La population totale d'une commune est égale à la somme de la population municipale et de la population comptée à part de la commune. Il s'agit du chiffre de référence pour de très nombreux textes législatifs ou réglementaires.

Chaque année, l'Insee publie ainsi, pour chaque commune, sa population municipale, sa population comptée à part et sa population totale. Ces trois chiffres sont authentifiés par décret au plus tard le 31 décembre de l'année et rentrent en vigueur au 1er janvier suivant. Ce sont les « populations légales ».

Les résultats statistiques et l'exploitation complémentaire

De la même manière que pour le calcul des populations légales, les résultats statistiques du recensement sont produits chaque année à partir des données recueillies lors des cinq dernières enquêtes annuelles. Ils décrivent l'ensemble des réponses individuelles ainsi que celles propres au logement et à la famille. Les pondérations utilisées sont celles obtenues par le calcul des populations légales, en répartissant sur l'ensemble des individus des cinq dernières collectes l'agrégat obtenu à l'issue du calcul de population légale décrit dans la partie précédente, commune par commune. Cette utilisation « empilée » des données collectées pendant cinq années ne pose pas de problème dans l'analyse de la majorité des comportements démographiques. Cependant, elle peut conduire à une vision biaisée pour étudier des disparités spatiales sur des indicateurs conjoncturels du type taux de chômage (Rapport du Cnis, 2005).

Comme pour les recensements traditionnels, les résultats détaillés distinguent deux exploitations de données :

- L'une, dite principale, mobilise l'ensemble des données recueillies et porte sur la plupart des questions, à l'exception de celles liées à la profession, l'activité ou la famille, de nature plus complexe ;

- L'autre, dite complémentaire, permet précisément de décrire la structure familiale des ménages ainsi que la catégorie socioprofessionnelle des individus, notamment des personnes qui exercent un emploi. Cette exploitation, plus coûteuse, est réalisée sur un sous-échantillon constitué de la façon suivante pour les ménages[15] :

  Pour les communes de 10 000 habitants ou plus : l'ensemble des résidences principales recensées et de leurs habitants, soit environ 40 %.

  Pour les communes de moins de 10 000 habitants : 25 % des résidences principales recensées et tous les habitants des logements sélectionnés font partie de l'échantillon de l'exploitation complémentaire[16]. Ce principe général admet deux exceptions en Corse et à Saint-Pierre-et-Miquelon, où l'exploitation complémentaire porte sur l'ensemble des données, comme l'exploitation principale.

Les résultats statistiques du recensement en référence au 1er janvier *n* sont mis en ligne sur le site insee.fr chaque année en juin *n*+3. Les données au niveau infra-communal sont diffusées en octobre *n*+3 et sont accompagnés d'indicateurs de précision, pour chaque question et au niveau géographique le plus fin. Ces informations sur la

---

14. Le concept de population comptée à part est défini par le décret n° 2003-485 publié au Journal officiel du 8 juin 2003, relatif au recensement de la population. La population comptée à part comprend certaines personnes dont la résidence habituelle (au sens du décret) est dans une autre commune mais qui ont conservé une résidence sur le territoire de la commune :
- Les mineurs dont la résidence familiale est dans une autre commune mais qui résident, du fait de leurs études, dans la commune ;
- Les personnes ayant une résidence familiale sur le territoire de la commune et résidant dans une communauté d'une autre commune, dès lors que la communauté relève de l'une des catégories suivantes : services de moyen ou de long séjour des établissements publics ou privés de santé, établissements sociaux de moyen ou de long séjour, maisons de retraite, foyers et résidences sociales ; communautés religieuses ; casernes ou établissements militaires.
- Les personnes majeures âgées de moins de 25 ans ayant leur résidence familiale sur le territoire de la commune et qui résident dans une autre commune pour leurs études.
- Les personnes sans domicile fixe rattachées à la commune au sens de la loi du 3 janvier 1969 et non recensées dans la commune.
15. Une sélection existe aussi pour la population hors ménages : elle se fait directement en sélectionnant des individus pour la population des sans-abris et pour celle qui réside dans les communautés. Pour les habitations mobiles terrestres, on sélectionne des logements et on retient tous les individus des logements tirés.
16. À partir de la collecte 2014, le taux du complémentaire dans ces communes est de 20 %.

précision se présentent sous la forme d'une table donnant, pour chaque variable, le coefficient de variation (en %) en fonction de la tranche d'effectif (voir plus loin). Cette table est utilisable pour les Iris des communes de 10 000 habitants ou plus de France métropolitaine et ne porte que sur la population des ménages.

Les estimations de précision du recensement rénové avant 2009

Les premiers travaux d'estimation de la précision liée au sondage sur le recensement de la population rénové s'inscrivaient dans le cadre des réflexions méthodologiques permettant de fixer définitivement les différents paramètres du plan de sondage. Ils ont donc été réalisés en l'absence de données d'un cycle du *RP* rénové complet (car avant l'achèvement des cinq premières années de collecte) et reposaient sur les données du recensement de la population de 1999.

Ces calculs ne concernent que la partie échantillonnée de la population (les ménages des communes de 10 000 habitants ou plus) et ne portent pas sur la partie complémentaire. Néanmoins ils sont assez lourds car ils doivent être faits sur l'équivalent d'un cycle de cinq ans de collecte (soit 45 millions d'observations) en tenant compte à la fois du plan de sondage équilibré et de la procédure de calage.

Méthodologie

Les premières estimations de variance ont été obtenues par simulations sur les données du dernier recensement exhaustif (le recensement de 1999, *RP* 99), notamment parce que la méthode de calcul des populations légales n'était pas fixée précisément et parce qu'on ne disposait pas encore d'un cycle complet (cinq ans) de données du recensement rénové.

À partir de la réalisation d'un certain nombre de tirages dans le recensement traditionnel de 1999 et en suivant le plan de sondage envisagé pour le recensement rénové, plusieurs calculs ont été réalisés afin d'estimer la précision du recensement rénové. On faisait varier plusieurs paramètres, notamment :

- La méthodologie retenue (strate des grandes adresses : pertinence de la strate et détermination du seuil permettant de caractériser une grande adresse, calage : niveau géographique retenu, etc.) ;

- Les variables d'intérêt (population seule ou jeu de plusieurs questions collectées au *RP*) ;

- Le nombre d'échantillons simulés pour le calcul de variance ;

- Le niveau géographique retenu (plusieurs communes, quelques Iris).

En général, dans chacun des cas, près de 500 tirages dans les données du recensement de 1999 ont été réalisés. Cependant, étant donné

Tableau 1
La précision des résultats pour les communes de 10 000 habitants ou plus

Tranches d'effectif En nb	Précision (CV) En %
50 000 ou plus	< 1,0
20 000-49 999	1,5
10 000-19 999	2,0
6 000-9 999	2,5
3 000-5 999	3,0
2 000-2 999	3,5
1 000-1 999	4,5
500-999	6,0
250-499	8,0
Moins de 250	> 8,0

Lecture : si le nombre d'enfants de moins de 5 ans d'une commune donnée de 10 000 habitants ou plus est de 2 700, la précision de ce résultat est de l'ordre 3,5 %.
Champ : communes métropolitaines de 10 000 habitants ou plus.
Source : Rapport du Cnis « Utilisation des données produites par le recensement rénové de la population et leur diffusion », *décembre 2005.*

le volume et la lourdeur de ces calculs, ceux-ci n'ont pour la plupart pas pu être réalisés sur la totalité des « grandes » communes. Par exemple, les calculs sur les variables d'intérêt n'ont été réalisés que sur 60 communes de 10 000 habitants ou plus et ont été extrapolés aux autres communes. Ces simulations ont permis d'optimiser au mieux le plan de sondage retenu pour les communes de 10 000 habitants ou plus[17].

Résultats obtenus

Le résultat le plus connu et utilisé de ces travaux est le tableau (voir tableau 1) présenté dans le Rapport du Cnis de décembre 2005, disponible sur le site du Cnis : pour toute commune de 10 000 habitants ou plus, ce tableau fournit une estimation du coefficient de variation lié au sondage en fonction de la taille de la population estimée.

Il a été établi pour le chiffre de l'ensemble de la population d'une commune. Il permet aussi de juger de la précision des résultats pour des populations ciblées, si celles-ci se répartissent de façon homogène sur l'ensemble de la commune. Si ce n'est pas le cas, la précision est moins bonne. En toute rigueur, ce tableau diffère selon les communes, la part des adresses enquêtées exhaustivement (adresses de grande taille et adresses neuves) dans l'ensemble des adresses variant d'une commune à une autre.

Ces indications ont permis de rassurer sur l'impact lié à l'introduction du sondage dans la précision du recensement. Ils ont également permis de définir le détail d'information à la fois utile et fiable au niveau infra-communal, moins riche que pour les recensements traditionnels du fait d'une partie enquêtée par sondage. Ceci a permis de préparer la diffusion sur le site insee.fr.

Le contexte plus riche à partir de 2009 permet d'affiner les calculs de précision

Méthodologie

À partir de 2009, le contexte change puisque la méthode de calcul des populations légales et des résultats statistiques est alors fixée et l'ensemble des résultats du premier *RP* rénové (*RP* 2006) sont disponibles. Dans la perspective de la mise à disposition de ces données, un groupe de travail interne à l'Insee s'est mis en place. Sa principale mission portait sur la validation statistique des données infra-communales afin d'en déterminer le niveau de finesse et les conditions d'utilisation. Son champ d'expertise se limitait aux communes de 10 000 habitants ou plus de métropole, c'est-à-dire aux communes pour lesquelles la collecte procède par sondage, hors DOM dans un premier temps.

Pour calculer la précision des résultats en métropole, une macro SAS a été développée par G. Chauvet. Elle est basée sur les formules de précisions en présence d'échantillons équilibrés issues des travaux de Deville et Tillé (2000) et tient compte de l'équilibrage des adresses en cinq groupes de rotation et du calage à l'Iris *via* la méthode des résidus (voir encadré 3). L'effet du calage a ainsi été pris en considération en remplaçant dans la formule de l'estimation de variance la variable d'intérêt Y par le résidu de la régression linéaire de la variable Y sur les différentes variables de calage (Deville et Särndal, 1992). La macro permet également de calculer la précision sur des variables de l'exploitation complémentaire, donc en tenant compte de l'échantillonnage réalisé dans l'exploitation complémentaire pour les communes de moins de 10 000 habitants.

Comme expliqué au début de cet article, le calcul de précision fourni par cet outil sous-estime la variance : en effet, il omet la perte de précision engendrée par la non prise en compte de la correction de la non-réponse totale, par l'existence de redressements importants pour certaines variables, etc. Cette perte de précision n'est d'ailleurs pas propre au nouveau recensement, elle existe également dans tout recensement traditionnel. De plus, du fait de la complexité du plan de sondage, le calcul réalisé se base nécessairement sur un certain nombre de simplifications, qui sont en partie listées ci-dessous :

- En premier lieu, le calcul néglige la dimension temporelle (étalement sur cinq ans des collectes utilisées pour calculer les populations légales d'un millésime *RP*) et utilise les données de cinq collectes annuelles comme si elles avaient toutes été collectées en même temps.

- Le calcul se base de plus sur une vision simplifiée des frontières entre grandes adresses, nouvelles adresses, petites adresses connues, alors que ces frontières peuvent être franchies par un certain nombre d'adresses au cours d'un

17. Ces travaux de simulation n'ont pas été diffusés en dehors de l'Insee.

cycle de cinq ans (révision du seuil des grandes adresses, reclassement avec un poids unitaire des « petites » adresses qui se sont révélées « grandes » à la collecte, etc.). Tous les effets de la dynamique des logements et de la population sur la variance ne sont donc pas retenus.

Encadré 3

LES CALCULS D'ESTIMATIONS DE VARIANCE

On se place dans une « grande commune » ; on note U la population des adresses présentes dans cette « grande commune » une année donnée, $\boldsymbol{Y}$ une variable d'intérêt, qui prend la valeur y_k pour l'adresse k de U et S l'échantillon obtenu par cumul des cinq années de collecte.

On s'intéresse à l'estimation du total Y de la variable $\boldsymbol{Y}$ pour la « grande commune » considérée.

L'estimateur d'Horvitz-Thompson est :

$$\widehat{Y} = \sum_{k \in S} d_k y_k$$

où d_k est le poids de sondage de l'adresse k, qui vaut 1 pour les « grandes adresses » et les « adresses nouvelles », et est de l'ordre de 2,5 (1/40 %, 40 % étant le taux de sondage) pour les « petites adresses connues ». L'estimateur de Y après calage sur le nombre de logements au 1er janvier de l'année n est noté :

$$\widehat{Y}^* = \sum_{k \in S} w_k y_k$$

où w_k est le poids de l'adresse k après calage.

La variance de cet estimateur est approximée (Deville et Särndal, 1992) par la formule $V\left(\widehat{Y}^*\right) \approx V\left(\widehat{E}\right)$ où $\widehat{E}$ est le total estimé de la variable résidu e_k obtenue par la régression (pondérée par les poids de sondage) de la variable y_k sur le nombre de logements de l'adresse k.

Or, les « grandes adresses » et les « adresses nouvelles » sont recensées exhaustivement en cinq ans : leur contribution à la variance totale est donc considérée comme nulle ; seules les « petites adresses connues » contribuent.

En cinq ans, des adresses sont enquêtées dans tous les groupes de rotation. La variance de l'estimateur obtenu par cumul des cinq années de collecte est donnée par celle de l'estimateur dans la catégorie des « petites adresses connues » qui peut s'écrire sous la forme :

$$V\left(\widehat{E}\right) = V\left(\sum_{h=1}^{5} \widehat{E}_{PAC,h}\right)$$

où $\widehat{E}_{PAC,h} = \sum_{k \in s_h} d_k e_k$.

L'échantillon à enquêter chaque année pour les « petites adresses connues » est issu d'un plan de sondage à deux phases : dans la première phase, les adresses sont réparties en cinq groupes de rotation par un échantillonnage à probabilités égales équilibré ; dans la seconde phase, un échantillon de « petites adresses connues » s_h est sélectionné chaque année dans le groupe de rotation concerné R_h, à probabilités égales et en équilibrant sur un vecteur Z_k de variables auxiliaires.

En utilisant la formule de décomposition de la variance d'un plan de sondage en deux phases, l'indépendance conditionnellement à la première phase des tirages de seconde phase ainsi que le fait que l'estimateur soit basé sur le cumul des cinq années de collecte, nous obtenons :

$$V\left(\widehat{E}\right) = E_1\left(\sum_{h=1}^{5} V_{2/1}\left(\widehat{E}_{PAC,h}\right)\right)$$

où E_1 est l'espérance liée à la première phase et $V_{2/1}$ est la variance liée à la seconde phase conditionnellement à la première phase.

En approximant la variance du plan de sondage équilibré de la seconde phase conditionnellement à la première par la formule proposée par Deville et Tillé (2005), on obtient la formule suivante comme estimation de variance :

$$\widehat{V}\left(\widehat{E}\right) = \sum_{h=1}^{5} \frac{1-\pi_2}{(\pi_2)^2} \sum_{k \in s_h} (u_k)^2$$

où π_2 est la probabilité de sélection d'une adresse (qui est constante au sein de chaque groupe de rotation de la commune et qui est proche de 40 %) et $u_k = e_k - (\widehat{B}_z^h)^t z_k$ avec $\widehat{B}_z^h = \left(\sum_{k \in s_h} z_k (z_k)^t\right)^{-1} \left(\sum_{k \in s_h} z_k e_k\right)$

L'utilisation de la formule de variance proposée par Deville et Tillé (2005) pour le calcul de précision d'un échantillon équilibré repose en particulier sur l'hypothèse que le plan de sondage garantit un équilibrage exact en respectant les probabilités de sélection. Cette hypothèse n'est pas forcément vérifiée dans le cas du plan de sondage du recensement mais de nombreuses simulations ont montré que l'estimation de variance proposée par Deville et Tillé (2005) est généralement correcte et stable.

L'estimation de variance pour l'estimateur du total de la variable $\boldsymbol{Y}$ pour l'ensemble des grandes communes s'obtient en sommant les estimations de variance obtenues pour chacune d'entre elles calculées avec la formule ci-dessus.

- L'équilibrage initial des adresses en cinq groupes de rotation n'a pas entièrement été réalisé aléatoirement (avec la macro Cube). En effet, lorsqu'il y avait moins de 50 grandes adresses ou moins de 50 nouvelles adresses, celles-ci ont été réparties de manière déterministe dans les cinq groupes de rotation. Il arrive également que l'équilibrage n'ait pas pu être atteint dans chacun des Iris. En effet, il peut arriver qu'il n'y ait pas toujours d'échantillons vérifiant exactement les équations d'équilibrage ou que l'équilibrage conduise à un choix déterministe. On se contente alors d'un équilibrage dit « approximatif ». Ce cas de figure n'est pas non plus pris en compte dans les formules de variance utilisées.

- On procède à quelques retirages lors de la confection des échantillons, dans les cas rares où l'on est trop loin des 40 % visés pour le groupe de rotation de l'année. Ceci peut être nécessaire lorsque la commune se distingue par un taux important de grandes adresses, ce qui réduit le nombre d'adresses échantillonnées.

- Ce calcul de précision considère que la source de calage (la base de sondage des adresses ou BSA) est exacte. Or, cette base de calage au 1er janvier n'est qu'une construction, calculée comme la moyenne de deux BSA, datant respectivement des mois de juillet précédant et suivant : elle n'a pas de réalité propre. En outre, cette BSA « médiane » est mise à jour par les résultats de collecte, lesquels portent uniquement sur l'échantillon de l'année, c'est-à-dire sur un seul des cinq groupes d'adresses et seulement sur les adresses enquêtées.

- L'ajustement sur le nombre de logements de la base de calage se fait tous types d'adresses confondus (grandes adresses, nouvelles adresses, petites adresses connues). Ce choix est lié au fait que les changements de « strates » entre grandes et petites adresses étaient mal maîtrisés (difficiles à suivre exhaustivement et potentiellement fréquents), si bien qu'on a choisi de définir les strates d'estimation comme étant les Iris tous types d'adresses confondus ; cela induit que les grandes adresses peuvent avoir un poids différent de 1 après calage.

- Dans les calculs de variance concernant des variables du complémentaire, pour lequel les données des petites communes sont elles aussi échantillonnées (à hauteur de 25 %), ce qui induit une variance d'échantillonnage liée aux petites communes, on ne tient pas compte du fait que ces données ont subi divers traitements (interpolation ou extrapolation) pour être ramenées à l'année de référence des populations légales.

Par rapport aux simulations réalisées jusqu'en 2009 à partir des données du recensement 1999, la macro SAS permet de faire des calculs beaucoup plus systématiques : à différents niveaux communal, supra-communal (région, département) et infra-communal (Iris) et sur toutes les variables. Les résultats obtenus peuvent ensuite être comparés à ceux donnés précédemment par les simulations antérieures à 2009.

La macro a été utilisée fin 2009 pour toutes les variables du *RP* 2006 au niveau Iris, puis début 2010 pour la variable population du *RP* 2006 au niveau région, département et commune de métropole. En 2010, ce programme a été adapté à la diffusion infra-communale du recensement sur des zones à façon, offre ouverte aux organismes ayant une mission de service public (une collectivité territoriale, un service de l'État, etc.) : la précision est alors calculée en incluant un calage supplémentaire sur le parc de logements des zones d'intérêt.

Résultats obtenus pour la métropole

En utilisant la macro SAS évoquée ci-dessus pour la variable « population » c'est-à-dire le nombre de personnes du ménage au niveau national, on montre que, du fait de l'existence de sondage pour la population des ménages des grandes communes, la population française de métropole est connue à plus ou moins 15 800 personnes soit à 0,05 % près (15800*2 / population française = 0,05 %).

Au niveau régional (voir tableau 2), le coefficient de variation (CV) de cette même variable varie entre 0,03 % et 0,45 % selon les régions. Les régions ayant le CV le plus important correspondent logiquement aux régions les moins peuplées. Le coefficient de variation médian est de 0,16 %.

Au niveau départemental, l'étendue des coefficients de variation est plus élevée : il varie entre 0,03 % et 1,15 %. Ce sont également les départements les moins peuplés qui ont les CV les plus importants.

Au niveau communal, pour 876 communes de 10 000 habitants ou plus sur 892 au *RP* 2006 (soit 98 % des grandes communes), la précision obtenue avec ce nouveau calcul est inférieure ou égale à celle estimée selon le rapport Cnis de décembre 2005 à partir des données de 1999, ce qui permet de confirmer, tout en l'affinant, l'information dont on disposait jusque là en termes

de précision liée au sondage du recensement rénové (voir tableaux 3 et 4).

Pour l'ensemble des variables de diffusion du *RP* 2006, les coefficients de variation ont été calculés au niveau Iris. On distingue trois grands types d'Iris :

- Les Iris d'habitat : leur population se situe en général entre 1 800 et 5 000 habitants. Ils sont homogènes quant au type d'habitat et leurs limites s'appuient sur les grandes coupures du tissu urbain (voies principales, voies ferrées, cours d'eau, etc.). Ils formaient 92 % des Iris au 1er janvier 2008, géographie de diffusion du *RP* 2006.

- Les Iris d'activité : ils regroupent plus de 1 000 salariés et comptent au moins deux fois plus d'emplois salariés que de population résí-

Tableau 2
Distribution des coefficients de variation de la variable population aux niveaux régional et départemental

	Distribution du CV de la variable population au niveau région	Distribution du CV de la variable population au niveau département
100 % Max	0,45	1,15
99 %	0,45	1,15
95 %	0,26	0,94
90 %	0,25	0,70
75 % Q3	0,20	0,53
50 % Médiane	0,16	0,35
25 % Q1	0,12	0,25
10 %	0,09	0,15
5 %	0,05	0,13
1 %	0,03	0,03
Min	0,03	0,03
Moyenne	0,17	0,40

Lecture : au niveau régional, le CV de la variable population varie entre 0,03 % et 0,45 % selon les régions. 95 % des régions ont un CV de la variable population inférieur à 0,26 % ; pour la moitié d'entre elles, le CV est inférieur à 0,16 %, etc. En moyenne, le CV est de 0,17 %.
Champ : communes métropolitaines de 10 000 habitants ou plus.
Source : RP *2006, Insee.*

Tableau 3
Distribution des coefficients de variation de la variable population pour les communes de 10 000 habitants ou plus

	Distribution du CV de la variable population au niveau commune
100 % Max	6,62
99 %	2,07
95 %	1,42
90 %	1,22
75 % Q3	1,05
50 % Médiane	0,88
25 % Q1	0,71
10 %	0,55
5 %	0,46
1 %	0,32
Min	0,23
Moyenne	0,91

Lecture : au niveau communal, le CV de la variable population varie entre 0,23 % et 6,62 % selon les communes. 95 % des communes ont un CV de la variable population inférieur à 1,42 % ; pour la moitié d'entre elles, le CV est inférieur à 0,88 % (médiane), etc. En moyenne, le CV est de 0,91 %.
Champ : communes métropolitaines de 10 000 habitants ou plus.
Source : RP *2006, Insee.*

dente. Au 1er janvier 2008, ils représentaient 5 % des Iris.

- Les Iris divers : il s'agit de grandes zones spécifiques peu habitées et ayant une superficie importante (parcs de loisirs, zones portuaires, forêts, etc.).

Les calculs ont montré que le coefficient de variation est très variable d'un Iris à un autre et d'une variable à une autre. Ainsi la moyenne des CV calculée au niveau des Iris s'établit à près de 3 % pour la variable population et à 13 % pour le nombre de chômeurs. Pour l'essentiel, ces différences dépendent des effectifs concernés mais sont également fortement liées à la répartition spatiale des caractéristiques mesurées sur la population en question, donc de leur plus ou moins grande concentration.

Il a été décidé de diffuser sur le site insee.fr[18] en même temps que les données une table par variable donnant un CV « résumé » pour une dizaine de tranches d'effectifs. En revanche, diffuser un CV pour chaque variable de chaque Iris n'a pas été retenu car l'information calculée à partir du recensement risquait d'être fluctuante d'une année sur l'autre.

Deux options ont été envisagées pour obtenir le CV « résumé » :

- La première option consistait pour une variable donnée à déterminer une relation linéaire entre le CV et le logarithme de l'effectif de la variable à partir des CV de tous les Iris. Cette relation était alors appliquée aux effectifs des bornes inférieures des tranches retenues en termes de diffusion pour la variable considérée ;

- La seconde consistait à calculer le CV moyen des Iris pour une variable donnée pour chacune des tranches d'effectifs retenues.

La première option a été testée pour certains thèmes de variables mais pas tous car elle conduisait parfois à des CV négatifs, la liaison entre le CV et l'effectif n'étant pas suffisamment forte. C'est donc la seconde option reposant sur le CV « moyen » qui a été privilégiée pour la plus grande partie des thèmes. Elle conduit d'ailleurs à des résultats proches de ceux obtenus avec la première option et repose sur un concept plus simple à expliciter aux utilisateurs.

Ainsi, des tableaux contenant les CV résumés, et dont un extrait est disponible ci-dessous (voir tableau 5), ont été mis à disposition sur le site insee.fr. Les données des tableaux 5 et 6 présentés ci-dessous ne correspondent pas exactement à celles actuellement disponibles sur le site insee.fr, une actualisation des calculs de précision des résultats du recensement de la population ayant été réalisée par l'Insee en 2015, après la rédaction de cet article.

La table des CV « résumés » par variable et tranche de taille ne peut cependant pas servir de référence universelle sur la précision pour tous les Iris. En effet, certains d'entre eux ne sont pas de qualité suffisante. Par conséquent, un second indicateur appelé « label » a été calculé pour chaque Iris afin d'indiquer à l'utilisateur si les CV peuvent être utilisés ou non pour les Iris qui l'intéressent, et dans certains cas lui signaler qu'un regroupement d'Iris est nécessaire s'il souhaite avoir une précision suffisante. Le label d'un Iris donné permettant de qualifier l'utilisation possible des données de niveau Iris est construit à partir de la précision relative de la variable population des ménages de cet Iris, et de la stabilité de la population de

18. Lien au 1er janvier 2014 : http://www.insee.fr/fr/bases-de-donnees/default.asp?page=recensements.htm.

Tableau 4
Distribution des coefficients de variation de la variable population par tranche de taille pour les communes de 10 000 habitants ou plus

	Communes de 10 000 à 19 999 hab. 452 communes	Communes de 20 000 à 49 999 hab. 318 communes	Communes de 50 000 à 99 999 hab. 79 communes	Communes de 100 000 hab. et plus 43 communes
75 % Q3	1,16	0,87	0,71	0,43
50 % Médiane	1,02	0,78	0,56	0,39
25 % Q1	0,90	0,68	0,50	0,34

Lecture : pour 75 % des communes de 10 000 à 19 999 habitants, le CV de la variable population est inférieur à 1,16% ; pour 50 % d'entre elles, il est inférieur à 1,02 % et pour 25 % d'entre elles à 0,90 %.
Champ : communes métropolitaines de 10 000 habitants ou plus.
Source : RP *2006, Insee.*

l'Iris appréhendée au travers de celle du nombre de personnes par logement.

Trois modalités sont prises par l'indicateur appelé « Label » (voir tableau 6) :

- Les Iris avec « Label = 1 » : l'utilisation de la table pour estimer la précision d'une donnée à l'Iris est possible. Ces Iris sont les plus nombreux, ils représentent 88 % des Iris d'habitat des communes de 10 000 habitants ou plus. Au vu de la précision et de la stabilité de l'estimation de leur population, leur échantillon a été jugé d'une qualité qui autorise le calcul et l'utilisation de la table de précision ;

- Les Iris avec « Label = 2 » : la préconisation est de les regrouper en ensembles de plusieurs Iris équivalents à un Triris[19]. La table est alors utilisable pour estimer la précision des résultats ainsi regroupés. Pour ces Iris, l'échantillon n'a pas été jugé suffisamment représentatif. C'est en général lié à une structure de l'habitat peu homogène. En regroupant cet Iris avec d'autres, il perd de sa spécificité et l'échantillon de l'ensemble recouvre une représentativité correcte. Environ 12 % des Iris d'habitat sont dans cette classe ;

- Les Iris avec « Label = 3 » : ils sont atypiques, mal représentés par l'échantillon parfois en

19. Le Triris a été créé en 1999 pour la diffusion de variables sensibles du recensement pour lesquelles l'Iris apparaît insuffisant pour garantir le secret statistique. Un Triris est un regroupement d'Iris (en général 3 Iris).

Tableau 5
Extrait de table mise à disposition sur le site insee.fr donnant des éléments sur la précision des données

Données infra-communales – Diplômes – Formation							
Précision des variables							
Pour les Iris appartenant à une commune de France métropolitaine de 10 000 habitants ou plus							
© Insee Source : Insee, *Recensement de la population* 2007 exploitation principale.							
Variable	Libéllé	Précision (coefficient de variation en %) selon la tranche d'effectif de la variable					
VAR_ID	VAR_LIB	VAR00_49	VAR50_99	VAR100_199	VAR200_299	VAR300_499	VAR500_699
P07_SCOL0205	Pop scolarisée 2-5 ans	> 18	18	13	8	< 8	so
P07_SCOL0614	Pop scolarisée 6-14 ans	> 18	18	15	11	9	7

Lecture : cette table est utilisée de la façon suivante : par exemple, si pour un Iris donné l'effectif de « la population scolarisée des 6-14 ans » est de 324, la table donne, pour cette variable et la tranche d'effectif « 300-499 », un coefficient de variation de 9 %. L'effectif a donc 95 % de chances de se trouver dans l'intervalle [267; 381].
Champ : communes métropolitaines de 10 000 habitants ou plus.
Source : RP *2007, Insee.*

Tableau 6
Extrait de table donnant le label des Iris

Données infra-communales – Diplômes – Formation							
France - Iris							
Découpage géographique au 01/01/2009							
© Insee Source ; Insee, *Recensement de la population* 2007 exploitation principale.							
Iris	Région	Libellé commune ou ARM	Libellé de l'Iris	Type d'Iris	Label de l'Iris en 2007	Pop scolarisée 2-5 ans en 2007 (princ)	Pop scolarisée 6-14 ans en 2007 (princ)
	REG	LIBCOM	LIBIRIS	TYP_IRIS	LAB_IRIS	P07_SCOL0205	P07_SCOL0614
010040101	82	Ambérieu-en-Bugey	Les Perouses-Triangle d'Activite	H	1	56	110
010040102	82	Ambérieu-en-Bugey	Longeray-Gare	H	1	117	393
010040201	82	Ambérieu-en-Bugey	Centre-St Germain-Vareilles	H	1	114	345
010040202	82	Ambérieu-en-Bugey	Tiret-Les Allymes	H	1	152	559

Lecture : l'Iris n° 010040101 libellé « Les Perouses – Triangle d'activité » situé dans la commune d'Ambérieur-en-Bugey est un Iris d'habitation dont le label est de 1.
Champ : communes métropolitaines de 10 000 habitants ou plus.
Source : RP *2007, Insee.*

raison du mode de tirage mais surtout en raison de leurs caractéristiques. Les informations les concernant peuvent être entachées d'une très forte imprécision et la table n'est pas utilisable. Il s'agit essentiellement d'Iris d'activité et d'Iris divers. Quelques dizaines d'Iris d'habitat dont la structure est très particulière sont également dans ce cas.

Cas particulier des DOM

Pour compléter la macro SAS de calcul de précision disponible depuis 2009 pour la métropole, une seconde macro SAS a été écrite à l'été 2010 afin de tenir compte de la spécificité du plan de sondage des DOM. Pour les Iris appartenant à une commune de 10 000 habitants ou plus des DOM, le calcul des coefficients de variation est disponible pour la seule variable de population. L'équivalent des tables de précision pour les autres variables calculées pour les Iris des communes de 10 000 habitants ou plus de France métropolitaine n'existe pas encore pour les Iris des communes de 10 000 habitants ou plus des DOM.

Néanmoins, un label de qualité a également été attribué aux Iris des communes de 10 000 habitants ou plus des DOM selon le même principe que pour ceux de France métropolitaine ; il permet de distinguer trois cas parmi ces Iris :

- Pour les Iris avec « Label = 1 » (environ 70 % des Iris de ces communes), la précision et la stabilité de l'estimation de la population sont d'une qualité suffisamment bonne. Ceci autorise l'utilisation des informations fournies au niveau de ces Iris ;

- Pour les Iris avec « Label = 2 » (environ 25 % des Iris de ces communes), la précision et la stabilité de l'estimation de la population ne sont pas d'une qualité suffisante pour que les données fournies au niveau de ces Iris soient « utilisables ». C'est en général lié à une structure de l'habitat peu homogène. La préconisation est de regrouper ces Iris en un ensemble de plusieurs Iris, ce qui permet l'utilisation des informations au niveau de ces regroupements ;

- les Iris avec « Label = 3 » (environ 5 % des Iris de ces communes) sont atypiques. Ils sont trop peu peuplés pour pouvoir être correctement représentés par l'échantillon. La précision et la stabilité de l'estimation de la population sont de mauvaise qualité, même en regroupant ces Iris. Il s'agit essentiellement d'Iris d'activité ou d'Iris divers. Quelques Iris d'habitation dont la structure est très particulière sont également dans ce cas.

Il faut souligner que les trois modalités proposées pour les communes de 10 000 habitants ou plus des DOM n'ont pas tout-à-fait le même sens que pour celles de France métropolitaine, le niveau d'exigence sur la précision de l'estimation de la population étant moindre pour les Iris des communes de 10 000 habitants ou plus des DOM. Les particularités de l'habitat des DOM, qui pouvait évoluer plus vite qu'en métropole lorsque le recensement rénové a été mis en place, font que le plan de sondage du recensement rénové dans les DOM n'inclut pas de phase de calage et que la précision obtenue est globalement moins bonne que pour les communes de métropole (le CV pour la variable de population au niveau de la commune est en moyenne de l'ordre de 1,1 dans les DOM contre 0,9 pour les communes de métropole) : on ne peut de ce fait pas utiliser exactement les mêmes seuils en termes de CV et de stabilité de l'estimation que ceux retenus pour déterminer les labels des communes de métropole.

* *
*

Mises en perspective des calculs de précision

Les programmes développés en 2009-2010 afin d'estimer au mieux la précision liée au sondage ont permis de confirmer les premiers calculs effectués à partir de simulations de 2005 sur le recensement de 1999 et de les généraliser à l'ensemble de toutes les variables. Cependant, les résultats donnés dans les deux parties précédentes proviennent pour la plupart de calculs menés sur un seul millésime de données du recensement rénové (le premier recensement rénové en référence au 1er janvier 2006, le *RP* 2006), essentiellement pour des raisons de charge de travail de l'équipe en charge des calculs de populations légales et parce que les données du deuxième cycle complet (*RP* 2011) n'étaient pas encore disponibles à ce moment-là. Les coefficients de variation de la population des ménages des grandes communes de métropole ont toutefois été calculés pour plusieurs millésimes de recensement et les résultats obtenus sont stables d'un millésime à l'autre. Ces calculs sont en cours d'actualisation grâce à la disponibilité du recensement de la population de 2011. En effet, avec ce système de recensement tournant

basé sur des cycles de collecte de cinq années, les données qui ont servi à calculer les populations du recensement de 2007 sont en partie les mêmes que celles qui ont servi à calculer celles de 2006. De ce fait, les comparaisons ne sont rigoureusement possibles qu'entre des millésimes espacés d'au moins cinq ans, donc avec la comparaison entre les *RP* 2011 et 2006.

D'autres travaux permettraient d'étoffer la mesure de l'impact de l'introduction du sondage sur la qualité des résultats du recensement rénové. En particulier :

- Un calcul systématique pour toutes les variables au niveau Iris dans les DOM reste à faire, afin de construire la table de précision qui fait défaut pour l'instant ;

- Un calcul de la précision des résultats d'une enquête annuelle de recensement (et non du cumul de cinq enquêtes) pourrait être intéressant pour les utilisateurs de ces données, internes à l'Insee.

D'un point de vue méthodologique, il serait également intéressant, d'une part, de s'interroger sur l'apport en termes de précision de la phase de calage à l'Iris qui est réalisée et, d'autre part, de comparer les calculs de précision obtenus avec ceux que l'on aurait obtenu sous un plan de sondage aléatoire simple, afin d'estimer l'ordre de grandeur de l'effet lié au plan de sondage du recensement rénové.

Toutefois, il n'a pas été jugé prioritaire jusqu'ici d'affiner ces travaux de mesure de la précision liée au sondage, relativement coûteux à réaliser en termes de temps étant donné la taille de la base de données et la complexité du plan de sondage, car l'ordre de grandeur obtenu pour cette imprécision est négligeable et ne dégrade pas la qualité des résultats du recensement. □

BIBLIOGRAPHIE

Bertrand Ph., Chauvet G., Christian B. et Grosbras J.-M. (2002), « Les plans de sondage du nouveau recensement & Données produites par le recensement rénové de la population », *Actes JMS*, 2002.

Cnis (2005), « Utilisation des données produites par le recensement rénové de la population et leur diffusion », *Rapport du Cnis*, n° 98.

Coeffic N. (1993), « L'enquête post-censitaire de 1990, une mesure de l'exhaustivité du recensement », *Population*, n° 6, pp. 1655-1682.

Deville J.-C. et Särndal C.-E. (1992), « Calibration Estimators in Survey Sampling », *JASA*, n° 87, pp. 376-382.

Deville J.-C. et Tillé Y. (2005), « Variance approximation under balanced sampling », *Journal of Statistical Planning and Inference,* vol. 128, pp. 569-591.

Durr J.-M. et Dumais J. (2002), « La rénovation du recensement français », *Techniques d'enquête de Statistique Canada*, vol. 28, n° 1.

Godinot A. (2003), « La rénovation du recensement de la population », *Courrier des statistiques*, Insee, n° 105-106, pp. 5-12.

Godinot A. (2005), « Pour comprendre le recensement », *Insee méthodes*, mai.

Henry L. (1949), « Le contrôle des recensements », *Population*, 4e année, n° 2, 1949, pp. 231-248.

Martal E. et Jacquin Y. (2012), « Qualité du codage de l'activité et de la profession dans le recensement de la population », *Actes JMS*, 2012, Insee.

Tillé Y. (2011), « Ten years of balanced sampling with the cube method: an appraisal », *Survey Methodology*, vol. 7, n° 2, pp. 215-226.

Valente P. (2010), « Comment la population est-elle recensée dans les pays européens en 2010 ? », *Population et sociétés*, n° 467.