La repondération des enquêtes annuelles de recensement pour une diffusion complémentaire du RP

Pierre-Arnaud Pendoli[1], Sébastien Hallépée[2] & Olivia Sautory[3]

[1] *INSEE - Direction Générale, 88 Avenue Verdier, 92120 Montrouge - Timbre F520,* pierre-arnaud.pendoli@insee.fr
[2] *INSEE - Direction Générale, 88 Avenue Verdier, 92120 Montrouge - Timbre F520,* sebastien.hallepee@insee.fr
[3] *INSEE - Direction Générale, 88 Avenue Verdier, 92120 Montrouge - Timbre F520,* olivia.sautory@insee.fr

Résumé. Le dernier recensement de la population (RP) date de 1999. Depuis 2004, la France a rénové son opération de recensement qui repose désormais sur un cycle de cinq ans : chaque année, tout ou partie d'un groupe est recensé au cours des enquêtes annuelles de recensement (EAR). Un RP repose ainsi sur l'agrégation des données de 5 EAR.
Les groupes de rotation, sur lesquels sont construites les EAR, sont parfois très déséquilibrés en termes de nombre de logements et de population. On observe ainsi dans certaines régions d'importantes fluctuations dans les séries de populations calculées à partir des EAR, freinant jusqu'à présent une utilisation plus poussée de cette source.
L'exploitation des données d'une seule EAR offre pourtant plusieurs avantages, notamment la fraîcheur des données, l'unicité de la période d'observation et une plus grande réactivité en cas de modifications du questionnaire. Par exemple, une modification du questionnaire de la feuille de logement a lieu en 2018. Une nouvelle analyse ménage-famille pourra conduire à des études dès 2019 à partir de l'EAR, contre 2023 si on devait utiliser le RP.
Nous présentons une nouvelle méthodologie de calage des EAR visant à améliorer la précision de ces données. Celle-ci repose notamment sur l'utilisation de variables auxiliaires issues du Fichier démographique des logements et des individus (Fidéli) basé sur des sources fiscales.
La nouvelle pondération réduit les écarts entre les estimations issues de l'EAR 2014 et celles du RP 2014. Elle permet aussi de lisser les fluctuations des séries de population et de nombre de logements dues aux déséquilibres des groupes de rotation.

Mots-clés. Calage, recensements, fichiers administratifs

Abstract. The last Population Census was held in 1999. Since 2004, French census has been renewed and is based on a 5-year cycle : each year, all or part of a group is collected during the annual census surveys. A Population Census is based on the aggregation of 5 annual census surveys data.
The rotation groups, on which the annual census surveys are built, are sometimes imbalanced in terms of number of dwellings and inhabitants. In some regions, there has also been considerable fluctuations in the series of population calculated from the annual census surveys. This problem has so far prevented a more elaborate use of these surveys.
However, using the data of the annual census survey provides several advantages, in particular the timeliness of data, the uniqueness of the observation period and an enhanced responsiveness in case of a modification of the questionnaire. For instance, a modification of the housing questionnaire happened in 2018. By using the annual census surveys data, studying complex family relationships would be possible from 2019. On the other hand, by using the Population

Census, that is a complete cycle of five annual census surveys, theses studies would be postponed to 2023.
We will present a new calibration methodology of the annual census surveys whose aim is to improve the estimations accuracy. This new method is based on the use of auxiliary variables from the new tax system used at INSEE and called "Demographic file of housings and individuals".
The new calibrated weights reduce the gap between the estimations from 2014 annual survey and the estimations from 2014 Population Census. Besides, the new calibration methodology enables to smooth the fluctuations in the series of population that are caused by the imbalance of the rotation groups.

Keywords. Calibration, Census, administrative files

1. Introduction

Le recensement de la population est basé sur un cycle de cinq ans depuis 2004. Ce changement a permis d'une part de répartir les coûts de collecte dans le temps et d'autre part de pouvoir diffuser de nouvelles populations et de nouveaux résultats statistiques basés sur le recensement chaque année et pour chaque commune depuis 2008.

Pour ce faire, on réalise chaque année la collecte du recensement d'un cinquième des communes de moins de 10 000 habitants (petites communes) de manière exhaustive et, sur toutes les communes de 10 000 habitants et plus (grandes communes), d'un échantillon d'adresses représentant 8 % des logements de la commune. Les résultats du RP sont établis en cumulant les informations collectées sur un cycle de cinq années successives.

Le protocole de collecte est adapté aux catégories de population à enquêter. Les individus faisant partie de ménages ordinaires représentent la très grande majorité (97 %) de la population. Les individus résidant dans des communautés sont recensés selon un protocole spécifique. En particulier, la collecte des communautés est la plupart du temps concentrée sur une seule des cinq années du cycle pour la plupart des grandes communes. De même, les individus vivant dans des habitations mobiles et les sans-abris ne sont pas recensés de la même façon. Enquêtés exhaustivement tous les cinq ans, ils l'ont été par exemple en 2016 pour toutes les grandes communes.

L'opération du recensement comporte des spécificités dans les départements d'outre-mer où la mise à jour de la base de sondage des adresses à enquêter dans les grandes communes fait appel à une enquête cartographique menée sur une partie du territoire alors qu'en métropole, les mises à jour du répertoire d'adresses sont initiées par le suivi des permis et les échanges avec les communes.

Ces particularités impliquent que la méthode d'estimation doit être adaptée selon la taille de la commune (plus ou moins de 10 000 habitants), la catégorie de population (ménages ordinaires, communautés, habitations mobiles et sans abris) et le territoire (métropole ou DOM).

L'enquête annuelle de recensement (EAR) a donc été considérée pendant longtemps surtout comme un élément du cumul de cinq années successives de recensement. Néanmoins, l'EAR constitue depuis 2004 un produit de diffusion à part entière. Il est actuellement peu mis en avant à l'extérieur de l'Insee et du service statistique public et son utilisation est assortie de recommandations d'utilisation restrictives. En effet, le produit est largement perfectible et conduit à des séries de résultats très volatils, même pour des indicateurs assez agrégés comme la population des régions. L'exploitation des données d'une seule EAR offre pourtant plusieurs avantages, notamment la fraîcheur des données, l'unicité de la période d'observation et une plus grande réactivité en cas de modifications du questionnaire. Par exemple, une modification du questionnaire de la feuille de logement a lieu en 2018. Une nouvelle analyse ménage-famille pourra conduire à des études dès 2019 à partir de l'EAR, contre 2023 si on devait utiliser le RP.

2. Équilibrage imparfait des groupes de rotation

2.1 Plans de sondage équilibrés pour les groupes de rotation

Le tirage des cinq groupes de rotation du recensement en continu a été réalisé au début des années 2000 selon un plan de sondage spécifique pour les petites communes, d'une part, et pour les grandes communes, d'autre part. Modulo les changements de géographie et la création d'adresses nouvelles en grandes communes, les groupes de rotation ne sont pas modifiés au cours du temps.

- **Pour les petites communes**

La commune constitue l'unité statistique. Les petites communes sont stratifiées selon leur région d'appartenance.

Au sein de chaque région les petites communes sont réparties aléatoirement en cinq échantillons, par tirage équilibré à probabilités égales (1/5) sur des variables de type logement et des variables socio-démographiques données par le RP 1999, notamment la population associée à chaque département.

- **Pour les grandes communes**

Chaque grande commune fait l'objet d'un plan de sondage indépendant. L'adresse constitue l'unité statistique.

Les adresses sont réparties aléatoirement en cinq échantillons, par tirage aléatoire équilibré à probabilités égales (1/5) sur les mêmes variables que celles utilisées pour les petites communes, hors la population par département.

2.2 Déséquilibre empirique des groupes de rotation

Si l'équilibrage des groupes de rotation vis-à-vis de ces variables était exact, on devrait pouvoir estimer parfaitement la population au 1er janvier 1999 de chaque département.

Toutefois, lorsque l'on tente d'estimer pour chaque département la population des ménages au RP 1999 à partir des groupes de rotation, il apparaît que ces estimations ne sont pas exactes. D'après la figure 1, la majorité des estimations de populations départementales par groupe de rotation (GR) s'écartent de plus de 4% par rapport au vrai total au 1er janvier 1999.

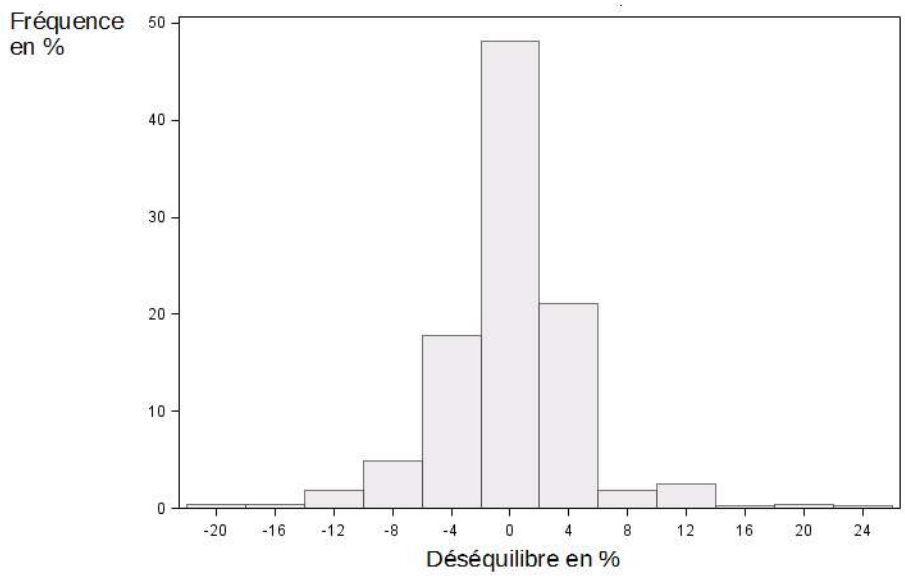

Figure 1 : Déséquilibres (en %) des GR vis-à-vis de la population par département au 01/01/1999.

Pour 9 départements sur 10, les petites communes contribuent à plus de 80% du déséquilibre des groupes de rotation vis-à-vis de la population totale du département. Cela s'explique par le fait qu'au niveau départemental, chaque groupe de rotation peut contenir un nombre relativement faible de petites communes, en tout cas plus faible que le nombre d'adresses de chaque groupe de rotation en grandes communes. Calculées à partir d'échantillons de tailles plus faibles qu'en grandes communes, les estimations de population en petites communes se caractérisent par une variance plus forte au niveau départemental.

2.3 Volatilité des estimations de population à partir de l'EAR

Comme le montre la figure 2 pour la région Bourgogne-Franche-Comté, ces déséquilibres des groupes de rotation en termes de population sont à l'origine d'importantes fluctuations des estimations du nombre d'habitants par région entre les EAR. Dans l'état actuel des choses, il est donc impossible d'estimer le niveau de la population et son évolution à partir de l'EAR. Pour l'instant, seules les estimations du RP, qui cumulent les résultats de cinq EAR successives, peuvent être utilisées à cet effet.

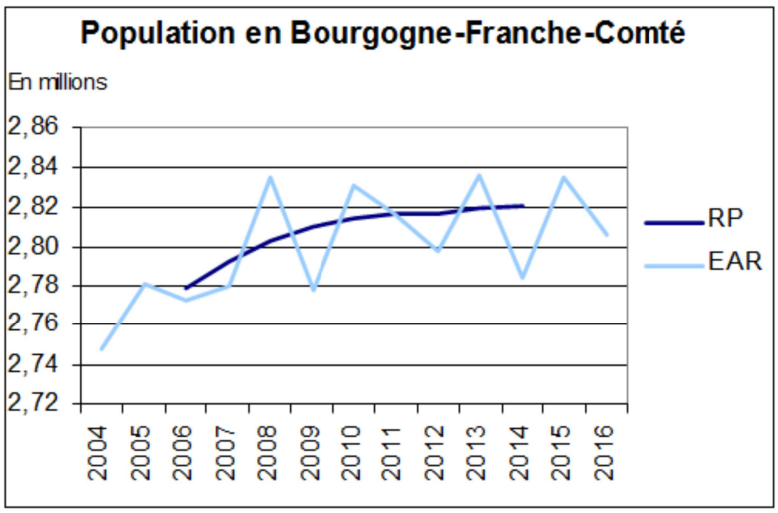

Figure 2 : Évolution des estimations de la population de la Bourgogne-Franche-Comté entre 2004 et 2016, à partir du RP d'une part et de l'EAR d'autre part.

3. Nouvelle méthode de calage des EAR

3.1 Identification des champs de l'EAR

Les plans de sondage de l'EAR diffèrent selon les champs. Pour corriger les déséquilibres associés au plan de sondage initial, il est naturel que les unités de calage coïncident avec les unités d'échantillonnage.

- **Échantillons de petites communes**

Pour résorber les déséquilibres constatés, nous proposons une méthode de calage des groupes de rotation régionaux des petites communes sur ces mêmes variables d'équilibrage actualisées.

Dans le cadre des travaux de repondération de l'EAR 2018 par exemple, les petites communes de chaque région administrative[1] recensées en 2018 forment un échantillon de communes à caler. On a donc autant d'échantillons de petites communes qu'il existe de régions.

[1] Il s'agit des régions administratives issues de la réforme territoriale de 2015.

- **Échantillons d'adresses en grandes communes**

En revanche, dans chaque grande commune, on sélectionne pour chaque enquête annuelle de recensement un échantillon d'adresses à recenser. Dans le cadre des travaux de repondération de l'EAR 2018, les adresses d'habitation recensées en 2018 forment un échantillon d'adresses à caler pour chaque grande commune. On a donc *a priori* autant d'échantillons d'adresses qu'il existe de grandes communes.

3.2 Détermination de la source auxiliaire

Une fois déterminés les échantillons de l'EAR, il est nécessaire d'identifier une source auxiliaire permettant d'améliorer les estimations d'une EAR. Cette base de calage doit répondre aux critères suivants :

- Des variables de type socio-démographique telles que celles présentes dans le recensement doivent être disponibles et mises à jour régulièrement de manière exhaustive.
- Les totaux de ces variables auxiliaires doivent pouvoir être calculés à un niveau agrégé (i.e. les régions pour l'échantillon de petites communes ; les communes pour les échantillons d'adresses des grandes communes).
- Les variables auxiliaires doivent être connues pour chaque unité de l'échantillon (i.e. pour chaque petite commune d'une part et pour chaque adresse échantillonnée en grande commune d'autre part).

Pour le calage de l'EAR 2018 par exemple, il serait possible de recourir au dernier millésime de RP disponible, i.e. le RP 2015 qui agrège les cinq EAR de 2013 à 2017. Mais cela impliquerait de caler l'échantillon de l'EAR sur des totaux ayant trois ans d'ancienneté. En revanche le Fichier démographique des logements et des individus (Fidéli) rassemble des données exhaustives issues des sources fiscales (cadastre, taxe d'habitation, impôt sur le revenu…) n'ayant qu'une seule année d'ancienneté par rapport à l'EAR. Le calage de l'EAR 2018 utilise ainsi le millésime 2017 de Fidéli. Par ailleurs, on retrouve dans Fidéli des variables auxiliaires similaires à celles qui ont été utilisées dans les plans de sondage des groupes de rotation, telles que le sexe et l'âge des individus.

3.3 Calage de l'échantillon de petites communes

En petites communes, le calcul des variables auxiliaires est simple puisque le code officiel de la commune est toujours présent dans Fidéli. Pour le calage de l'EAR 2018 par exemple, la méthode consiste à suivre les étapes suivantes.

- Pour chaque petite commune recensée en 2018, on calcule les variables auxiliaires suivantes à partir des données de Fidéli 2017 : nombre total de logements ; nombre total de logements en immeubles collectifs ; nombre total de personnes selon certaines classes d'âge ; nombre total de femmes ; nombre total d'hommes ; nombre total d'habitants.
- Pour chaque région administrative, on calcule les marges de calage associées à chaque variable auxiliaire à partir des données de Fidéli 2017. Concernant le nombre total d'habitants, on calcule une marge pour chaque département de la région.
- Un calage de petites communes est réalisé pour chaque région. Afin de nous prémunir contre une trop grande dispersion des poids calés, nous utilisons la méthode Logit qui permet un contrôle des déformations maximales des poids. Les paramètres de la macro SAS %Calmar sont choisis de façon à ce que les rapports de poids soient compris entre 0,25 et 2, ce qui garantit que les nouveaux poids soient compris dans l'intervalle [1,25 ; 10][2], évitant ainsi d'avoir des unités trop influentes.

[2] Chaque petite commune recensée a en effet un poids de sondage égal à 5.

3.4 Calage de l'échantillon d'adresses en grandes communes

En grandes communes, la constitution des variables auxiliaires pour chaque unité de l'échantillon n'est pas immédiate. En effet, il n'existe pas d'identifiant commun entre Fidéli et l'échantillon d'adresses du recensement. Le tirage de l'échantillon annuel d'adresses pour le recensement est réalisé parmi les adresses de la Base de Sondage d'Adresses (BSA) issue du Répertoire d'Immeubles Localisés (RIL) mis à jour chaque année, et qui appartiennent au groupe de rotation de l'année. Pour être en mesure de créer les variables de calage à partir de Fidéli pour chaque adresse de l'échantillon, un appariement est réalisé avec la BSA. Pour le calage de l'EAR 2018 par exemple, la méthode consiste à suivre les étapes suivantes :

- **Appariement de la BSA 2018 avec Fidéli 2017**. La diversité et la complexité des types d'adresses nécessitent de recourir à un algorithme permettant d'apparier à Fidéli environ 90 % des adresses de la BSA tout en évitant les faux appariements. L'appariement est réalisé sur la base de variables d'adressage que l'on trouve à la fois dans Fidéli et dans la BSA : code commune, identifiant de la voie (code rivoli), numéro dans la voie, suffixe (bis, ter…), référence cadastrale.
- **Calcul des variables auxiliaires composites** : nombre total de personnes selon certaines classes d'âge ; nombre total de femmes ; nombre total d'hommes ; nombre total d'habitants ; nombre total de logements.
    - Dans tous les cas, le nombre total de logements que l'on utilise en tant que variable auxiliaire est le nombre de logements de la BSA corrigé de la collecte. En effet, il s'agit de la seule variable auxiliaire dont on dispose de manière exhaustive dans la BSA et qui est mise à jour annuellement dans le RIL. Il n'est donc pas utile d'utiliser le nombre de logements issu de Fidéli.
    - Pour la fraction (89%) de la BSA 2018 appariée à Fidéli, les autres variables auxiliaires sont calculées pour chaque adresse à partir de Fidéli 2017.
    - Pour la fraction qui n'a pas été appariée, les variables auxiliaires sont calculées à partir de la dernière collecte de recensement (9%) ou sont imputées (2%). Dans ce dernier cas, pour le nombre d'hommes par exemple, on calcule sur les adresses appariées avec Fidéli ou précédemment enquêtées au recensement la moyenne suivante :

$$\frac{\text{nombre total d'hommes de la commune}}{\text{nombre total de logements de la commune}}$$

      Pour chaque adresse sujette à imputation, on affecte cette valeur moyenne multipliée par le nombre de logements de l'adresse dans la BSA corrigé de la collecte.
- **Calcul des marges de calage** : pour chaque grande commune, les marges de calage sont les totaux au niveau communal des variables auxiliaires composites calculées sur la BSA.
- **Calage de l'échantillon**. L'échantillon à caler est l'ensemble des adresses recensées en 2018. À chaque adresse sont rattachées les variables auxiliaires composites issues de la BSA enrichie par les traitements décrits précédemment. Pour éviter de réaliser près de 1000 calages[3], on réalise finalement 13 calages, c'est-à-dire un par région, avec des marges communales. Les calages sont réalisés à l'aide de la macro SAS %Calmar dont on précise certains paramètres.
    - On utilise la méthode Logit avec des bornes de rapports de poids identiques pour chaque région : 0,3 et 2,5.
    - Le paramètre POIDS correspond aux poids de sondage des adresses de l'échantillon
    - On renseigne le paramètre PONDQK en entrée du calage :

[3] La France compte près de 1000 grandes communes.

$$\text{pondQK} = \frac{1}{\text{nombre de logements de l'adresse dans la BSA corrigé de la dernière collecte du recensement}}$$

Il s'agit d'une variable de pondération spécifique des adresses de l'échantillon, qui n'est pas liée aux poids de sondage. L'utilisation de ce paramètre est nécessaire pour que le calage converge pour certaines régions, et évite que les plus grandes adresses subissent des déformations de poids importantes. On associe ainsi mieux pour chaque adresse la dimension des bases de données de diffusion qui sont aux niveaux des ménages et des individus.

4. Impact sur la précision des estimations issues de l'EAR

4.1 Réduction de l'écart entre estimations EAR et estimations RP

Pour un paramètre d'intérêt donné, l'utilisation des nouveaux poids permet de réduire globalement les écarts entre l'estimation issue de l'EAR et celle issue du RP. En utilisant les anciennes pondérations l'écart entre l'estimation EAR 2014 et l'estimation RP2014 de la population départementale est supérieure à +/- 2% pour un quart des départements (cf. figure 3). En utilisant les nouvelles pondérations, seuls 9% des départements se caractérisent par de tels écarts.

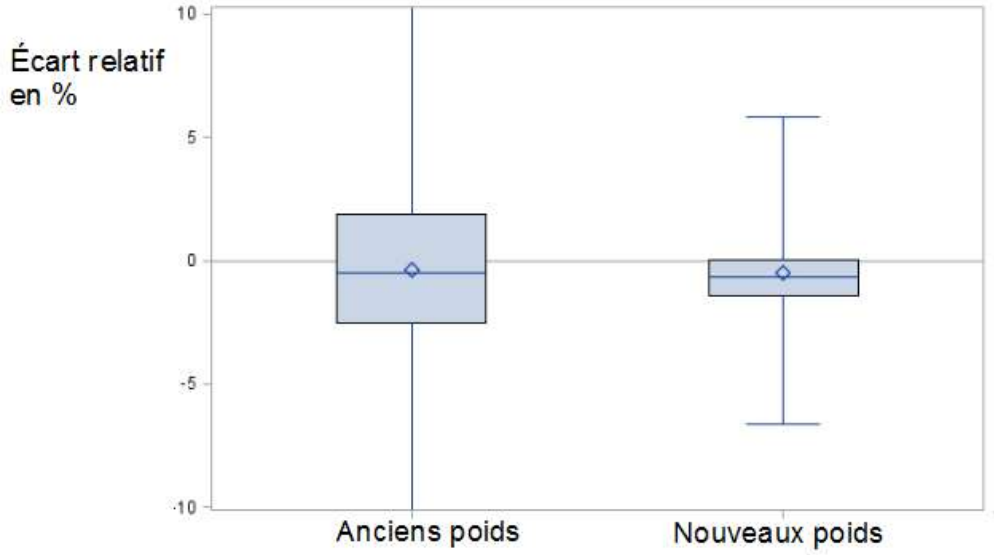

Figure 3 : Distribution des écarts relatifs (en %) entre les estimations de la population départementale obtenues par l'EAR 2014 d'une part et par le RP2014 d'autre part.

Cet « alignement » des estimations de l'EAR sur celles du RP obtenu grâce à l'utilisation des nouvelles pondérations peut également être observé pour l'estimation du total de populations plus spécifiques comme le nombre total d'actifs occupés à temps partiel par département. En revanche, de telles améliorations ne sont pas visibles pour l'estimation de pourcentages. En effet, les estimations issues de l'EAR 2014 utilisant les anciennes pondérations sont déjà proches des estimations du RP. Par exemple, quelque soit le jeu de poids utilisé, pour 95% des départements, l'estimation de la proportion de temps partiel parmi les actifs occupés d'après l'EAR ne diffère du pourcentage issu du RP que de 0,9 point au maximum.

4.2 Réduction de la volatilité des estimations EAR au cours du temps

La difficulté de l'utilisation des EAR pour la réalisation d'études statistiques est liée à la volatilité des estimations d'un paramètre d'intérêt donné au cours du temps, ce qui est dû aux déséquilibres des groupes de rotation. Au niveau régional, le recours aux nouvelles pondérations permet de réduire cette volatilité, introduisant ainsi la possibilité de dégager des évolutions tendancielles de ces paramètres d'intérêt. À l'aide des nouvelles pondérations on peut par exemple observer qu'une légère baisse de la population de la région Bourgogne-Franche-Comté est amorcée depuis 2014 (cf. figure 4). Une telle évolution n'aurait pas pu être observée à l'aide des anciennes pondérations qui ne permettaient de dégager aucune tendance dans la série de population de cette région.

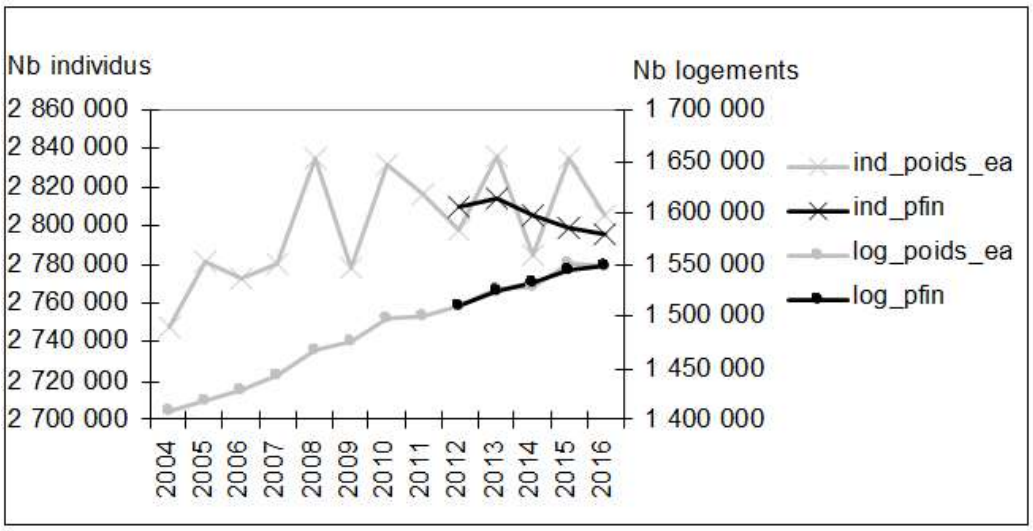

Figure 4 : Évolution des estimations EAR du nombre de logements et du nombre d'individus en Bourgogne-Franche-Comté entre 2004 et 2016, en utilisant les anciennes pondérations (_poids_ea) ou les nouvelles pondérations (_pfin).

Les données de Fidéli sont disponibles à partir du millésime 2011. Il est donc en théorie possible de calculer les nouvelles pondérations à partir de l'EAR 2012. Toutefois l'appariement de la BSA (*resp.* millésimes 2012 et 2013) avec Fidéli (*resp.* millésimes 2011 et 2012) n'est pas de bonne qualité en grandes communes. Ainsi, pour les EAR 2012 et 2013, les nouvelles pondérations ont été calculées uniquement pour le champ des petites communes. En grandes communes, on conserve les anciennes pondérations.

Au niveau communal également (pour les grandes communes), on observe que les nouvelles pondérations permettent de réduire la dispersion des estimations du nombre de logements et d'habitants au cours du temps, comme le montre la figure 5 à propos de la commune du Mée-sur-Seine (77285). Pour chaque grande commune, on ne peut calculer les estimations à l'aide des nouvelles pondérations qu'à partir de l'EAR 2014, étant donné que l'appariement entre la BSA et Fidéli est de qualité insuffisante pour les millésimes antérieurs.

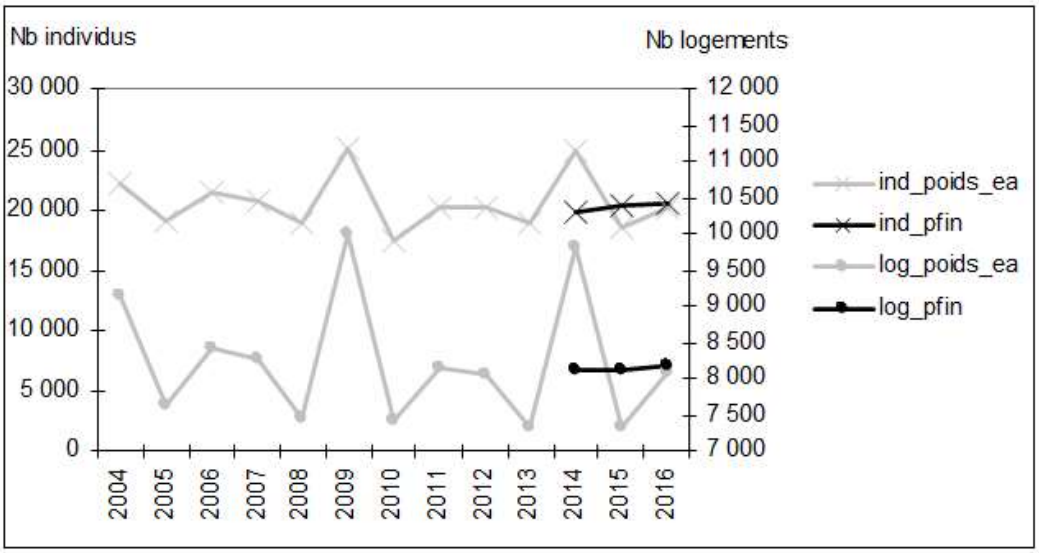

Figure 5 : Évolution des estimations EAR du nombre de logements et du nombre d'individus de la commune du Mée-sur-Seine (77285) entre 2004 et 2016, en utilisant les anciennes pondérations (_poids_ea) ou les nouvelles pondérations (_pfin).

4.3 Réduction de la variance des estimations

Pour compléter l'approche empirique développée ci-dessus, l'impact de l'utilisation de la nouvelle méthode a aussi été mesuré au travers du gain associé à la précision d'échantillonnage.

Comme on pouvait s'y attendre, ce gain est particulièrement important sur le champ des petites communes et pour les indicateurs statistiques qui sont les proxy directs des variables qui ont servi au calage. Le gain de précision est plus réduit sur des populations rares pour lesquelles la corrélation avec les variables de calage est plus faible. Le gain de précision est également plus élevé au niveau départemental qu'aux niveaux régional et national, étant donné que la taille de l'échantillon importante garantissait déjà une précision suffisante sur les niveaux les plus agrégés.

Bibliographie

Bertrand P., Chauvet G., Christian B., Grosbras J.M., « Les plans de sondage du nouveau recensement », VIIIèmes Journées de méthodologie statistique, 16-17 décembre 2002.
Godinot A., « Pour comprendre le recensement de la population », 2005.