

Partie 3 : Des analyses quantitatives de la qualité des estimations de population

Les contrôles présentés dans la partie 2 sont indispensables pour assurer une bonne qualité des estimations de population. Ils permettent notamment de s'assurer que les traitements statistiques sont réalisés de manière correcte. Ils permettent aussi d'encadrer le processus de collecte des données afin d'éviter de s'éloigner de la réalité du terrain. Toutefois, il n'est pas possible de maîtriser l'ensemble des aléas. Notamment, le recensement reste une enquête auto-administrée, et la réponse des personnes recensées est considérée comme exacte. On fait donc implicitement l'hypothèse que toutes les consignes ont été comprises et respectées.

Cette partie prend plus de recul par rapport au processus de production du recensement pour analyser la qualité des estimations produites. Elle cherche à repérer d'éventuelles sur-estimations ou sous-estimations. Pour ce faire, plusieurs moyens peuvent être imaginés. Le plus intuitif serait de réaliser une enquête de mesure de la qualité du recensement qui viserait à retourner dans un échantillon de logements recensés pour vérifier que les déclarations des personnes correspondent bien à ce que l'on a cherché à collecter : déclaration de la bonne catégorie du logement, déclaration du bon nombre d'habitants permanents du logement, bonne déclaration des personnes multi-résidentes, bonne exhaustivité des logements en petite commune, etc. De telles enquêtes ont été réalisées en France à la suite des recensements de 1962 et de 1990. Coûteuses, elles n'ont à ce jour pas été reproduites¹⁹ depuis le recensement rénové.

Le changement de la feuille de logement en 2018 a mis en évidence que l'ancien questionnaire conduisait à comptabiliser deux fois certaines personnes multi-résidentes. Ce phénomène concernait majoritairement des enfants en résidence partagée qui étaient déclarés en tant qu'habitant permanent du logement par chacun de ses deux parents. En 2018, la manière de « déclarer » les enfants en situation de résidence partagée a été modifiée et les consignes ont été éclaircies. L'Insee estime l'effet du changement de questionnaire en 2018 à une baisse d'environ 500 000 personnes sur la population française, soit 0,75 % (cf. notes techniques 2020, Insee).

Faute de disposer de résultats d'une enquête de mesure de la qualité, nous comparons dans cette partie les estimations de population produites à partir du recensement à d'autres bases de données : l'échantillon démographique permanent (EDP), les données de l'état civil et les bases de données fiscales. Ces comparaisons permettent de mettre en évidence certaines divergences entre les sources, qui tantôt s'expliquent par des différences de concept entre les sources et parfois peuvent laisser soupçonner des défauts du recensement par rapport à ce qu'il cherche à mesurer.

19 À l'exception des enquêtes de mesure de la qualité du RIL présentées dans la partie 2-A, qui fournissent un état des lieux de l'exhaustivité de la base de sondage en grande commune qui est utilisée pour les estimations de la population sur ces territoires. Elles ne constituent toutefois qu'une partie du dispositif d'enquête de mesure de la qualité du recensement qu'il faudrait réaliser.

A) Estimation des doubles comptes statistiques au recensement à partir de l'Échantillon démographique permanent

Cette partie a été rédigée par Laurent Toulemon (Institut national d'études démographiques [Ined]), Sébastien Durier (Insee, Direction des Statistiques Démographiques et Sociales, au moment de la rédaction de cet article) et Benjamin Marteau (Ined).

Ce texte reprend partiellement et met à jour des résultats présentés aux Journées de méthodologie statistique en 2018 (Toulemon, Durier et Marteau, 2018), ainsi que le fruit de discussions avec des collègues de l'Insee. Qu'ils en soient ici remerciés. Comme il est d'usage, les erreurs restent les nôtres. Ce travail a bénéficié du soutien de l'Agence nationale de la recherche (subvention "ANR-16-CE41-0007-01") dans le cadre du projet BigStat (<https://big-stat.site.ined.fr>).

A-1. Les doubles comptes dans le recensement rénové

La qualité statistique des recensements se mesure d'abord par la qualité du comptage des habitants. Les deux erreurs de comptage sont les doubles comptes (personnes recensées deux fois) et les omissions (personnes non recensées). À partir de l'enquête post-censitaire de 1990, la proportion de doubles comptes probables a pu être estimée à 0,7-0,9 %, celle des omissions à 1,5-2,0 %. Ces omissions concernaient principalement des logements occupés par une seule personne, qui sont moins facilement accessibles ou parfois moins visibles. Les personnes rattachées à plusieurs logements ordinaires font partie des personnes à risque d'être soit omises soit en double compte. Les doubles comptes ou les omissions touchaient majoritairement les jeunes adultes âgés de 20 à 24 ans (Coeffic, 1993).

Le but de cet article est d'estimer la proportion d'habitants en situation de double compte dans le recensement rénové. Après avoir défini les doubles comptes statistiques et les méthodes permettant de les identifier, nous présenterons rapidement l'Échantillon démographique permanent, source de données qui permet de réaliser plusieurs estimations, avant de discuter les limites et les prolongements de nos analyses.

Recensement exhaustif et recensement fondé sur des enquêtes annuelles

Dans un recensement exhaustif, les doubles comptes peuvent être identifiés facilement grâce à la présence de deux bulletins correspondant au même habitant. Les doubles comptes peuvent être évités grâce à une règle simple consistant à remplir « un bulletin et un seul par personne ». Avec les enquêtes annuelles de recensement (EAR), le taux de sondage varie entre 8 % dans les grandes communes et 20 % dans les petites, une petite commune sur cinq étant recensée intégralement chaque année (Desplanques, 2008). Les notions d'omissions et de doubles comptes doivent alors être adaptées pour pouvoir être mesurées. Dans le recensement rénové, les doubles comptes ne conduisent au remplissage effectif de deux bulletins la même année qu'environ une fois sur sept (entre 8 % et 100 % des cas, selon que les deux logements sont dans la même commune ou non) : il faut que les deux logements soient inclus dans la même vague d'enquête de recensement. Les doubles comptes sont donc le plus souvent invisibles pour les personnes recensées, seule une minorité donnant lieu au remplissage effectif de deux bulletins la même année, que l'on appelle un bulletin double dans le fichier du recensement (Mambetov, 2014). Nous parlerons de doubles comptes statistiques pour les distinguer des bulletins doubles.

En l'absence d'enquête de qualité du recensement, on n'observe pas les erreurs de mesure du recensement. On ne sait pas si une personne non présente dans le recensement aurait dû l'être et nous n'étudierons pas les omissions dans cette partie. On ne sait pas non plus si une personne présente dans le recensement y est recensée à tort. En revanche, on peut affirmer qu'une personne en bulletin double est recensée à raison dans un logement et à tort dans un autre, et on peut estimer les doubles comptes statistiques à partir des bulletins doubles la même année t , au prix de quelques hypothèses.

Le recensement regroupant les données de cinq vagues annuelles, la même personne peut figurer plusieurs fois dans le fichier si elle a déménagé. Il ne s'agit d'un double compte que si elle a été recensée une fois à tort et l'autre fois à raison. Nous utiliserons les paires d'années de collecte t et $t+1$ pour proposer une estimation améliorée des doubles comptes tenant compte du risque accru de double compte dû au fait que, le plus souvent, une personne ayant deux logements n'est recensée une année donnée que dans un seul de ses logements.

L'échantillon démographique permanent permet d'estimer les doubles comptes

L'identification des individus appartenant à l'Échantillon démographique permanent (EDP – cf. encadré 3-C.1) se fait à partir de chaque vague d'EAR. Avec le recensement rénové, il n'y a de suppression des bulletins doubles qu'au sein d'un même logement. La même personne peut apparaître plusieurs fois la même année (ou des années proches) au recensement dans deux logements différents, sans que l'élimination d'un éventuel bulletin « à tort » ne soit entreprise. On dispose ainsi, pour tous les individus EDP, des informations sur toutes les vagues d'enquête de recensement où ils sont présents. Nous avons calculé des scores d'appariement entre bulletins du recensement attribués au même individu EDP pour nous assurer que les bulletins doubles ne résultaient pas d'erreurs d'appariement. Le détail de ce calcul est fourni dans [Toulemon, Durier et Marteau, 2018] et nous supposons ici qu'il n'y a pas d'erreur de ce type à l'EDP. Pour les analyses fondées sur les données de deux années successives, t et $t+1$, nous appelons bulletins doubles la présence d'un bulletin de recensement une année t et d'un autre bulletin l'année $t+1$, dans lequel, à la question sur la résidence un an avant, « Où habitiez-vous le 1^{er} janvier [t]? », la personne répond « Dans le même logement que maintenant ». Toutes nos analyses sont fondées sur les données brutes du recensement, pour éviter que les imputations ou corrections ne créent des faux bulletins doubles. Bien que l'EDP contienne des variables socio-fiscales très riches (Insee, 2018), nous n'utilisons ici que l'information issue des vagues de recensement.

Méthodes d'estimations, à partir des données d'une année

On se place dans la situation où il n'y a pas d'omission. On suppose également que les habitants sont recensés au plus dans deux logements. En pratique, presque aucun individu n'est recensé trois fois ou plus la même année dans des logements différents. On cherche à estimer la proportion de doubles comptes statistiques une année donnée à partir des bulletins doubles observés cette même année. Dans les calculs, on tient compte du fait que les probabilités d'inclusion sont différentes (20 % pour les logements des petites communes et pour les logements neufs ou les grandes adresses des grandes communes, 8% pour les autres) ainsi que du fait que deux logements ne sont pas indépendants s'ils appartiennent à la même petite commune ou à la même adresse d'une grande commune.

La probabilité qu'une personne ayant deux logements habituels soit interrogée dans ces deux logements une année donnée varie donc de $0,2 \times 1 = 20\%$ pour les habitants ayant deux logements dans la même petite commune à $0,08 \times 0,08 = 0,64\%$ pour ceux vivant dans deux adresses différentes en grande commune.

Tous les calculs tiennent compte précisément du plan de sondage des EAR dans le calcul de la probabilité de présence simultanée de deux logements dans la même vague de recensement (Toulemon, 2018). Cependant, pour simplifier la présentation de la méthode, on suppose que tous les logements ont la même probabilité d'inclusion et que, pour deux logements différents, les probabilités de sélection sont indépendantes. En appelant p la probabilité de sélection d'un logement, on peut écrire $w = 1/p$ la pondération de l'individu dans son logement.

Quatre catégories d'habitants

On peut définir quatre catégories au sein de la population, dont on va estimer l'importance et le poids dans la population recensée, classées ici par risque croissant de double compte. Le recensement est fondé sur cinq vagues annuelles. On ne s'intéresse qu'aux personnes rentrant dans le champ du recensement (personnes

ayant leur résidence principale en France) : une personne ayant un logement répond au recensement ; une personne ayant plusieurs logements répond au moins une fois ; on suppose également que les habitants ont deux logements au maximum. On omet également la pondération EDP tenant compte du fait que l'échantillon EDP couvre une personne sur 100 jusqu'en 2008 et une personne sur 25 à partir de 2009 (Jugnot, 2014) ; nous calculons des proportions de doubles comptes dans le champ EDP et nous l'appliquons à l'ensemble de la population.

Catégorie A : individus a ayant un seul logement habituel. Pour les individus ayant un seul logement, on définit $Y(a,i)$ comme l'indicatrice de présence de l'individu a dans le fichier du recensement dans le logement i , et $Y(a)$ la somme des $Y(a,i)$ pondérée par le poids $w(i)$ de l'individu a dans chaque logement i . En pratique, le poids est calculé pour le logement i et affecté à tous les habitants du logement.

$$\text{Pour } a \text{ dans } i : Y(a,i) = 1$$

Soit J l'ensemble de tous les logements recensés une année donnée. En écrivant $Pop(A)$ la taille de la population appartenant à la catégorie A , et $Rec(A)$ la taille de la catégorie A au recensement :

$$Rec(A) = \sum_{i \in J} p \sum_{a \text{ dans } i} w(a) Y(a,i) = p w Pop(A) = Pop(A)$$

Logiquement, le recensement estime correctement l'effectif des habitants ayant un seul logement habituel. Les personnes ayant une résidence secondaire dans laquelle elles ne sont pas susceptibles de remplir un bulletin du recensement (le logement étant repéré sans ambiguïté comme une résidence secondaire au recensement) appartiennent à cette catégorie.

Pour les individus ayant deux logements habituels dans lesquels un agent recenseur est susceptible de leur laisser un bulletin individuel, on définit $Y(a,i)$ l'indicatrice de présence au recensement de a dans le logement i en t . On suppose que la personne habite dans deux logements i et j . On suppose que le logement i est leur résidence principale au sens du recensement, la personne est donc recensée à raison dans le logement i et à tort dans le logement j , mais on ne connaît pas la résidence principale dans le fichier. On peut envisager 4 cas pour les individus ayant deux logements i et j :

Cas 1 : Ni i ni j ne sont dans J : a n'est pas interrogé par le recensement.

Cas 2 : i est dans J , et pas j : a est interrogé dans i .

Cas 3 : j est dans J , et pas i : a est interrogé dans j .

Cas 4 : i et j sont dans J : a est interrogé deux fois, dans i et dans j .

Dans un recensement exhaustif, J représente l'ensemble des logements et seul le cas 4 existe ; avec un recensement par enquête, on peut écrire pour une année donnée, si chaque logement a une probabilité p d'inclusion et si les logements sont indépendants les uns des autres :

$$P(\text{cas 1}) = (1-p)^2 ; P(\text{cas 2}) = p(1-p) ; P(\text{cas 3}) = p(1-p) ; P(\text{cas 4}) = p^2$$

La probabilité pour une personne ayant deux logements de participer au recensement dans un seul logement (cas 2 ou 3) est donc de $2p(1-p)$, tandis que la probabilité de se voir proposer deux bulletins est de p^2 . Nous allons distinguer trois catégories parmi les habitants ayant deux logements habituels.

Catégorie B : individus ayant deux logements et capables d'identifier le logement où ils doivent être recensés. Seuls ceux qui doivent être recensés dans i remplissent un bulletin en t :

$$\text{Pour } a \text{ dans } i \text{ et } i \text{ résidence principale de } a : Y(a,i) = 1$$

Les habitants de la catégorie B sont également interrogés dans leur logement j , mais ils ne remplissent pas de bulletin, car j n'est pas leur résidence principale et ils ont lu, compris et appliqué les consignes (le logement j est pour eux une résidence habituelle, mais pas leur « résidence principale » au sens du recensement).

$$\text{Pour } a \text{ dans } j \text{ et } j \text{ non résidence principale de } a : Y(a, j) = 0$$

Pour les habitants de la catégorie B , $Y(a)$ vaut w dans les cas 2 et 4, 0 sinon

À partir de l'échantillon J , on estime au recensement la population de B , notée $Rec(B)$ par :

$$Rec(B) = \sum_{i \in J} p \sum_{a \text{ dans } i \text{ et } a \in B} wY(a, i) = (p(1-p) + p^2) w Pop(B) = Pop(B)$$

Les habitants de la catégorie B ont deux logements, mais une seule résidence principale. Du point de vue du recensement, ils ne produisent pas de double compte.

Catégorie C : individus ayant deux logements, ne lisant pas les consignes, mais sensibles au fait qu'ils ne doivent pas être recensés deux fois. Ces personnes ne lisent pas les consignes et ont deux résidences habituelles dans lesquelles elles sont prêtes à remplir un bulletin. Cependant si les deux logements sont présents la même année, elles ne remplissent qu'un seul bulletin (pour éviter le double compte, avec le raisonnement qu'elles n'ont pas à remplir deux bulletins). On suppose que le logement i est leur résidence principale au sens du recensement. Si i et j font partie de la même vague de recensement, on suppose qu'elles répondent dans le premier logement avec une probabilité 1 et dans le deuxième avec une probabilité 0 : elles évitent de participer deux fois au recensement la même année. Par contre, si un seul logement est inclus dans l'échantillon une année donnée, elles remplissent un bulletin dans le logement. En particulier, si un logement est inclus dans le recensement en t et un autre en $t+1$, elles remplissent deux bulletins, car elles ont oublié avoir déjà répondu au recensement ou ne perçoivent pas le double compte.

Dans le cas où les deux logements sont interrogés l'année donnée, seuls les habitants qui sont d'abord recensés dans i remplissent un bulletin en i ; les autres remplissent un bulletin en j ; finalement, tous ne remplissent qu'un seul bulletin :

$$\text{Pour } a \text{ recensé dans } i \text{ et } i \text{ avant } j : Y(a, i) = 1$$

$$\text{Pour } a \text{ recensé dans } j \text{ et } j \text{ avant } i : Y(a, j) = 1$$

Si un seul logement est interrogé l'année donnée, les habitants de la catégorie C remplissent un bulletin en i ou en j :

$$\text{Pour } a \text{ recensé dans } i : Y(a, i) = 1$$

$$\text{Pour } a \text{ recensé dans } j : Y(a, j) = 1$$

Pour les habitants de la catégorie C , $Y(a)$ vaut w dans les cas 2, 3 et 4, 0 sinon.

À partir de l'échantillon J , on a :

$$Rec(C) = \sum_{i \in J} p \sum_{a \text{ dans } i \text{ et } a \in C} wY(a, i) = (2p(1-p) + p^2) w Pop(C) = (2-p) Pop(C)$$

Les habitants de la catégorie C ne sont jamais en bulletin double la même année ; cependant, ils sont le plus souvent en double compte statistique, et leur poids total (effectif pondéré de la catégorie C) n'est pas $Pop(C)$ mais $(2-p) Pop(C)$. On attribue à chaque bulletin des habitants de la catégorie C un poids w , car on ignore que le même habitant est également susceptible d'être interrogé et de remplir un autre bulletin dans son autre logement.

Catégorie D : individus ayant deux logements, ne lisant pas les consignes et répondant au recensement dans tous leurs logements. Ces personnes ne lisent pas les consignes et ont deux résidences habituelles dans lesquelles elles sont prêtes à remplir un bulletin. Les habitants de la catégorie *D* remplissent un bulletin en *i* et un bulletin en *j*, même si ces deux logements sont recensés la même année :

Pour *a* recensé dans *i* : $Y(a, i) = 1$

Pour *a* recensé dans *j* : $Y(a, j) = 1$

Pour les habitants de la catégorie *D*, $Y(a)$ vaut *w* dans les cas 2 et 3, $2w$ dans le cas 4, 0 sinon.

À partir de l'échantillon *J*, on a :

$$Rec(D) = \sum_{i \in J} p \sum_{a \text{ dans } i \text{ et } a \in D} wY(a, i) = (2p(1-p) + 2p^2) wPop(D) = 2Pop(D)$$

Pour un recensement exhaustif, $p=1$, les catégories *B* et *C* sont identiques et ne donnent pas lieu à des doubles comptes ; les doubles comptes correspondent à la seule catégorie *D*. Dans un recensement par enquête annuelle, avec $p < 1$, les habitants de la catégorie *C* s'ajoutent à ceux de la catégorie *D*, et la proportion de doubles comptes augmente. Les habitants de la catégorie *D* sont en double compte statistique (leur poids total dans le fichier du recensement est $2 Pop(D)$ au lieu de $Pop(D)$) ; on peut parler pour les habitants de la catégorie *C* de doubles comptes partiels, leur poids total étant inférieur à $2 Pop(C)$, car quelques habitants de la catégorie *C* « évitent » le double compte : s'ils sont interrogés dans leurs deux logements, ils ne répondent qu'une seule fois.

Estimation finale des doubles comptes. Les habitants de la catégorie *D* sont en double compte dès que l'un de leurs logements est recensé, puisqu'ils répondent et sont inclus dans les fichiers dans chaque logement ; les habitants de la catégorie *C* sont en double compte uniquement si un seul de leurs logements est recensé une année donnée. Quel est le poids relatif de ces catégories dans les fichiers du recensement ? Une personne susceptible d'être recensée dans deux logements et effectivement recensée dans un logement une année donnée a respectivement les probabilités $2p(1-p)$ et p^2 d'être recensée une ou deux fois une année donnée. Dans le recensement fondé sur 5 EAR, les mêmes formules s'appliquent, à ceci près que, parmi les personnes dont les deux logements sont inclus dans l'échantillon du recensement, une personne sur 5 est recensée deux fois la même année et 4 sur 5 sont recensées deux années de suite ou à deux ans d'écart. Enfin, certaines personnes ne sont recensées que dans un logement sur les cinq ans et d'autres ne sont pas recensées du tout. Avec $p^2 = 1/36$, probabilité observée pour les doubles comptes en $t; t^{20}$, on obtient les ordres de grandeur présentés au tableau 3-A.1. Les probabilités exactes, calculées pour chaque strate de type de commune et de logement, conduisent à la même estimation qu'en appliquant une probabilité moyenne de sélection.

20 Les doubles comptes sont plus fréquents dans les grandes communes et, dans les grandes communes, dans les grandes adresses et les logements neufs, ce qui aboutit à un poids moyen observé pour les personnes en bulletin double plus faible, de l'ordre de 36 (soit un taux de sondage moyen de 1/6) au lieu de 50 (pour un taux de sondage de 14 %, environ 1/7).

Tableau 3-A.1 : Probabilités de sélection de deux logements pour une EAR ou au recensement (cumul de cinq EAR)

	Nombre d'EAR		Avec $p=1/6$	
	Une EAR	Cinq EAR	Une EAR	Cinq EAR
Deux logements recensés...	p^2	$25p^2$	1/36	25/36
- la même année	p^2	$5p^2$	1/36	5/36
- deux années de suite		$10p^2$		10/36
- à deux ans d'intervalle		$10p^2$		10/36
Un seul logement recensé	$2p(1-p)$	$10p(1-5p)$	10/36	10/36
Aucun logement recensé	$(1-p)^2$	$(1-5p)^2$	25/36	1/36
Compte : habitant A ou B	1	1	1	1
Compte : habitant C	$2-p$	$2-p$	11/6	11/6
Compte : habitant D	2	2	2	2

Note : ces valeurs sont estimées sous l'hypothèse d'indépendance d'inclusion de chaque logement. Elles fournissent des ordres de grandeur. En pratique, nous avons calculé les probabilités pour les strates de type de commune et de logement (Toulemon, Durier et Marteau, 2018). Les comptes (poids moyen pondéré) valent 1 pour les personnes comptées une fois et 2 pour les doubles comptes. Les comptes pour les 5 EAR sont divisés par 5. On voit que les habitants de la catégorie C sont en double compte partiel ($11/6=1,83$), car une fois sur six ils sont recensés deux fois la même année et ne remplissent qu'un seul bulletin.

Estimation des effectifs des différentes catégories

Bulletins doubles en t et catégorie D : doubles comptes systématiques. À partir des bulletins doubles observés l'année t , on estime directement l'effectif de la catégorie D, qui donne lieu à un bulletin double $BD(t)$ dans le cas 4, à partir de la probabilité du cas 4. En notant $1(E)$ la variable indicatrice de l'événement E (qui vaut 1 si E est vrai, 0 sinon) et EDP l'ensemble des individus EDP, on a :

$$Pop(D) = \sum_{a \in EDP} \frac{1}{p^2} 1(Y(a)=2w) = w^2 BD(t)$$

Les individus de D ayant une probabilité double d'apparition, leur poids total dans les fichiers du recensement est le double de leur effectif :

$$Rec(D) = 2w^2 BD(t)$$

Bulletins doubles en $t; t+1$ et catégorie C. Sous l'hypothèse que la somme des $Y(a)$ en t et $t+1$ vaut $2w$ quand un logement est inclus dans le recensement en t et l'autre en $t+1$ pour les personnes catégories C et D qui n'ont pas déménagé entre t et $t+1$, on peut estimer la proportion de bulletins doubles $BD(t; t+1)$ en t , à partir des bulletins doubles en t et $t+1$: parmi les habitants des catégories C ou D, certains peuvent être recensés en i en t et en j en $t+1$. Ils seront observés comme des doubles comptes en t sous deux conditions supplémentaires :

- les logements i et j appartiennent à deux vagues annuelles successives. Cela restreint le champ aux doubles comptes des habitants dont les deux logements ne sont ni dans la même petite commune ni à la même adresse dans une grande commune. Pour ceux dont les deux logements sont dans la même petite commune ou à la même adresse, seuls les habitants de la catégorie D sont en double compte, puisque leurs deux logements sont nécessairement recensés la même année ;

- à l'enquête de $t+1$, l'habitant recensé en j doit y résider depuis plus d'un an pour avoir été en double compte en t , quand il a été recensé dans le logement i . Il faut éviter de compter à tort comme doubles comptes les personnes résidant en i en t et ayant déménagé pour j en $t+1$, qui ne sont pas en double compte. On ne conserve donc que les personnes recensées en t dans le logement i et qui, à la question sur

la résidence un an avant, posée en $t+1$ dans le logement j , « Où habitiez-vous le 1^{er} janvier [t] ? » ont répondu « Dans le même logement que maintenant ». Ils habitaient donc en t dans les deux logements i et j .

La première condition limite le nombre de doubles comptes au recensement ; la seconde limite leur possibilité d'observation : une personne en double compte a deux fois moins de chances d'être observée comme bulletin double en $t;t+1$ si elle a quitté un de ses logements que si elle n'a pas déménagé : il faut qu'entre t et $t+1$, elle ait déménagé de i (où elle a été recensée en t) mais pas de j où elle est recensée en $t+1$ (et répond en $t+1$ qu'elle résidait déjà dans j en t) ; si elle a quitté ses deux logements, elle ne peut pas être observée comme bulletin double en $t;t+1$. On peut estimer la proportion s , parmi les habitants en double compte en t , de ceux qui vivent dans un ou dans les deux logements depuis plus d'un an, et qui sont susceptibles d'être observés à partir des vagues d'enquête t et $t-1$. Sous l'hypothèse de stationnarité, on utilise cette proportion pour les années $t;t+1$ pour estimer finalement la proportion de doubles comptes²¹. On définit alors $Y(a,j,t+1)$ l'indicatrice de résidence de a dans j en t , repérée dans la vague d'enquête en $t+1$. On a :

$$Pop(C+D) = \sum_{a \in EDP} \frac{1}{2p^2} \frac{1}{s} 1 \left(\sum_{i \in J(t), j \in J(t+1), i \neq j} Y(a,i) + Y(a,j,t+1) = 2 \right) = \frac{w^2}{2s} BD(t;t+1)$$

Bulletins doubles sans double compte. Par construction, si une personne figure deux fois dans le fichier de la même vague d'enquête en t , elle est en double compte. À l'inverse, certains habitants peuvent figurer deux fois dans le fichier, une fois en t et une fois en $t+1$, sans être en double compte. Tout d'abord, environ 14 % des personnes recensées en t et ayant déménagé entre t et $t+1$ ont emménagé dans un logement présent dans la vague annuelle de $t+1$: il est très important de les distinguer des personnes en double compte qui vivaient déjà en t dans le logement où elles sont recensées en $t+1$. Nous utilisons l'information brute (non imputée) sur la résidence un an auparavant et nous considérons tous les cas douteux comme des déménagements. Les changements de géographie de la collecte du recensement (en lien notamment avec les fusions de communes) peuvent également rendre quelqu'un éligible dans le même logement deux années de suite. Nous avons éliminé ces cas après avoir vérifié qu'ils étaient très peu nombreux. Enfin, dans certains cas de multi-résidence, certaines personnes pourraient avoir conservé leurs logements, mais changé de résidence principale au recensement entre t et $t+1$. Nous avons identifié trois possibilités :

- Enfants en résidence partagée 50-50 entre les logements M et P de leurs mère et père séparés, qui auraient passé la nuit de référence du recensement en M l'année t et en P l'année $t+1$, M et P faisant partie des échantillons des enquêtes annuelles t et $t+1$ respectivement.

- Jeunes adultes recensés en t dans leur logement d'étude et en $t+1$ chez leurs parents dans leur logement familial (où ils revenaient les week-ends en t) ; en $t+1$, ils peuvent déclarer qu'ils habitaient dans le logement familial un an avant, même s'ils ont été recensés ailleurs.

- D'une manière générale, tous les adultes qui ont deux logements et passent plus de la moitié de l'année dans un logement l'année t et plus de la moitié dans l'autre l'année $t+1$ peuvent figurer à raison dans deux vagues annuelles de suite. La question sur la résidence un an auparavant ne permet pas d'identifier les personnes qui vivaient dans le même logement mais dont la résidence principale était ailleurs : on suppose que les personnes recensées en $t+1$ dans un logement et répondant qu'elles y habitaient déjà en t auraient pu être recensées dans ce logement si le recensement y avait eu lieu en t .

²¹ L'hypothèse de stationnarité suppose que les doubles comptes sont constants, en nombre et en structure. Elle permet d'identifier les transitions entre t et $t-1$, que l'on observe au recensement à partir de la question sur la résidence un an avant, aux transitions entre t et $t+1$ qui servent à estimer la proportion s de doubles comptes observables. Dans l'analyse fondée sur les EAR de t et $t+1$, l'échantillon est multiplié par deux et le nombre de bulletins doubles est également deux fois plus élevé, puisque les combinaisons (i en t et j en $t+1$) et (j en t et i en $t+1$) sont différentes, tandis que la combinaison (i en t et j en t) est unique. Le produit des poids est donc divisé par $2s$, et non par s .

Dans tous ces cas, un bulletin double identifié en $t;t+1$ conduit à « estimer » à tort $w^2/2$ doubles comptes de catégorie C, puisque le bulletin double est observé à partir de deux vagues successives. Seul le premier cas correspond à des informations cohérentes avec les consignes du recensement, mais on peut le considérer comme assez improbable, d'autant que la date de référence du recensement ne figure ni sur les bulletins individuels ni sur la feuille de logement²². On estime à 400 000 au maximum le nombre d'enfants en garde partagée 50-50 en 2016, soit 3 % des enfants mineurs (Algava *et al*, 2019 ; Chardon et Vivas (2004) l'estimaient à moins de 100 000, sur une période plus ancienne, tandis que Toulemon et Denoyelle (2012) estiment la population d'enfants au risque de double compte entre 250 000 et 500 000 pour la décennie 2000). Une année donnée, on compte, d'après l'EDP, 200 000 enfants en doubles comptes. Selon que ces enfants sont des enfants en résidence alternée 50-50 ou non, restent entre 200 000 et 400 000 enfants « au risque d'être en bulletin double » en $t;t+1$ sans être en double compte si les consignes sont lues et respectées. Comme l'alternance est *a priori* indépendante de la date de référence du recensement, ces enfants ont chaque année une chance sur deux de passer la nuit dans le logement recensé, soit une chance sur quatre d'être comptés deux fois à raison. Cela représente donc entre 50 000 et 100 000 enfants, soit 0,4 % à 0,7 % des enfants mineurs, et 0,16 % de la population au maximum, sous l'hypothèse que les parents dont les enfants sont en garde partagée lisent les consignes et les appliquent, connaissent la date de la nuit de référence du recensement. Il est possible que ce soit le cas pour certains d'entre eux, sachant que les enfants en résidence partagée ont relativement plus de chances d'appartenir à la catégorie D que C, car ils ne remplissent pas eux-mêmes leur bulletin : si deux bulletins sont remplis pour eux, ils ne le sont pas par la même personne. Or le rapport des catégories C et D varie peu avec l'âge (Toulemon, Durier et Marteau, 2018). La correction à apporter serait de l'ordre de 0,1 % sur les doubles comptes $t;t+1$. Nous ne l'avons pas effectuée.

Estimation de la variance des estimations. Compte tenu de la grande complexité du plan de sondage du recensement et du fait que les probabilités de bulletin double varient comme le produit des probabilités de sélection de chaque logement, l'écart-type de chaque estimation est difficile à estimer. Par ailleurs, les effets de grappe sont limités par le fait que les doubles comptes sont rares et dispersés sur le territoire (à l'exception des fratries en doubles comptes dans les mêmes deux logements). Nous avons donc adopté la méthode d'estimation suivante : à partir d'un nombre n d'individus dans l'EDP représentant un groupe (ici les individus présents deux fois au recensement) correspondant à une proportion u de la population, on estime l'effectif N dans la population P et $V(N)$ et $V(u)$ la variance de u à partir des formules suivantes, avec w le poids des individus du groupe (produit du poids EDP et du produit des poids de sondage de l'EAR (inverse de la probabilité d'être sélectionné deux fois, voir tableau 3-A.1 et (Toulemon, 2018) pour le détail des probabilités selon la combinaison des deux logements entre petites et grandes communes) et $\bar{w}$ le poids moyen des individus du groupe :

$N = \bar{w} n$ estimation du nombre d'individus en double compte

$u = N / P$ estimation de la proportion d'individus en double compte

$V(n) = n$ estimation poissonnienne de la variance en première approximation

$V(u) = u^2 n z$ estimation de la variance de la proportion d'individus en double compte tenant compte de la variance des poids (poids EAR * poids EDP pour les bulletins doubles).

On estime le facteur z d'augmentation de la variance dû à la variance des poids des membres du groupe comme $[1 + (1/\bar{w}^2) V(w)]$. Cette formule est obtenue en estimant la variance de l'effectif pondéré comme la variance du produit des deux variables X (être en bulletin double) et w (pondération), sous les hypothèses d'absence de covariance entre X et w et de faiblesse de X devant w : $V(uX) = V(X) V(u) + V(X) + V(u) [E(X)]^2 \approx V(X) [1 + V(u)]$ avec u poids ramené à la moyenne 1 ($u = w/\bar{w}$). Sur un groupe d'effectif similaire (les

22 La date de référence a été ajoutée sur la feuille de logement en 2018, mais nos analyses portent sur des années antérieures.

couples de même sexe repérés dans l'enquête Famille et logements) et également réparti dans l'ensemble du territoire, nous avons montré (Toulemon et al., 2014) que cette estimation reproduit très bien l'ordre de grandeur de la variance, par rapport à une estimation tenant compte de l'ensemble du plan de sondage (strates par région, tirage équilibré dans les régions, grappes de logements).

Entre 1,5 et 3 % de doubles comptes au recensement

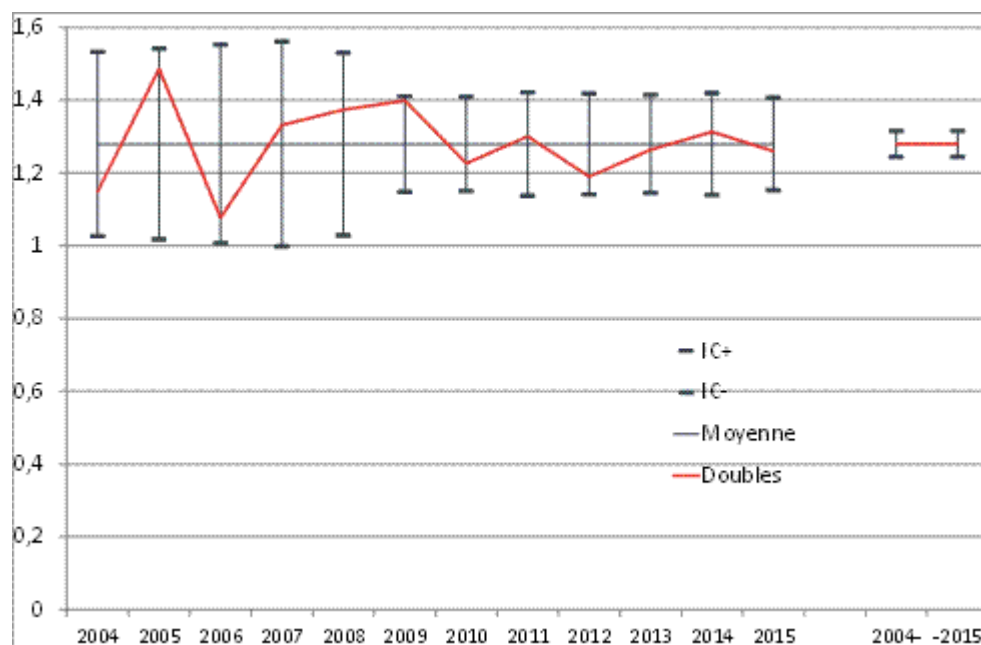
1,3 % de doubles comptes estimés à partir des bulletins doubles la même année

On compte dans l'EDP environ 210 bulletins doubles chaque année jusqu'en 2008, et 910 de 2009 à 2015, après l'augmentation du taux de sondage de l'EDP de 1 % à 4 % (de 4 à 16 jours de naissance par an). Ces effectifs sont tout à fait suffisants pour estimer les doubles comptes de la catégorie *D*, ceux qui peuvent donner lieu à un bulletin double la même année. Le poids moyen dans la vague annuelle de recensement (poids EAR) des personnes recensées deux fois la même année est de 6 en moyenne dans chacun de leurs logements : dans les fichiers du recensement, ils comptent dans les grandes communes pour 1,5 % et dans les petites pour 1,1 %, souvent dans des logements neufs ou des grandes adresses (Toulemon, Durier et Marteau, 2018)²³. Pour les années 2004 à 2015, la proportion de doubles comptes s'établit à 1,3 %. Cette proportion est stable d'une année à l'autre (Graphique 3-A.1).

On peut estimer l'intervalle de confiance en tenant compte des différentes catégories de commune et de la variance des poids : sur l'ensemble des années observées, l'intervalle de confiance est de [1,23 % – 1,33 %], autour d'une valeur centrale de 1,28 %, soit 800 000 personnes en double compte au recensement (entre 770 000 et 830 000), pour une population métropolitaine moyenne de 62,6 millions en 2009. Cet intervalle de confiance est indicatif, car le recensement est une opération complexe, avec un plan de sondage du recensement qui combine découpage en strates, équilibrage des EAR et effets de grappe, et auquel s'ajoute le tirage de l'EDP. Cependant, cette incertitude est très inférieure à la différence avec l'estimation issue des doubles comptes repérés à partir de deux vagues successives (la catégorie *C* s'ajoutant alors à la catégorie *D*).

23 Leur pondération est égale au produit des poids (pondérations de l'EAR) dans chacun des logements où elles habitent, multiplié par leur poids EDP. Leur part est calculée comme le rapport de leur effectif estimé à celui de l'estimation de la population totale d'après la vague EAR de l'année considérée (poids de l'EAR x poids EDP de l'ensemble des personnes recensées).

Graphique 3-A.1 : Proportion d'individus en double compte par année de recensement



Lecture : en 2004, 1,15 % des individus recensés sont en double compte. L'intervalle de confiance est de [1,01 %-1,53 %] autour de la valeur moyenne de 1,28 %, au seuil de 5 % bilatéral.

Champ : bulletins individuels (BI) des individus nés un jour EDP, à l'exclusion des BI avec un poids nul. EAR 2004-2015.

Source : Base étude 2015 de l'EDP.

2,4 % de doubles comptes estimés à partir des bulletins doubles deux années de suite

Pour les années de recensement qui s'étendent de 2011 à 2015, la proportion de doubles comptes statistiques mesurés la même année atteint 1,27 % (graphique 3-A.1). Quelle est la proportion de personnes en double compte en t (catégorie D) susceptibles d'être repérées à partir des EAR t et $t+1$? Il faut pour cela que leurs deux logements ne soient pas recensés la même année et qu'elles n'aient pas emménagé depuis moins d'un an dans le logement où elles sont recensées en $t+1$. Le tableau 3-A.2 présente le détail des estimations. Les personnes ayant leurs deux logements dans la même petite commune ou à la même adresse (0,06 %, tableau 3-A.2, ligne b1) ne peuvent être en bulletin double en t et $t+1$. Les autres doubles comptes en t (1,21 %, tableau 3-A.2, lignes b2 et b3) sont susceptibles d'être en bulletin double en t et $t+1$: recensées en t dans un logement et en $t+1$ dans un autre logement où elles déclarent qu'elles y résidaient déjà un an auparavant (donc en t). En supprimant parmi les doubles comptes en t les personnes résidant dans leurs logements depuis moins d'un an et la moitié de celles résidant dans un de leurs logements depuis moins d'un an (et depuis plus d'un an dans l'autre), la proportion descend à 0,76 % (tableau 3-A.2, ligne b2), ce qui est conforme à ce que l'on sait des épisodes de multi-résidences dans les enquêtes longitudinales : d'après l'enquête SRCV, la part des multi-résidences durant plus d'une année est de l'ordre de 50 % à 65 % (Toulemon et Pennec, 2009). Pour le même champ, l'estimation directe des doubles comptes statistiques à partir de deux vagues successives est de 1,60 % (tableau 3-A.2, ligne c2). On suppose ensuite que le rapport $c2/b2$, représentant le rapport entre le nombre de personnes de catégorie $C+D$ et le nombre de celles de catégorie D , s'applique aussi aux personnes ayant déménagé entre t et $t+1$: $c3 = b3 \ c2/b2 = 0,94$ %. Finalement, on estime à 2,54 % la proportion de doubles comptes en t repérés à partir des bulletins doubles en t ; $t+1$ (tableau 3-A.2, ligne c).

Chaque logement des habitants en double compte potentiel ayant une probabilité moyenne observée de 1/6 d'être sélectionné chaque année, au bout de 5 ans, quatre cas sont possibles (tableau 3-A.1) : deux logements interrogés la même année ($p=5/36$) ; deux logements interrogés deux années différentes

($p=20/36$) ; un seul logement interrogé ($p=10/36$) ; aucun logement recensé ($p=1/36$). Quand l'individu est présent au recensement, les deux logements sont interrogés la même année avec une probabilité de $5/35=1/7$, et un seul logement est interrogé, ou les deux logements sont interrogés deux années différentes avec une probabilité de $30/35=6/7$. Finalement, les personnes ayant deux logements habituels et recensées sont donc interrogées deux fois la même année une fois sur sept (dans ce cas, seule la catégorie D est en double compte, $p=1,27\%$) et interrogées une seule fois ou deux fois lors de deux EAR différentes (dans ce cas, les catégories C et D sont en double compte, à condition que les deux logements ne soient pas dans la même petite commune ou à la même adresse, $p=2,54\%$). L'estimation finale de la proportion de doubles comptes statistiques sur l'ensemble d'un recensement s'écrit $e = [d(t;t)+6d(t;t+1)] / 7$, soit $2,36\%$. Ramené à l'ensemble de la population (et non plus l'échantillon EDP), cela conduit à une estimation finale de 1 501 000 personnes en double compte, pour une population métropolitaine de 63 650 000 au 1^{er} janvier 2013. Les variations aléatoires autour de cette estimation sont du même ordre que précédemment (on observe 4 643 bulletins doubles en $t;t+1$ de 2011 à 2015), ce qui conduit à un intervalle de confiance final autour de l'estimation de $2,36\%$ de l'ordre de $[2,27 ; 2,45]$, en tenant compte uniquement des variations liées au sondage du recensement et de l'EDP.

Les doubles comptes sont à peu près deux fois plus probables à partir des bulletins doubles dans deux vagues successives, ce qui confirme notre hypothèse selon laquelle les doubles comptes sont plus souvent visibles et donc évités quand les deux logements sont interrogés la même année (tableau 3-A-2).

On retrouve le même résultat en estimant la proportion d'habitants de la catégorie D à $1,27\%$, puis ceux de la catégorie C à $1,39\%$, dont $1,10\%$ en double compte. La proportion s , parmi les habitants de la catégorie D, de ceux qui n'ont pas déménagé et ont leurs deux logements dans des adresses ou des petites communes différentes (lignes b2/b, $0,76/1,27$) est de 60% . À la ligne c2, on observe $1,60\%$ des habitants, dont $0,76\%$ de la catégorie D et donc $0,84\%$ de la catégorie C. En supposant que la proportion s est identique pour les habitants de la catégorie C, on estime leur effectif à $0,84/60\% = 1,39\%$. Ces habitants sont en double compte partiel (voir tableau 3-A.1) : une chance sur six d'être compté une fois, 5 chances sur 6 d'être compté deux fois ; poids moyen $11/6$). On compte donc $1,39 \times 5/6 = 1,16\%$ doubles comptes théoriques, auxquels on enlève ceux dont les deux logements sont dans la même petite commune ou à la même adresse ; finalement, on compte $1,10\%$ d'individus en double compte de la catégorie C, et $2,36\%$ de doubles comptes au total.

Tableau 3-A.2 : Comparaison en $t;t$ et $t;t+1$ des individus déclarant ne pas avoir déménagé (2011-2015)

	Effectifs pondérés		Proportions (en %)	
	$t;t$	$t;t+1$	$t;t$	$t;t+1$
a. Individus EDP recensés (5 vagues, pondération EAR)	13 804 821	13 804 821	100,00	100,00
b. Individus recensés deux fois en t (catégorie D), dont...	174 896		1,27	
b1. Individus à la même adresse ou dans la même petite commune	7 615		0,06	
b2. Individus déclarant ne pas avoir déménagé (hors même petite commune ou même adresse)	105 462		0,76	
b3. Individus ayant déménagé entre $t-1$ et t (hors même petite commune ou même adresse)	61 819		0,45	
c2. Individus (catégories C+D) déclarant ne pas avoir déménagé (hors même petite commune ou même adresse)		221 033		1,60
c3. Individus (catégories C+D) ayant déménagé entre t et $t+1$ (hors même petite commune ou même adresse)				0,94
c. Estimation d'individus en doubles comptes en $t;t+1$				2,54
d1. Estimation des doubles comptes: catégorie D en $t;t$ et en $t;t+1$			1,27	1,21
d2. Estimation des doubles comptes: catégorie C en $t;t$ et en $t;t+1$			0,00	1,33
d. Estimation des doubles comptes en $t;t$ et en $t;t+1$			1,27	2,54
e. Estimation finale (chaque logement a une probabilité 1/6 d'être inclus en t)			2,36	
e. Estimation finale de doubles comptes au recensement, pour une population totale de 63 650 000 habitants début 2013	1 501 000		2,36	

* Pour $t;t$, sont également inclus les individus déclarant avoir déménagé dans un des deux logements recensés (on divise alors par deux leur poids pour avoir des champs comparables). Pour faciliter la cohérence des totaux, on estime $t;t$ avec t de 2011 à 2015, et $t;t+1$ avec $t+1$ de 2011 à 2015.

Champ : $t;t$: bulletins doubles des individus nés un jour EDP. $t;t+1$: bulletins individuels des individus recensés lors de deux vagues successives. EAR 2011-2015.

Source : base étude 2015 de l'EDP.

Les scores d'appariement montrent que la proximité entre bulletins d'une même personne est plus forte quand les deux bulletins ne sont pas recueillis la même année, mais deux années de suite : dans ce cas, il est plus probable que les deux bulletins ont été remplis par la personne elle-même, tandis que, quand deux bulletins sont remplis la même année, cela peut être par deux personnes différentes, la personne elle-même dans un logement et un parent dans l'autre, par exemple (Toulemon, Durier et Marteau, 2018).

Doubles comptes et estimation de la population

Une estimation finale de 2,4 % de doubles comptes, assez fragile

Ces estimations des doubles comptes sont assez fragiles. Les effectifs sont suffisants pour que les variations aléatoires soient très faibles ; par contre, les sources d'erreur sont nombreuses. Des erreurs d'appariement à l'EDP ou des réponses erronées à la question sur la résidence un an avant créent un bulletin double une fois sur sept (si le nouveau logement en $t+1$ est recensé) et on multiplie ensuite son poids par sept, donc chaque erreur d'appariement ou réponse erronée conduit à un faux double compte. Les

erreurs d'appariement à l'EDP pourraient construire des « faux bulletins doubles », mais nos investigations nous conduisent à penser que les procédures d'appariement sont très rigoureuses et les scores d'appariement sont très élevés pour les différents bulletins du même habitant. De même, nous avons utilisé la variable brute du recensement pour limiter au maximum les erreurs de codage au recensement (Toulemon, Durier et Marteau, 2018).

Les réponses erronées peuvent être plus fréquentes, mais sont difficiles à identifier. D'après les équipes en charge du recensement, dans près d'un cinquième des cas où l'année d'emménagement dans le logement correspond à l'année précédente ou l'année en cours (t), les personnes déclarent être présentes dans le logement depuis plus d'un an à la question « Où habitiez-vous le 1^{er} janvier [$t-1$] ? ». Il est possible que, malgré les instructions, certaines personnes répondent à cette question en prenant comme date de référence le 1^{er} janvier de l'année en cours et non le 1^{er} janvier de l'année précédente. C'est probablement la plus grande source de fragilité de l'estimation en $t; t+1$.

Les hypothèses de comportement sont également assez fortes. Pour une année donnée, une personne ayant deux logements et présente au recensement ne verra ses deux logements recensés la même année qu'une fois sur onze : presque tous les doubles comptes sont invisibles pour les habitants, car ils ne donnent pas lieu à des bulletins doubles ; en ce sens, on peut parler de doubles comptes statistiques : les personnes qui répondent dans leurs deux logements ont deux fois plus de chances que les autres d'être recensées, mais sont très rarement confrontées à deux bulletins dans leurs deux logements. Quand on considère le recensement cumulant cinq EAR, la probabilité que les deux logements soient inclus la même année est un peu plus élevée : elle est de $5/35 = 1/7$; la probabilité d'être recensé deux années de suite est de $10/35 = 2/7$, et la probabilité qu'il y ait plus d'un an entre les deux inclusions ou que seul un logement soit inclus dans l'échantillon du recensement est de $20/35 = 4/7$ (tableau 3-A.2). Nous avons fait l'hypothèse que les personnes de la catégorie C remplissaient un bulletin en t et un autre en $t+1$ sans se soucier du double compte. Sous l'hypothèse que certains doubles comptes seraient évités en $t; t+1$ mais pas dans les cas où les deux logements sont recensés à un écart temporel plus grand qu'un an ou après un déménagement, notre estimation devrait en être augmentée. En l'absence d'une enquête post-censitaire, tous les doubles comptes sans bulletin double restent invisibles.

Les données de l'EDP permettent donc de proposer une première estimation des doubles comptes au recensement depuis le passage à un recensement par enquête : 2,4 %. Les intervalles de confiance dus aux aléas de sondage sont serrés [2,2 % ; 2,5 %] et difficiles à estimer précisément, les aléas de sondage de l'EDP et du recensement se combinant. On peut cependant les estimer grossièrement et de nombreuses autres sources d'incertitude existent ; l'incertitude totale sur l'estimation est beaucoup plus large. Cependant, la proportion de doubles comptes statistiques est certainement supérieure à 1,5 % et inférieure à 3 %. D'une part, la proportion déduite des bulletins doubles en $t; t$ (1,3 %) est certainement inférieure à la vraie proportion, puisque la grande majorité des doubles comptes ne donnent pas lieu à deux bulletins remplis la même année ; d'autre part, il est probable que peu d'habitants éviteraient un double compte si leurs deux logements sont recensés deux années de suite, mais pas si la durée entre les deux interrogations est supérieure. Nous proposons donc l'estimation de 2,4 % avec un intervalle de confiance de [1,5 % ; 3,0 %], soit entre 950 000 et 1 900 000 personnes.

Autres possibilités offertes par l'EDP

Notre estimation est plus élevée que la proportion de 0,8 % de doubles comptes estimée au recensement de 1990 (Coeffic, 1993), et comparable à ce que l'on obtient dans les recensements étrangers avant correction (Census Bureau, 2012 ; Toulemon, Durier et Marteau, 2018).

L'EDP permet, au-delà du comptage des doubles comptes, un très grand nombre d'analyses que nous avons esquissées ailleurs grâce aux nombreuses informations sur les individus EDP et les autres membres de leur ménage recensés avec eux. Ce sont surtout les jeunes adultes qui sont comptés deux fois, leur profil étant très proche de celui des personnes ayant deux résidences habituelles (Toulemon et Pennec, 2010 ;

Imbert *et al.*, 2014). De plus, on dispose pour ces habitants présents deux fois la même année des informations sur leurs deux logements et leurs deux ménages. La correction des doubles comptes conduirait à diminuer l'effectif total de la population recensée, mais dans une mesure partielle puisque, si l'on supprimait les personnes recensées seules à tort dans un logement, il faudrait également supprimer le logement de la liste des résidences principales (environ 0,8 % de ménages en moins). L'estimation totale de la population étant fondée sur la liste des résidences principales et le nombre d'habitants par logement, elle ne serait modifiée que de 1,5 %. L'écart est plus sensible sur des indicateurs comme la situation familiale des enfants : la proportion d'enfants vivant avec leurs deux parents (Buisson et Lapinte, 2017) passe de 75 % à 78 % si l'on corrige du fait que les enfants comptés deux fois partagent le plus souvent leur temps entre les logements de leurs deux parents séparés. Les définitions des fratries ou des situations familiales méritent d'être analysées pour ces enfants, dans la mesure où les gardes partagées deviennent plus nombreuses.

Conclusion

Compte tenu des nombreuses sources d'incertitude dans les méthodes d'estimation, nous proposons une fourchette large de 1,5 % à 3 % autour de notre estimation centrale de 2,4 % des habitants comptés deux fois au recensement, proportion stable entre 2004 et 2015. Pour certains groupes spécifiques ayant plus que les autres deux résidences habituelles (enfants de parents séparés, jeunes adultes revenant de temps en temps vivre chez leurs parents, personnes recensées seules dans leur logement), cette proportion est plus élevée et la description des situations familiales peut en être affectée (Toulemon, Durier et Marteau, 2018). Les situations de multi-résidence sont très diverses (Toulemon, Pennec, 2010) et il en va de même pour les situations familiales des personnes recensées deux fois (Toulemon, Durier et Marteau, 2018). La modification de la feuille de logement du recensement en 2018 a conduit à une diminution de la population estimée et à un ajustement des mouvements de la population (Papon et Beaumel, 2019 ; Insee, 2020). Il est probable que ce soient surtout des doubles comptes d'enfants qui ont été évités (Toulemon, 2018b), mais cela reste à confirmer. L'amélioration de la feuille de logement a été concentrée sur les enfants, sans question explicite sur l'ensemble des situations de multi-résidence ; l'ajustement prévu, de l'ordre de 500 000 personnes au total, pourrait correspondre à un tiers des doubles comptes.

La comparaison avec les données administratives, qui posent elles-mêmes des difficultés de comptage, est compliquée. L'EDP fournit des informations très riches sur les doubles comptes, qui pourront continuer à permettre des analyses précises ; l'utilisation de sources de données complémentaires au sein de l'EDP, telles que les données socio-fiscales, pourrait faire avancer les réflexions sur les caractéristiques socio-démographiques des individus en double compte ou omis dans le recensement, et est une piste prometteuse à l'avenir.

B) Omissions possibles de jeunes enfants au recensement : que nous apprennent les comparaisons avec l'état civil ?

Les comparaisons entre les effectifs des jeunes enfants et ceux des naissances enregistrées en France à l'état civil suggèrent de possibles omissions dans le recensement, plus particulièrement pour les nouveau-nés, mais également, de manière plus générale, pour les enfants en bas âge.

Ces omissions ne sont pas spécifiques à la méthode actuelle de recensement annuel, fondée sur la collecte d'Enquêtes annuelles de recensement depuis 2004, qui, cumulées sur cinq années successives, donnent les résultats du recensement. Elles s'observent aussi dans les recensements généraux de la population, qui étaient fondés sur une collecte exhaustive environ tous les dix ans.

Le déficit de jeunes enfants au recensement n'est pas spécifique non plus à la France. Il s'observe aussi dans les recensements étrangers, notamment en Grande-Bretagne et aux États-Unis. Ces cas sont connus et documentés, à l'image d'un commentaire rédigé à partir du cas de la France par Laurent Toulemon, chercheur à l'Institut national d'études démographiques (Ined), dans une revue étrangère ayant consacré un numéro au sous-dénombrement des jeunes enfants dans les recensements (Toulemon, 2017).

L'obligation d'attribuer un numéro d'identification au répertoire (NIR) à toutes les personnes nées sur le territoire français et l'obligation de déclarer les nouveau-nés dans des délais brefs après la naissance, ainsi que les contrôles mis en œuvre pour ne pas avoir de doubles enregistrements, font que les statistiques d'état civil sont considérées comme une référence sur le dénombrement des naissances.

On compare donc les effectifs de naissances à l'état civil avec les effectifs des jeunes enfants des générations correspondantes dans le recensement de la population.

Les concepts recueillis dans les deux sources n'étant pas strictement les mêmes, il convient tout d'abord de les préciser et de définir un champ de comparaison le plus proche possible (partie 3-B-1), avant d'exposer une estimation des écarts (partie 3-B-2), quelques caractéristiques des écarts (partie 3-B-3) et l'impact sur la pyramide des âges (partie 3-B-4).

B-1. État civil et recensement : des concepts et champs de couverture différents

Le recensement fournit des statistiques sur les naissances d'enfants dont la mère réside habituellement en France

Le recensement apporte des informations statistiques sur les personnes qui résident habituellement en France. L'analyse est menée ici à partir des enquêtes annuelles de recensement de la population de 2004 à 2018.

Champ géographique

Chaque année, ces enquêtes interrogent un échantillon de personnes en France métropolitaine et dans 4 DOM²⁴ pour fournir des données statistiques sur les personnes résidant habituellement en France. On dispose ainsi d'informations sur les enfants nés au plus tard l'année précédant la collecte et résidant habituellement en France, c'est-à-dire dont la mère est recensée. Par exemple, pour la collecte de l'EAR 2011 réalisée début 2011, les enfants les plus jeunes observés²⁵ sont nés en 2010 et ils ont été recensés, parce que leur mère réside habituellement en France.

24 Le recensement à Mayotte sur cette période est quant à lui exhaustif et quinquennal.

25 Les enfants nés au début de l'année de l'EAR (2011 dans le cas présent) sont également recensés, mais sont écartés pour les exploitations statistiques dans la mesure où la date de référence est le 1^{er} janvier de l'année.

L'état civil fournit des statistiques exhaustives sur les enfants nés en France

Les communes transmettent en continu à l'Insee les informations relatives à l'état civil sur leur territoire. En particulier, elles lui adressent de façon exhaustive les bulletins de naissance d'enfants nés dans la commune. Le bulletin de naissance comporte également des informations relatives à la commune de domicile des parents.

Le champ géographique

Les statistiques concernent l'ensemble du territoire. Mais pour les générations nées avant 1994, l'Insee ne dispose des fichiers statistiques de l'état civil que pour la France métropolitaine.

Naissances enregistrées / naissances domiciliées

L'état civil enregistre toute naissance ayant lieu en France, quel que soit le lieu de résidence de la mère. Il s'agit des **naissances enregistrées**. Le total des naissances en France publié dans les bilans démographiques est le total des naissances enregistrées.

Les parents des nouveau-nés peuvent ne pas habiter sur le territoire. On peut compter les naissances intervenues en France dont la mère réside en France d'après la déclaration de naissance : **on parle alors de naissances domiciliées en France**.

Se ramener à des concepts et champs les plus proches possibles pour comparer les naissances d'après le recensement à celles de l'état civil

On va comparer les effectifs d'enfants à chaque âge, entre 0 et 15 ans, présents dans les enquêtes annuelles de recensement à l'effectif dans l'état civil des naissances de la génération correspondante.

Quelques précautions sont nécessaires, pour raisonner sur des champs et concepts les plus comparables possibles.

Concept retenu pour le recensement : les enfants recensés en France et qui y sont nés

Le recensement porte sur les personnes résidant habituellement en France.

Pour se rapprocher des données de l'état civil, on considère, pour le recensement, uniquement les enfants recensés en France et qui y sont nés. En effet, les enfants nés hors du territoire français ne seront, par définition, pas inclus dans les statistiques de l'état civil. L'analyse est menée en comparant les enfants nés et recensés en France dans les EAR aux naissances enregistrées en France dans l'état civil.

Le champ géographique : France métropolitaine

La première EAR date de 2004, et donne des informations sur les naissances ayant eu lieu au cours de l'année 2003. Les enfants de 15 ans recensés en 2004 sont nés au cours de l'année 1988. En 1988, les fichiers statistiques de l'état civil disponibles couvrent uniquement la France métropolitaine.

Pour raisonner sur le même champ géographique dans les deux sources, l'analyse est donc menée en comparant les naissances d'**enfants nés en France métropolitaine et recensés en France métropolitaine dans les EAR aux naissances enregistrées en France métropolitaine d'après l'état civil**. Les constats dressés restent valables pour la France dans son ensemble (métropole + DOM).

La prise en compte des décès

Bien que ces cas soient rares, des bébés décèdent l'année même de leur naissance ou dans les années qui suivent. Ils ne seront mécaniquement pas comptabilisés dans l'EAR qui suit leur décès. Ainsi, **pour estimer le nombre des naissances de l'année précédant l'EAR à partir de l'EAR, on soustrait aux naissances de l'état civil les décès qui ont affecté la génération considérée**. On procède ainsi à chaque âge. Par

exemple, le nombre attendu d'enfants de moins d'un an de la génération N au 1^{er} janvier de l'EAR N+1 correspond au nombre de naissances de l'année N, moins les décès des enfants de cette génération survenus cette même année. On obtient ainsi, pour chaque âge, un nombre théorique d'enfants comparable à celui de l'EAR.

Ce traitement est rendu possible par les données d'état civil : en effet, les bulletins de décès ayant lieu en France sont transmis exhaustivement à l'Insee par les communes, au même titre que les bulletins de naissance.

Par adéquation au champ géographique retenu, les décès pris en compte sont les décès enregistrés en métropole.

La prise en compte des migrations

Des bébés nés en France, comptabilisés donc dans l'état civil, ont pu quitter ensuite le territoire, par exemple si leurs parents sont partis s'installer à l'étranger : ils ne sont alors pas comptabilisés au recensement après le départ des parents. En toute rigueur, en plus des décès, il faudrait soustraire aux naissances de l'état civil le nombre d'enfants ayant quitté la métropole. **Il n'est toutefois pas tenu compte ici du solde migratoire**²⁶. De manière générale, les enfants, en particulier les plus jeunes, sont peu concernés par les migrations hors de France (Le Penven, 2020). En 2016, 2000 enfants d'1 an nés en France vivaient à l'étranger (source : EDP).

La comparaison de l'effectif d'une génération recensée à un âge donné dans une EAR à l'effectif de cette génération à l'état civil n'est donc pas ajustée pour tenir compte des migrations aux jeunes âges²⁷. À noter que le correctif macro de la pyramide des âges proposé ensuite (cf. annexe 2) tient compte des migrations.

B-2. Moins d'enfants en bas âge au recensement que dans l'état civil

Comparaison du nombre des naissances d'après le recensement au nombre des naissances de l'état civil

Pour des raisons de comparaison avec l'état civil, seuls sont donc considérés dans le recensement les enfants recensés en France et qui y sont nés. Dans l'exploitation qui suit, le champ est restreint à la France métropolitaine.

Le recensement fournit par ailleurs le nombre d'enfants nés hors métropole et vivant en métropole à un âge donné. On en estime les effectifs par âge de manière séparée, pour voir comment l'arrivée des enfants nés hors métropole compense une partie de l'écart observé sur ceux nés en métropole (cf. annexe 2).

26 Pour estimer les effectifs à 0 an, on ne tient pas compte à ce stade du solde migratoire. Le solde migratoire apparent à 0 an, calculé par différence entre les effectifs à 0 an au 1^{er} janvier de l'année N+1 et les naissances de l'année N, corrigées des décès à 0 an de cette même année, est très négatif, du fait de la sous-estimation des bébés au recensement (cf. graphique A2.4 en annexe 2), mais il ne reflète *a priori* pas la réalité, compte tenu des omissions possibles de jeunes enfants.

27 Par ailleurs, le solde migratoire aux jeunes âges, calculé comme différence entre deux pyramides des âges successives, elles-mêmes estimées à partir des EAR, est sujet à caution aux jeunes âges, du fait des omissions probables de jeunes enfants au recensement. Ce point est abordé en annexe 2 sur le correctif macro.

Pour chaque âge entre 0 et 15 ans, on calcule l'écart entre le nombre d'enfants nés et recensés en France métropolitaine et le nombre de naissances enregistrées en France métropolitaine à l'état civil, diminué des décès. On procède ainsi pour chacune des Enquêtes annuelles de recensement (EAR) de 2004 à 2018. Pour chaque âge, on peut ensuite estimer l'écart moyen entre les deux sources, calculé comme la moyenne des écarts constatés dans chaque EAR sur la période 2004-2018 (cf. tableau 3-B.1).

Tableau 3-B.1: Écarts entre les effectifs nés et recensés en métropole et les naissances enregistrées à l'état civil (diminuées des décès)

Enquête annuelle de recensement	Age atteint au 1 ^{er} janvier											
	0 an	1 an	2 ans	3 ans	4 ans	5 ans	6 ans	7 ans	8 ans	9 ans	10 ans	15 ans
2004	-44 035	-42 563	-31 908	-27 987	-23 414	-26 589	-24 211	-23 856	-30 639	-25 332	-21 727	-17 378
2005	-34 354	-39 001	-28 449	-24 733	-14 866	-20 000	-19 093	-19 186	-15 876	-17 877	-17 210	-9 716
2006	-30 857	-35 760	-30 962	-27 708	-15 697	-12 786	-13 797	-10 679	-18 043	-12 883	-18 048	-14 620
2007	-38 987	-40 309	-40 690	-32 608	-22 627	-24 752	-16 379	-16 477	-18 688	-16 285	-17 416	-20 431
2008	-40 283	-43 599	-33 994	-29 599	-23 378	-23 167	-22 212	-13 796	-12 816	-10 664	-17 888	-13 680
2009	-41 788	-45 381	-36 730	-33 145	-26 641	-29 327	-23 302	-21 754	-11 916	-13 130	-18 285	-18 266
2010	-35 320	-50 253	-36 954	-31 579	-27 207	-26 605	-22 865	-17 051	-15 838	-8 199	-10 762	-10 387
2011	-36 760	-40 866	-34 873	-33 276	-29 847	-19 953	-24 631	-18 504	-13 731	-11 964	-3 946	-10 297
2012	-43 345	-43 060	-42 380	-39 773	-35 552	-34 131	-27 160	-27 677	-26 549	-17 618	-18 552	-14 074
2013	-48 041	-49 901	-39 591	-38 264	-34 450	-30 741	-25 473	-30 400	-17 077	-17 069	-16 110	-14 569
2014	-47 016	-53 315	-43 543	-37 314	-29 183	-30 213	-22 793	-21 081	-14 979	-12 656	-14 085	-11 885
2015	-53 288	-57 657	-43 808	-44 637	-39 183	-26 819	-32 718	-23 637	-19 593	-19 338	-17 509	-10 321
2016	-55 208	-68 356	-54 942	-47 482	-43 370	-32 984	-31 702	-23 585	-29 795	-25 737	-17 729	-2 749
2017	-55 303	-64 879	-56 991	-49 627	-45 567	-36 954	-26 031	-31 443	-29 591	-24 682	-21 890	-12 828
2018	-58 043	-63 596	-58 995	-59 237	-54 083	-45 807	-47 321	-35 086	-43 249	-41 593	-37 599	-26 109
Écart moyen	-44 175	-49 233	-40 987	-37 131	-31 004	-28 055	-25 312	-22 281	-21 225	-18 335	-17 917	-13 821

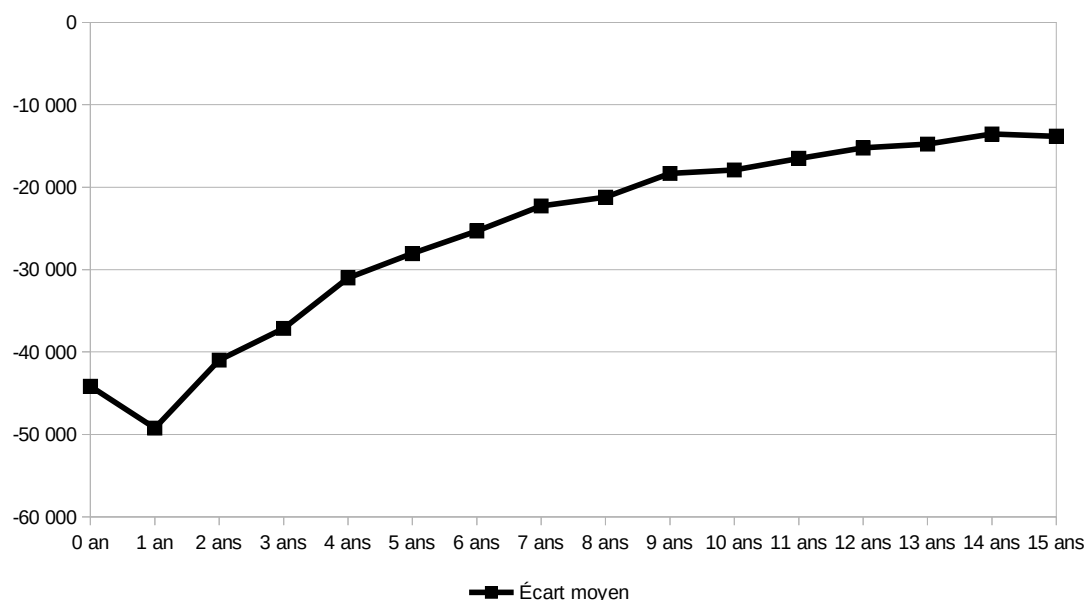
Champ : enfants recensés et nés en France métropolitaine – Naissances enregistrées en métropole, diminuées des décès enregistrés en métropole.

Source : Insee, enquêtes annuelles de recensement de 2004 à 2018 et statistiques d'état civil.

Les écarts entre les deux sources sont les plus importants aux jeunes âges. À 0 an, il y a en moyenne 44 000 enfants de moins dans les EAR que dans l'état civil. À 1 an, le déficit moyen du recensement est de 49 000 enfants. Les écarts sont encore importants aux âges suivants : - 41 000 à 2 ans, - 37 000 à 3 ans, - 31 000 à 4 ans et - 29 000 à 5 ans. Les omissions de jeunes enfants au recensement décroissent avec l'âge. À 15 ans, l'écart moyen entre l'état civil et le recensement s'établit à - 14 000 enfants.

Par ailleurs, les écarts entre les deux sources sont variables selon les années, et ont globalement tendance à s'accroître. Pour les enfants de 0 an par exemple, on trouve entre 31 000 (EAR 2006) et plus de 50 000 enfants (EAR 2015 et après) de moins au recensement qu'à l'état civil.

Graphique 3-B.1 : Les écarts entre le recensement et l'état civil se réduisent avec l'âge



Champ : enfants de 0 an recensés et nés en France métropolitaine (âge atteint au 1^{er} janvier) – naissances enregistrées en métropole, diminuées des décès enregistrés en métropole.

Source : Insee, enquêtes annuelles de recensement de 2004 à 2018 et statistiques d'état civil.

Il est naturel que des écarts existent entre les deux sources, état civil et EAR, du fait de concepts différents. Même en définissant des concepts les plus proches possibles, on ne peut réduire à zéro les écarts dus aux différences de concepts. Pour autant, cela n'invalide pas les constats présentés ici, les effectifs en jeu n'étant pas du même ordre : les écarts que l'on pourrait imputer aux différences de concept ou de temporalité²⁸ entre les deux sources sont vraisemblablement inférieurs aux écarts estimés ici. Par ailleurs, comme mentionné plus haut, une petite partie de ces écarts peut être due à des enfants partis vivre à l'étranger. Enfin, la fiabilité du nombre de naissances de mères résidant en France peut aussi être questionnée : n'y aurait-il pas des mères résidant à l'étranger qui choisiraient d'accoucher en France en déclarant une adresse en France au moment de la naissance de leur enfant ? Cela reste difficile à mesurer. Pour autant, il est probable qu'au moins une partie de ces écarts résulte d'un sous-dénombrement des jeunes enfants au recensement.

B-3. Quels déterminants possibles de ces écarts ?

D'où peuvent provenir ces omissions de jeunes enfants au recensement ? Quels en sont les déterminants ? Il est difficile d'identifier les facteurs d'omission potentiels, de mesurer leur impact et d'établir les facteurs prédominants. Dans la suite, on se concentre sur les enfants de 0 an et on les analyse selon différents critères : région, taille de la commune, nationalité des parents.

Les résultats suivants portent sur l'écart entre :

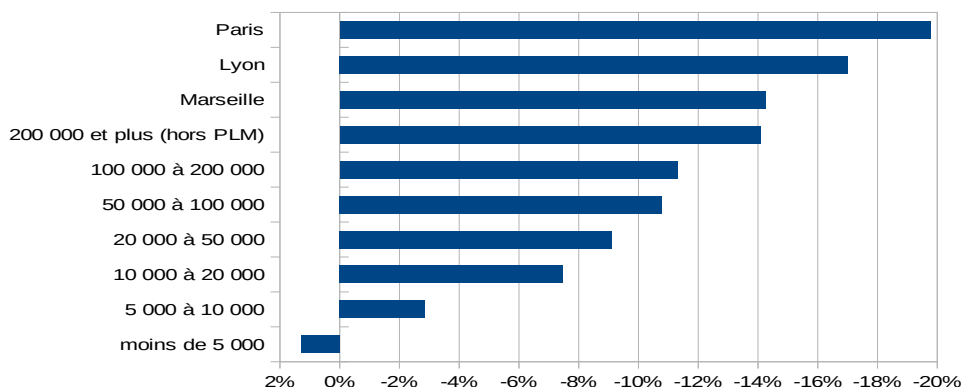
- les naissances enregistrées à l'état civil en métropole, d'enfants de mère résidant en métropole
- et les effectifs d'enfants âgés de 0 an au 1^{er} janvier dans l'EAR, nés et recensés en métropole.

²⁸ Le fait de comparer, par exemple, les effectifs d'enfants de 5 ans dans l'EAR aux naissances survenues 6 ans auparavant.

Des différences accentuées dans les métropoles, en Île-de-France et en Provence-Alpes-Côte d'Azur

Les communes de plus de 10 000 habitants concentrent la quasi-totalité de l'écart (cf. graphique 3-B.2). Plus la taille de la commune est importante, plus l'écart relatif l'est également. L'écart est ainsi particulièrement conséquent pour Paris, Lyon et Marseille.

Graphique 3-B.2 : Écarts de bébés de 0 an entre l'état-civil et l'EAR, selon la taille de la commune

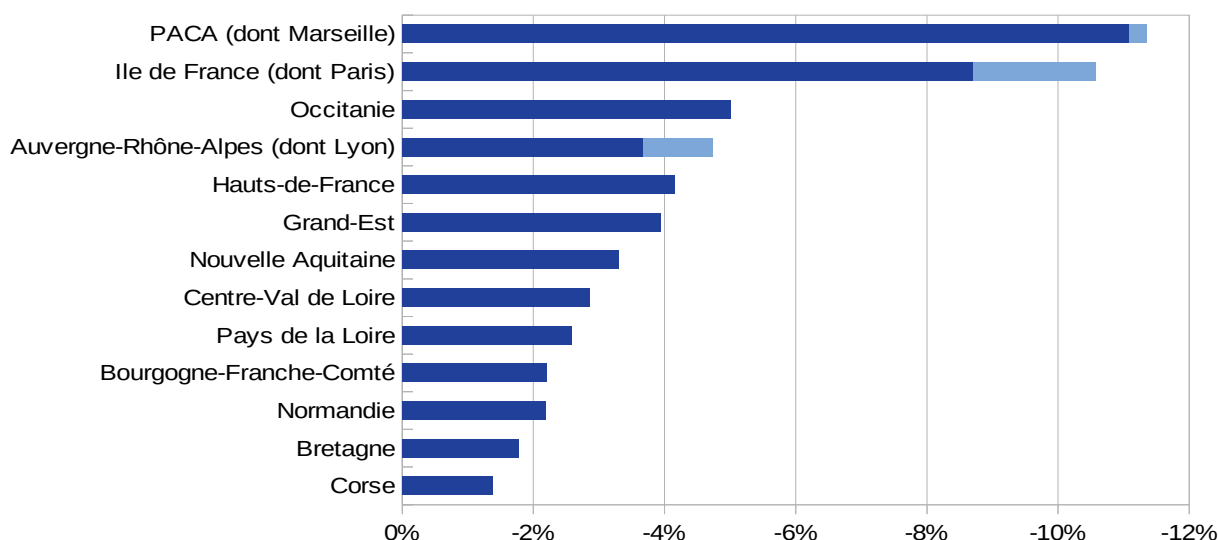


Source : Insee, enquêtes annuelles de recensement de 2004 à 2018 et statistiques d'état civil (comptabilisées au lieu de résidence de la mère).

Ces écarts, plus importants pour les grandes communes, s'expliquent en partie par la mobilité des familles. Car, si les enfants en bas âge sont peu concernés par les migrations hors de France, il n'en est pas de même au niveau local, où la migration des jeunes parents entre en jeu. La mobilité, plus forte aux bas âges, est marquée par un transfert des familles des grandes vers les petites communes : les parents, juste après la naissance, déménageraient souvent des métropoles vers le périurbain pour élever leurs enfants.

Les régions pour lesquelles l'écart entre les naissances au recensement et celles à l'état civil est le plus marqué sont les régions **Provence-Alpes-Côte d'Azur et Île-de-France** (cf. graphique 3-B.3). Les DOM présentent aussi un écart important, hormis la Guyane. Toutes les régions sont néanmoins touchées par le déficit d'enfants en bas âge.

Graphique 3-B.3 : Écarts du nombre de bébés de 0 an entre l'état civil et l'EAR, selon la région



Note : pour les régions Provence-Alpes-Côte d'Azur, Île-de-France et Auvergne-Rhône-Alpes, l'intégralité de la barre représente l'écart du nombre de bébés de 0 an entre l'état civil et l'EAR pour l'ensemble de la région, tandis que la barre foncée représente cet écart sur la région hors la principale commune (Marseille, Paris ou Lyon).

Lecture : pour l'Île-de-France, l'écart est de - 10,6 %, tandis qu'il est de - 8,7 % pour l'Île-de-France hors Paris.

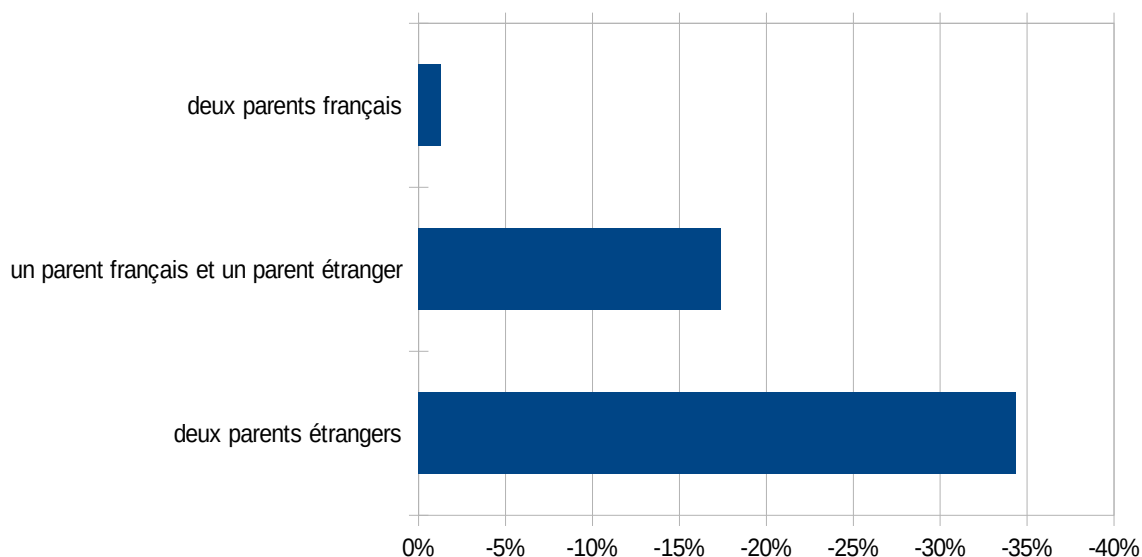
Source : Insee, enquêtes annuelles de recensement de 2004 à 2018 et statistiques d'état civil.

Au niveau régional, il semblerait aussi que, plus les grandes communes pèsent dans l'effectif régional, plus le déficit est grand. En effet, l'Île-de-France et Provence-Alpes-Côte d'Azur sont des régions plus fortement urbanisées que la moyenne des régions françaises.

Des « omissions » d'enfants au recensement plus fréquentes dans les familles étrangères ?

On constate un écart nettement plus conséquent pour les enfants nés de parents étrangers.

Graphique 3-B.4 : Écarts du nombre de bébés de 0 an entre l'état-civil et l'EAR, selon la nationalité des parents



Note de lecture : 1 % des naissances enregistrées dans l'état civil ne sont pas présentes dans le recensement suivant, pour les bébés dont les deux parents sont de nationalité française.

Source : Insee, enquêtes annuelles de recensement de 2004 à 2018 et statistiques d'état civil.

L'écart pour les bébés nés de deux parents français est de près de 1 %, alors qu'il s'élève en moyenne à 34 % pour les bébés nés de deux parents étrangers. Lorsqu'un seul parent est étranger, l'écart est de près de 17 %.

En conclusion, les différents facteurs étudiés permettent de repérer là où les omissions de jeunes enfants semblent les plus nombreuses, mais leur cause n'est pas vraiment identifiée. **Les hypothèses sont donc nombreuses, mais aucune n'a pu être validée ou ne pourrait expliquer à elle seule un sous-dénombrement des jeunes enfants au recensement. Pour les recensements étrangers, les mêmes questions se posent et restent sans réponse.**

Faute de diagnostic plus approfondi, une correction des données du recensement au niveau individuel n'est pas envisageable. Pour effectuer de tels redressements, il faudrait en effet connaître avec précision les déterminants de ces omissions.

Pour affiner ce diagnostic, l'Insee envisage de réaliser une enquête sur la qualité de la mesure de la population consécutive à une enquête annuelle de recensement (cf. conclusion). De telles enquêtes qualité sur le RP ont été menées à deux reprises par le passé : en 1962 et en 1990.

B-4. Impact sur les pyramides des âges

Les pyramides des âges, qui ventilent la population par sexe et âge, sont construites à partir des données des recensements. Le déficit potentiel de jeunes enfants influe donc directement sur ces pyramides. Une méthode de correction des pyramides des âges a été explorée (annexe 2), sur la base des comparaisons entre les données du recensement et celles de l'état civil. Rappelons tout d'abord comment sont construites ces pyramides.

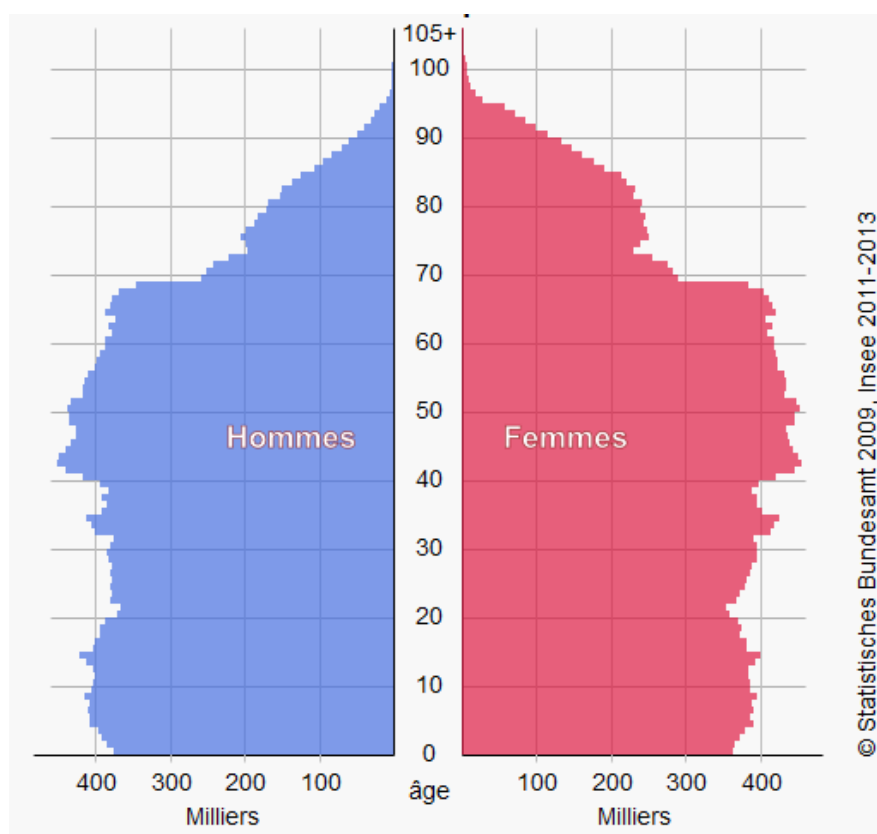
Depuis le recensement en continu, le recensement de la population (RP) cumule les résultats de la collecte de cinq EAR successives ramenées à la situation de l'année médiane : par exemple, le RP 2010 est issu du cumul des EAR 2008 à 2012. Les pyramides des âges définitives sont construites à partir des cinq EAR successives. Les effectifs par âge ne font intervenir que les résultats du recensement de la population. Elles répartissent la population par génération (année de naissance)²⁹, et non par âge atteint au recensement. Il ne s'agit donc pas d'une simple tabulation des données du recensement par âge. Le calcul de la pyramide des âges au 1^{er} janvier de l'année N s'effectue en plusieurs étapes :

- à partir des EAR, on calcule, pour chacune des cinq années de collecte, une répartition de la population par sexe et année de naissance en pondérant les effectifs par la pondération de l'EAR ;
- pour chaque sexe et année de naissance, on calcule les effectifs moyens sur les cinq EAR ;
- des corrections, d'ampleur très faible, sont réalisées avec l'aide des données d'état civil sur les décès, par exemple, pour ne pas tenir compte des personnes recensées en N-3 et N-4, mais décédées avant le 1^{er} janvier N-2 ;
- on applique un coefficient identique à tous les effectifs par sexe et âge pour obtenir le total de la population municipale ; c'est le calage sur la population du cumul RP.

Le graphique 5 présente la pyramide des âges définitive de la France métropolitaine au 1^{er} janvier 2016. Elle représente le nombre d'hommes (à gauche de la figure) et de femmes (à droite) de chaque âge résidant en France. On comprend la difficulté posée par le déficit de jeunes enfants au recensement en analysant la base de la pyramide des âges. Jusqu'à 10 ans environ, la pyramide des âges est rétrécie à sa base, les effectifs allant croissant avec l'âge. La pyramide reflète les naissances passées (les personnes de 10 ans par exemple sont la plupart nées en France, et le nombre de naissances varie d'une année à l'autre, donc le nombre de personnes résidant en France varie d'un âge à l'autre) et les mouvements migratoires (entrées et sorties du territoire). Pourtant, le solde migratoire ne peut expliquer à lui seul le rétrécissement de la pyramide aux jeunes âges. En effet, on constate que les enfants de 0 an au 1^{er} janvier 2015, nés donc en 2014, sont au nombre de 738 452 dans la pyramide des âges. Or, en 2014, l'état civil enregistre 781 167 naissances en métropole et 2 283 décès de nouveau-nés. La pyramide des âges aurait ainsi été « amputée » de 40 000 nouveau-nés, qui refléterait un solde migratoire « potentiel » en 2014 de – 40 000 à cet âge (les sorties du territoire à cet âge excèderaient de 40 000 les entrées). La variation de population entre deux pyramides des âges successives est la somme des naissances diminuée des décès et augmentée du solde migratoire. Ce solde est très élevé par rapport à celui estimé aux âges suivants (5 000 environ en moyenne entre 1 et 6 ans).

29 Document de travail n° F1003, « Construire une pyramide des âges pertinente pour le calcul des indicateurs démographiques à partir des enquêtes annuelles de recensement », S. Jugnot, M. Anguis et C. Beaumel, juin 2010.

Graphique 3-B.5 : Pyramide des âges au 1^{er} janvier 2015



Champ : France métropolitaine.
Source : Insee, estimations de population.

L'Insee a investigué sur les principes d'une méthode de redressement au niveau macro, c'est-à-dire directement au niveau de la pyramide des âges. Toute correction au niveau micro est en effet impossible à ce stade, puisqu'on ne peut estimer précisément ce déficit d'enfants, ni isoler les facteurs qui pourraient l'expliquer (cf. partie 3-B-3). Le principe serait de modifier la répartition par sexe et âge de la population, tout en gardant l'étape de calage sur la population issue du recensement de la population. Il est présenté en annexe 2.

Cette méthode s'appuie fortement sur l'état civil, en supposant donc que le nombre de naissances de mères vivant en France est une donnée robuste. Elle est basée uniquement sur une comparaison macro des effectifs entre l'état civil et les EAR, et ne peut être déclinée à tous les échelons géographiques.

Par ailleurs, si des comparaisons avec d'autres sources administratives tendent à conforter l'hypothèse d'une omission de jeunes enfants au recensement, l'ampleur de celle-ci diffère de celle évaluée à partir des données d'état civil (cf. partie 3-C). Des investigations spécifiques ont été réalisées sur une commune – celles de Grigny (91) – à partir des données de l'EAR 2015 et des données de l'Éducation nationale. Elles ne confortent pas l'hypothèse d'un sous-dénombrement d'enfants par le recensement, et ne permettent pas dès lors d'évaluer son ordre de grandeur ou ses déterminants (cf. encadré 3-C.2).

Au-delà du principe, il paraît alors difficile de savoir de combien il faudrait redresser la pyramide des âges. Dans l'immédiat, l'Insee n'envisage donc pas de mettre en œuvre un tel redressement.

C) Comparaison entre le recensement de la population et les données fiscales : Quelles différences ? Quels enseignements ?

La France n'est pas dépourvue de données administratives sur la population. Depuis 2016, l'Insee a construit et mis à disposition de la statistique publique une base de données statistique sur les logements et sur les individus mobilisant les sources fiscales. Cet assemblage de données administratives, baptisé Fidéli (Fichiers démographiques sur les logements et sur les individus), mobilise les fichiers de la taxe d'habitation, les fichiers des propriétés bâties, les fichiers d'imposition des personnes et les fichiers de déclaration des revenus. Ces données, assemblées au moyen d'identifiants fiscaux sur les foyers et les locaux, constituent le cœur de Fidéli. D'autres données complémentaires contextuelles sont également intégrées : l'appartenance à un IRIS ou un quartier de la politique de la ville, la présence de communautés ou de domiciliation administratives à l'adresse. Enfin, des informations sur le revenu disponible, les grandes masses de revenus déclarés ainsi que sur les montants des prestations sociales sont ajoutées. Fidéli est ainsi conçu pour permettre des analyses statistiques et sert désormais également pour l'échantillonnage des enquêtes ménages de l'Insee.

C.1 - Quelles différences de champ et de localisation entre Fidéli et le recensement ?

Avant de procéder à une comparaison des populations issues du recensement de la population et de Fidéli à différents échelons géographiques, il est important de rappeler comment est construite la source Fidéli et quelles sont les différences de champ ou de définition entre celle-ci et le recensement de la population, différences qui peuvent expliquer une partie des écarts mesurés.

Au niveau des logements

Le répertoire Fidéli couvre tout le bâti provenant des sources fiscales (propriétés bâties et taxe d'habitation), qu'il soit résidentiel ou non. Les locaux d'habitation sont repérés individuellement par un identifiant spécifique.

Par construction, les communautés ne contiennent pas de logements. Elles sont repérées dans Fidéli par un marquage spécifique à l'adresse. Les logements ordinaires non soumis à la taxe d'habitation ne sont pas non plus repérés en tant que tels dans Fidéli. C'est par exemple le cas des foyers logement pour jeunes travailleurs ou personnes âgées. Ces structures ressemblent à des communautés, mais n'en sont pas au sens du recensement, car chaque logement est équipé d'une cuisine et de sanitaires. Les personnes qui y habitent sont néanmoins souvent connues de Fidéli grâce à leur déclaration de revenu.

Le nombre de logements comptabilisés par le recensement est finalement proche au niveau national du nombre de logements de Fidéli : l'écart est de 0,7 % en 2017 en faveur de Fidéli, soit 260 000 logements en plus dans la source administrative. À un niveau plus fin, il y a plus de 3 % de logements de plus dans Fidéli que dans le recensement pour les départements de Seine-Saint-Denis, des Bouches-du-Rhône, du Vaucluse, du Jura et de la Meuse. En revanche, le nombre de logements dans Fidéli est plus faible que dans le recensement de plus de 4 % en Guyane, dans les deux départements de Corse, dans les Hautes-Alpes et en Guadeloupe.

Trop de logements vacants apparaissent cependant dans le Fidéli. Une partie de cet excédent est due à des mises à jour tardives dans les fichiers fiscaux de logements déjà détruits. Une autre partie tient aux règles fiscales : les occupants ont tout intérêt, par exemple en cas de déménagement, à faire en sorte qu'un logement soit considéré comme vacant au 1^{er} janvier afin d'être exonéré de la taxe d'habitation. Les personnes en question ne sont alors connues à l'adresse qu'au travers de l'impôt sur le revenu. Au total,

Fidéli compte environ 0,8 % de résidences principales (au sens de la taxe d'habitation) de moins que le recensement (soit 220 000 résidences principales).

Au niveau des habitants

Fidéli couvre toutes les personnes appartenant à un foyer fiscal sur l'ensemble du territoire français.

Au niveau des habitants, des filtres sont utilisés pour exclure au mieux les personnes qui ne résident pas en France, car cette information n'est pas directement disponible dans les sources fiscales. Par ailleurs, comme plusieurs sources fiscales sont mobilisées, Fidéli effectue des traitements d'appariements entre ces sources sur la base des noms, prénoms, lieu de naissance et identifiant fiscal, encadrés juridiquement. Ces appariements visent à repérer le plus précisément possible les individus distincts. Dans certains cas, ces appariements s'avèrent plus complexes, certains enregistrements étant plus flous (patronyme partiel, lieu de naissance non renseigné, etc.).

Par ailleurs, dans les fichiers fiscaux, il manque des jeunes adultes, majoritairement de 20 à 22 ans. Une des raisons est probablement qu'ils ont quitté le foyer parental et vivent en communauté, ou du moins dans un logement qui n'est pas soumis à la taxe d'habitation, et qu'ils n'apparaissent sur aucune déclaration de revenu. Lorsqu'ils peuvent être retrouvés dans les fichiers fiscaux des années antérieures, ils sont ajoutés au millésime courant. Leur localisation est sans doute médiocre, mais cela permet de corriger un léger creux dans la pyramide des âges au niveau national.

Outre les personnes vivant en communauté ou dans des logements non assujettis à la taxe d'habitation, certains contribuables ne sont connus qu'au travers d'une adresse de domiciliation, donc sans logement identifié (demandeurs d'asile, gens du voyage, sans domicile fixe...). Lorsque ces personnes ne font pas de déclaration de revenus, elles ne sont pas présentes dans les fichiers Fidéli.

En matière de localisation des personnes

Fidéli enregistre toutes les adresses connues pour une personne donnée. Dans les données fiscales de 2016, 7,3 millions d'adultes, soit 13 % d'entre eux, ont ainsi plus d'une résidence déclarée.

Ainsi, l'adresse déclarée comme principale dans les fichiers de la taxe d'habitation (TH) est différente de l'adresse déclarée à l'impôt sur le revenu dans 17 % des cas de multi-résidence avec environ 820 000 personnes localisées (soit 11 % des cas) dans des communes différentes entre ces deux fichiers. La localisation retenue comme prioritaire est alors celle enregistrée comme résidence principale dans les fichiers de la taxe d'habitation.

En matière de localisation géographique, si, dans la plupart des cas, la résidence principale déclarée au recensement coïncide avec la résidence principale au sens de la TH prise comme référence dans Fidéli, ce n'est pas le cas pour certaines populations spécifiques :

- les **étudiants majeurs** vivant dans un logement non soumis à la taxe d'habitation (cité universitaire par exemple) seront localisés chez leurs parents dans Fidéli si ces derniers les font figurer sur leur déclaration de revenu, alors qu'ils seront recensés dans leur logement situé sur leur lieu d'étude selon les règles du recensement³⁰ ;
- les **enfants en résidence partagée** sont comptés dans Fidéli pour moitié dans chacune des communes de leurs parents alors que le recensement les compte là où ils passent le plus de temps (et en cas d'égalité là où ils se trouvent la nuit précédant le début du recensement) ;
- les **personnes occupant un logement occasionnel** (par exemple les célibataires géographiques) pourront déclarer ce logement comme résidence principale à la TH alors que, pour le recensement, ils seront comptés dans leur domicile familial (celui-ci pouvant même éventuellement être à l'étranger).

30 Pour les étudiants mineurs, les règles du recensement prévoient de les compter dans la commune de domiciliation de leurs parents, ce qui est conforme à leur localisation dans Fidéli.

- La source fiscale contient également des personnes connues de l'administration fiscale à leurs **domiciliations administratives** qui ne correspondent pas nécessairement à la commune dans laquelle ils sont recensés (par exemple les gens du voyage, les sans-abris).
- pour des raisons fiscales, certaines personnes choisissent de **localiser fiscalement leur résidence principale dans un logement qui est une résidence secondaire** au sens du recensement.

C.2 - Au niveau national, seulement 0,6 % d'écart sur la population française mesurée par les deux sources en 2017

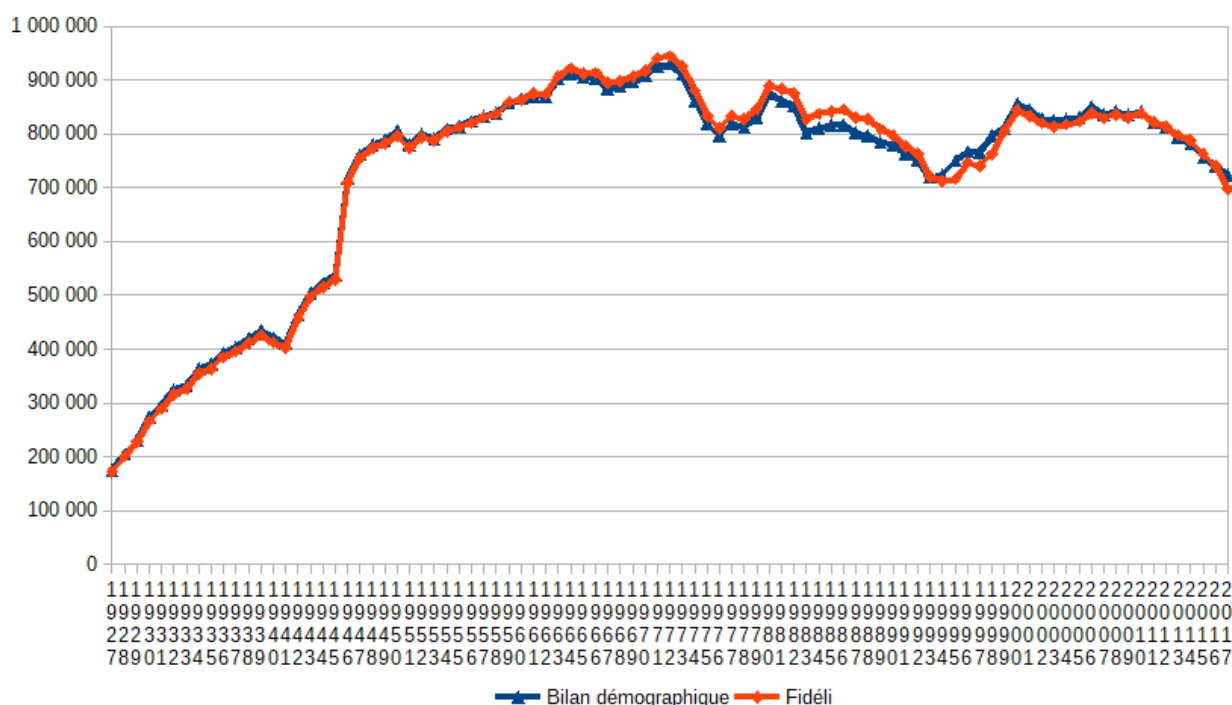
En 2017, d'après le recensement, la population française s'établit à 66 524 000 habitants (métropole et DOM hors Mayotte). Cette population respecte les critères européens et internationaux de population usuelle résidente (Unece, 2015). Selon ces critères, pour être compté comme habitant d'un pays (ou d'une commune), il faut que la personne ait sa résidence usuelle dans ce pays (ou commune), c'est-à-dire qu'elle y passe habituellement ses périodes de repos et qu'elle y réside ou envisage d'y résider au moins un an. En particulier, les étudiants Erasmus ou les travailleurs venus résider quelques mois en France ne sont pas comptabilisés dans la population recensée.

Fidéli comptabilise de son côté 66 948 000 personnes en 2017 sur le même champ, soit un écart de 0,6 %. Il y a donc environ 400 000 personnes de plus dans Fidéli par rapport au recensement.

Pour affiner la comparaison, le graphique 3-C.1 présente l'écart de population entre Fidéli et le bilan démographique par âge fin. La comparaison est faite sur l'année 2018. Il est plus pertinent de s'appuyer sur les données du bilan démographique que sur celles du recensement pour faire une comparaison par âge, dans la mesure où le recensement de la population s'appuie sur 5 enquêtes annuelles de recensement, ce qui tend à lisser la taille des populations entre les différentes générations (Insee, 2018b). Cela est retraité dans le cadre du bilan démographique.

Les profils par âge des populations issues des deux sources sont très proches (cf. graphique 3-C.1).

Graphique 3-C.1 : Population selon la génération dans Fidéli d'une part, et le bilan démographique d'autre part, au 1^{er} janvier 2018



Source : Insee, Fidéli 2018 et bilan démographique.
Champ : France entière (y compris Mayotte).

Plus précisément (cf. tableau 3-C.1), on constate que Fidéli compte un peu moins d'enfants que le bilan démographique (- 0,2 % pour les moins de 5 ans et - 0,4 % entre 6 et 9 ans) et sensiblement moins de jeunes de 10 à 17 ans (- 1,1 %) et de jeunes adultes de 18 à 24 ans (- 2,4 %). À l'inverse, il y a davantage de personnes âgées de 25 à 54 ans dans Fidéli que dans le bilan démographique.

Plus précisément, Fidéli enregistre un déficit de 27 000 enfants âgés de 0 an par rapport au bilan démographique. En comparaison avec les données de l'état civil, le déficit est alors encore plus important : de l'ordre de 70 000. L'écart est donc plus important qu'entre le recensement et l'état civil (cf. partie 3-B). Ensuite, le nombre d'enfants d'un an est très proche entre les deux sources, et Fidéli compte plus d'enfants que le bilan démographique entre 2 et 5 ans (18 000 d'écart en cumulé). Ce léger surplus pourrait correspondre à un défaut de déclaration des enfants au recensement de la population, défaut évoqué dans la partie 3-B et qui n'est pas propre à la France (Census bureau, 2019 et Toulemon, 2017). Cependant, d'après ces chiffres, la sous-estimation d'enfants au recensement pourrait être moins forte que l'estimation réalisée à la partie 3-B, à moins que le nombre d'enfants soit également sous-estimé à partir des sources fiscales, comme il semble l'être pour les enfants de 0 an.

Par ailleurs, dans certains cas, il est également possible que des enfants soient rattachés fiscalement à des foyers fiscaux sans être effectivement présents (cf. encadré 3-C.2).

Tableau 3-C.1 : Écarts de population par tranche d'âge entre Fidéli et le bilan démographique, au 1^{er} janvier 2018

	Bilan démographique	Fidéli	Écart absolu Fidéli/bilan démographique	Écart relatif Fidéli / bilan démographique (en %)
0 à 5 ans	4 607 468	4 599 414	-8 054	-0,2
6 à 9 ans	3 339 668	3 325 701	-13 967	-0,4
10 à 17 ans	6 694 070	6 618 784	-75 286	-1,1
18 à 24 ans	5 331 462	5 202 930	-128 532	-2,4
25 à 34 ans	7 920 402	8 154 805	234 403	3,0
35 à 44 ans	8 432 449	8 605 587	173 138	2,1
45 à 54 ans	9 051 352	9 154 681	103 329	1,1
55 à 64 ans	8 369 055	8 364 914	-4 141	0,0
65 à 74 ans	6 994 431	6 925 380	-69 051	-1,0

Source : Insee, Fidéli 2018 et bilan démographique.

Champ : France entière (y compris Mayotte).

La population issue de Fidéli apparaît également en déficit pour les jeunes adultes, comme cela était évoqué dans la partie précédente. Ce déficit de 128 000 personnes entre 18 et 24 ans peut s'expliquer notamment par le fait que ces jeunes adultes ne sont plus déclarés dans la déclaration de revenus de leurs parents et ne réalisent pas encore eux-mêmes leur propre déclaration de revenus. S'ils vivent dans un logement non soumis à la TH ou chez des amis, ils peuvent alors n'être présents dans aucun fichier fiscal.

La population Fidéli présente en revanche un excès de 408 000 personnes âgées de 25 à 44 ans. Une des causes de cet excès pourrait être que, malgré les filtres mis en place pour les repérer, les données fiscales contiennent des travailleurs non-résidents pour ces tranches d'âge.

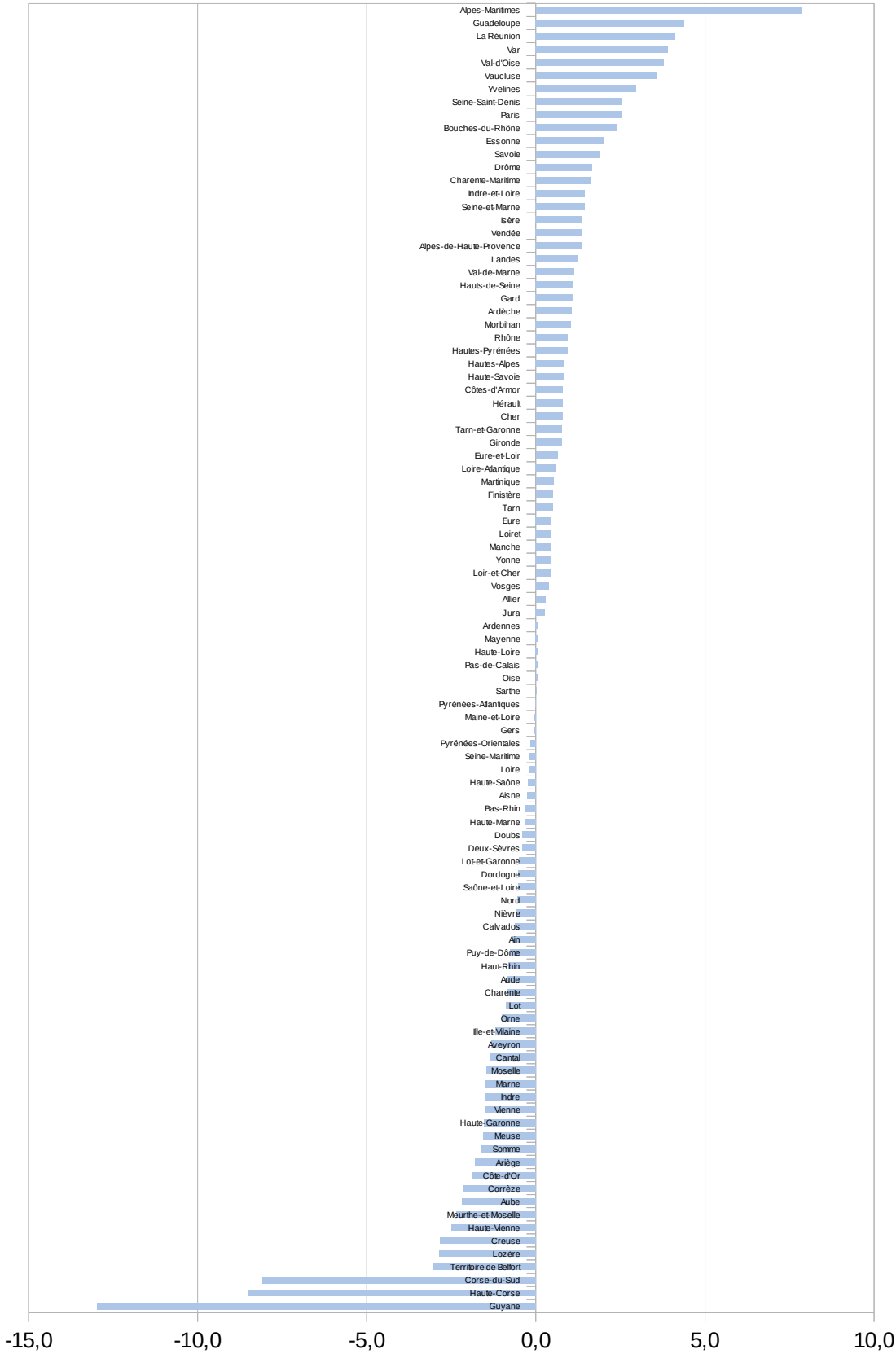
C.3 - Des écarts entre le recensement et Fidéli de signes et d'ampleur différents selon les départements

Les écarts de population selon les départements entre le recensement et Fidéli se situent entre - 13 % (Guyane) et + 7,8 % (Alpes-Maritimes). Hors Dom, Alpes-Maritimes et Corse, la fourchette est plus resserrée, comprise entre - 3 % pour le Territoire de Belfort et + 3,9 % pour le Var.

Ces écarts sont inférieurs en valeur absolue à 1 % dans la moitié des départements, et ils excèdent 3 % dans 10 départements.

En particulier, la population au recensement est supérieure à la celle de Fidéli de plus de 3 % dans quatre départements : la Guyane, la Corse du Sud, la Haute-Corse et le Territoire de Belfort. À l'inverse, les populations fiscales sont supérieures de plus de 3 % aux populations recensées dans six départements : les Alpes-Maritimes, la Guadeloupe, La Réunion, le Var, le Val-d'Oise et le Vaucluse.

Graphique 3-C.2 : écart entre le nombre d'habitants au recensement et dans Fidéli par département en 2017



Source : Insee, Fidéli 2017 et RP 2017.
 Champ : France hors Mayotte.

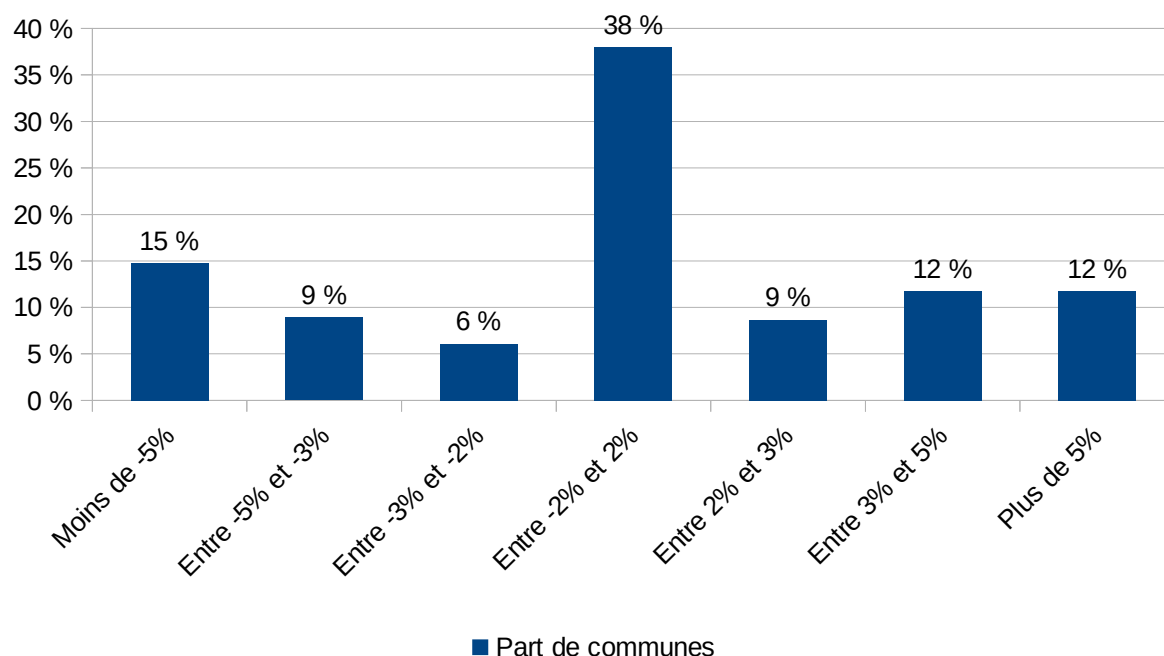
C.4 - Des écarts plus marqués au niveau communal, notamment dans les plus petites communes

Les écarts de population entre recensement et sources fiscales peuvent être plus importants au niveau communal, compte tenu du nombre élevé de communes et de leur hétérogénéité en matière de taille.

Les écarts semblent plus importants pour les plus petites communes, celles de moins de 2 000 habitants. La raison la plus généralement évoquée est que les différences de concepts entre les deux sources et les retards de mise à jour des données fiscales se compensent moins facilement dans les communes de petite taille.

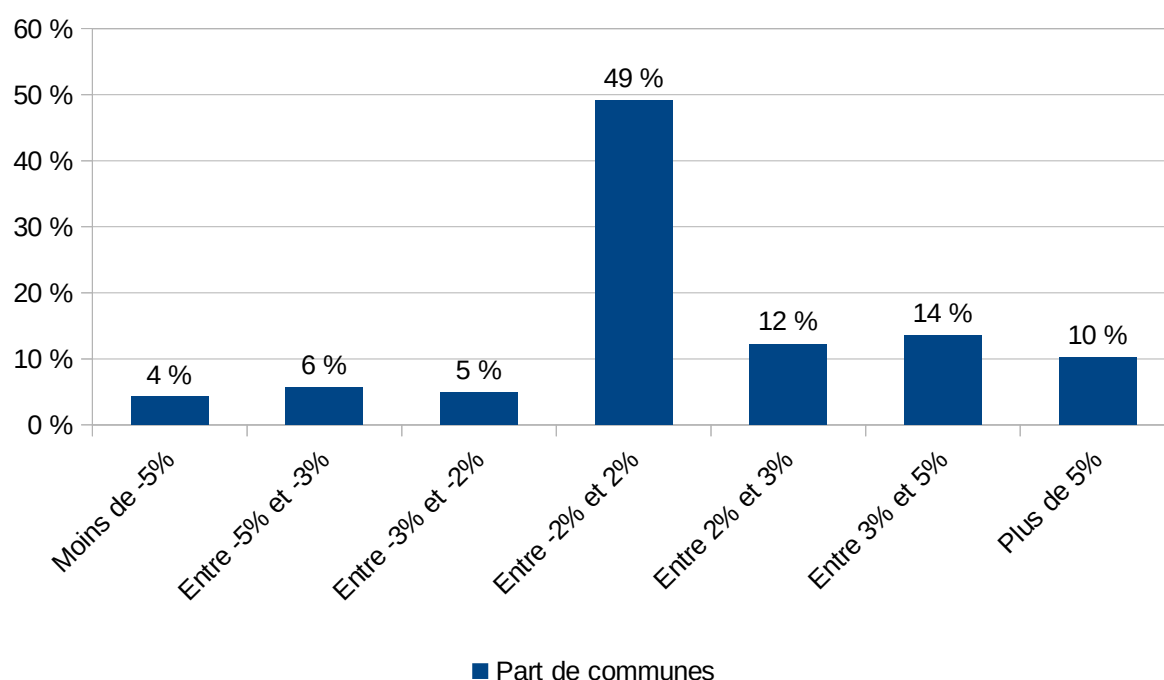
Ainsi, 38 % des 35 000 communes de France hors Mayotte ont un écart entre la population mesurée par Fidéli et le recensement de moins de 2 % en valeur absolue. Mais 62 % des communes présentent un écart de plus de 2 % et 27 % de plus de 5 %. Si on se restreint aux communes de plus de 2 000 habitants au recensement, la distribution des écarts entre les deux sources est plus concentrée autour de 0 %. Sur ce champ, 49 % des communes présentent un écart de moins de 2 % en valeur absolue, et 14 % des communes présentent un écart de plus de 5 %.

Graphique 3-C.3 : Distribution des écarts communaux entre Fidéli et le recensement en 2017 – Ensemble des communes



Source : Insee, Fidéli 2017 et RP 2017.
Champ : France hors Mayotte.

Graphique 3-C.4 :Distribution des écarts communaux entre Fidéli et le recensement en 2017 – Communes de plus de 2 000 habitants



Source : Insee, Fidéli 2017 et RP 2017.

Champ : France hors Mayotte.

Les travaux d'analyse ont montré par ailleurs qu'une partie des écarts entre sources était due aux personnes vivant en communautés, notamment les jeunes qui, lorsqu'ils ne font pas de déclaration de revenus, ne peuvent être localisés par la source fiscale qu'au domicile de leurs parents (puisque'ils ne paient pas de taxe d'habitation).

Même en ne s'intéressant qu'aux communes de plus de 100 000 habitants, on constate encore des écarts marqués entre sources, parfois positifs, parfois négatifs. En effet, sur les 50 communes concernées, seules 16 présentent des écarts inférieurs à 3 % en valeur absolue. Ce nombre double si on traite au cas par cas les populations vivant en communautés. Reste néanmoins une petite moitié de ces grandes communes pour lesquelles les écarts sont significatifs. Ainsi, par exemple, la population Fidéli est nettement inférieure à la population recensée à Caen (- 7,3 %), Dijon (- 7 %), Rennes (- 6,6 %) ou dans le 16e arrondissement de Paris (- 6,8 %). À l'inverse, la population Fidéli est souvent supérieure à la population recensée en région parisienne, par exemple à Paris 19e (10 %), 12e (8,8 %), à Argenteuil (7,1 %), Montreuil (6,8 %) ou encore à Saint Denis (5 %). C'est le cas aussi à Nice (10,9 %).

La comparaison entre recensement et sources fiscales montre ainsi que, si les résultats en nombre d'habitants sont proches au niveau national, ils peuvent sensiblement s'écarter au niveau local, probablement en raison de différences de localisation de la résidence principale retenue par les deux sources ou de champs pas complètement homogènes (cf partie 3-C.1). Les écarts ne sont ainsi pas systématiquement de même signe. On peut constater qu'en Île-de-France, la population recensée est plutôt inférieure à celle mesurée par Fidéli, tandis que dans les grandes agglomérations de province c'est souvent l'inverse.

Des analyses économétriques ont été menées au niveau des communes et des départements pour tenter d'extraire des facteurs explicatifs des écarts observés. Deux variables ont particulièrement été analysées : la part de résidences secondaires et la part d'étrangers. Ces analyses n'ont pas mis en évidence de corrélation forte entre les écarts mesurés entre sources et ces deux variables.

C.5 - Une autre approche possible pour comparer recensement et données fiscales : retrouver les personnes de l'échantillon EDP recensées en 2018 dans Fidéli 2018

L'objectif de notre analyse a été de partir des bulletins de l'enquête de recensement de 2018 et de voir si l'on retrouvait bien ces personnes dans Fidéli, puis pour les personnes retrouvées de comparer leur lieu de résidence.

L'appariement est conduit sur la totalité du champ des personnes « exploitables », i.e. dont le nom, le prénom, l'année de naissance et le lieu de naissance sont connus. L'appariement est réellement utilisé seulement pour les seules personnes majeures, nées avant 2001. Sur ce champ restreint, qui compte 771 884 individus, 95,3 % des individus disposent d'informations d'état civil exploitables.

Sur le champ des personnes mineures, l'appariement est moins bon, car, dans les fichiers fiscaux, les informations détaillées sur l'état civil ne sont pas toujours disponibles pour ces personnes.

De nombreuses opérations préalables sont nécessaires avant d'apparier les données en raison de problèmes lexicographiques, de l'utilisation du nom d'épouse et d'un seul prénom dans les réponses au recensement de la population.

Au total, 94 % des personnes appartenant à l'échantillon EDP (cf. encadré 3-C.1) et recensées ont pu être retrouvées dans les sources fiscales. Les disparités sont nettes selon l'âge. Sans surprise, les taux d'appariement sont moindres avant 30 ans, âges où des personnes encore en formation peuvent résider en communauté (résidence étudiante...) sans être connues des services fiscaux. Le taux d'appariement est de 86 % pour les moins de 30 ans. Les taux d'appariements sont un peu plus faibles chez les femmes au-delà de 30 ans, en raison des difficultés à repérer leur patronyme sous une forme équivalente dans les deux sources. Certains taux sont en revanche particulièrement élevés, aux alentours de 97 % pour les hommes proches de la cinquantaine. Il est ainsi difficile d'en tirer des conclusions claires sur l'absence de personnes recensées dans Fidéli, la majorité étant probablement due à des défauts de qualité des données servant à l'appariement.

Encadré 3-C.1 : présentation succincte de l'échantillon démographique permanent (EDP)

L'échantillon démographique permanent est un panel socio-démographique de grande taille, mis en place en France, pour étudier la fécondité, la mortalité, les parcours familiaux, les migrations géographiques au sein du territoire national, la mobilité sociale et la mobilité professionnelle, les carrières salariales et les niveaux de vie ainsi que les interactions possibles entre ces différents aspects.

L'échantillon démographique permanent a été créé en 1967.

Il contient des informations issues de cinq sources :

- des bulletins d'état civil de naissance, de mariage, de décès depuis 1968
- des recensements de 1968, 1975, 1982, 1990 et 1999 puis des enquêtes annuelles de recensement à partir de 2004
- du fichier électoral depuis 1967
- du panel " tous salariés " depuis 1967
- des données socio-fiscales depuis 2011 (revenus 2010), notamment venant du dispositif Filosofi.

Pour les personnes appariées se pose la question de la correspondance de la commune de résidence entre les deux sources (recensement et Fidéli). On se limite ici au seul champ des personnes EDP appariées

précédemment et pour lesquelles le code de la commune (ou d'arrondissement) est renseigné dans le recensement, soit environ 282 000 individus. Il en résulte que :

- **pour 94,1 % des personnes, le code commune (ou arrondissement) du recensement correspond au code commune (ou arrondissement) de Fidéli (issu de la taxe d'habitation ou à défaut impôt sur le revenu) ;**
- pour 1,6 % des personnes, le code commune du recensement est retrouvé dans les déclarations fiscales des individus, mais à une autre localisation que celle de Fidéli. Il s'agit dans 86 % des cas de la localisation d'une résidence secondaire, sinon d'une incohérence interne aux fichiers fiscaux (FIP et PLFC) sans doute due à un retard de mise à jour ;
- pour 0,8 % des personnes, le code commune du recensement est celui de la dernière adresse connue, en général une adresse destinée à remplacer l'adresse Fidéli au-delà du 1^{er} janvier (notamment en cas de déménagement lorsque la nouvelle adresse est connue par anticipation, par exemple en cas d'emménagement au-delà du 1^{er} janvier pour être exonéré de la taxe d'habitation) ;
- pour 1 % des personnes, le code commune de Fidéli est différent du code commune actuel du recensement, mais correspond au code commune du recensement l'année précédente ; les deux sources sont cohérentes, les écarts provenant d'un décalage de mise à jour des fichiers fiscaux ;
- pour 2,5 % des individus, la commune diffère entre les deux sources sans qu'il y ait trace de la commune au recensement présente ou passée dans les fichiers fiscaux pour la personne en question.

Comme on pouvait s'y attendre, les taux de correspondance sont meilleurs en ménage ordinaire.

Il en résulte que la localisation des personnes en logement ordinaire est proche dans le recensement et Fidéli, sans toutefois coïncider exactement. Une partie des écarts est sans doute due à des délais de mise à jour de la source fiscale, ou à des choix spécifiques conduisant à ce que la personne déclare au fisc comme résidence secondaire sa résidence principale au recensement. Au-delà de 40 ans, il est très rare que la localisation au recensement soit inconnue dans la source fiscale. Ces écarts tendent toutefois à expliquer la divergence de population observée entre les sources à des niveaux locaux.

Encadré 3-C.2 : Une comparaison recensement / fichiers de l'Éducation nationale sur la commune de Grigny

Une autre expérience de comparaison entre le recensement et un fichier administratif a été menée sur la commune de Grigny (91), suite à des doutes répétés émis par cette commune sur la qualité de la population recensée. Cette étude a permis de comparer au niveau individuel les résultats de l'enquête annuelle de recensement de 2015 et les informations figurant dans le fichier du Rectorat de Versailles pour l'année scolaire 2015-2016.

Dans la mesure où les comparaisons ont été effectuées à partir de données individuelles nominatives, la procédure d'obtention par l'Insee des fichiers de l'Éducation nationale a obéi aux dispositions réglementaires en vigueur, à savoir une demande d'accès aux données au titre de l'article 7bis de la loi du 7 juin 1951 modifiée adressée au Conseil national de l'information statistique (Cnis) et une autorisation de traitement accordée par la Commission nationale informatique et des libertés (Cnil) le 27 novembre 2015.

L'objectif de ce travail était d'identifier si des enfants manquaient au recensement par rapport aux informations présentes dans les bases de l'Éducation nationale en lien avec la problématique développée dans ce document dans la partie 3-B.

La comparaison a reposé sur l'enquête annuelle de recensement de 2015 de la commune de Grigny, dont l'échantillon était composé de 99 adresses et 792 logements. Dans ces logements, se trouvaient 365 bulletins individuels d'enfants nés entre le 1^{er} janvier 2005 et le 1^{er} janvier 2011 (c'est-à-dire âgés de 3 à 10 ans) ou nés en 2004 ou 2012 et scolarisés dans les classes observées (afin de prendre en compte les

enfants scolarisés précocement et ceux ayant redoublé). Ces enfants étaient domiciliés au sein de 228 logements dans 47 adresses.

La méthode de travail a consisté à comparer, pour toutes ces adresses, le nombre d'enfants recensés et le nombre d'enfants scolarisés d'après l'Éducation nationale. Les comparaisons ont aussi reposé sur l'identité des enfants présents dans les deux sources (nom, prénom, date de naissance).

Le premier résultat brut montre que 365 enfants ont été recensés à ces adresses alors que 252 sont enregistrés sur ces mêmes adresses dans la base de l'Éducation nationale de la commune de Grigny.

Plus précisément : 196 enfants ont été retrouvés aux adresses de l'échantillon dans les deux sources, 125 enfants étaient présents dans l'EAR et absents de la base Éducation nationale sur les adresses de l'échantillon et 56 enfants étaient absents de l'EAR et présents dans la base Éducation nationale. Par ailleurs, pour 44 enfants, la date de naissance n'était pas suffisamment complète dans l'EAR pour qu'on puisse les repérer dans le fichier de l'Éducation nationale.

Une recherche plus poussée a ensuite été menée sur les 56 enfants non présents dans le recensement. Quatre adresses regroupant 28 enfants ont été particulièrement analysées. Pour 4 enfants, le logement a bien été recensé avec un nom de famille correspondant à l'enfant tel que connu à l'Éducation nationale, mais aucun bulletin individuel n'a été rempli pour cet enfant³¹. Les 24 autres cas correspondent à des noms de famille non recensés à l'adresse. Il est possible que des logements n'aient pas été enquêtés et aient fait l'objet d'une procédure d'estimation sans connaître le nom de la famille (voir redressement des fiches de logement non enquêté - partie 2-D), que des logements aient été inscrits à tort en logement vacant ou aient été oubliés ou que l'adresse connue par l'Éducation nationale pour ces enfants ne soit plus d'actualité au moment de l'EAR en raison d'un déménagement en cours d'année (les adresses de la base éducation nationale étant renseignées en septembre 2014).

Parmi les 125 enfants présents dans le recensement mais pas dans la base Éducation nationale dans les adresses de l'échantillon, 20 ont été retrouvés sur une autre adresse de Grigny (ces enfants sont enregistrés par l'Éducation nationale à une adresse hors de l'échantillon de l'EAR 2015, peut-être en raison d'un déménagement), et 105 enfants sont absents du fichier de l'Éducation nationale de Grigny. En élargissant la recherche aux communes limitrophes, on en retrouve 88 dans une autre commune. Les écarts observés entre les données de l'EAR et celles de la base Éducation nationale ne doivent donc en aucun cas être interprétés en termes d'enfants déscolarisés.

Plus globalement, dans les fichiers de l'Éducation nationale, de la maternelle au CM1, le nombre d'enfants habitant Grigny quel que soit leur lieu de scolarisation est de 3 459. D'après le recensement, en 2015, on en a 4 088 entre 3 et 9 ans ou 4 028 entre 4 et 10 ans. **Le nombre d'enfants à Grigny est donc plus élevé dans le recensement que dans les fichiers de l'Éducation nationale.** Il convient toutefois de noter que certains enfants en âge d'être scolarisés en classe maternelle peuvent ne pas y aller, la scolarisation n'étant pas obligatoire à cet âge en 2015.

Si l'on se restreint au champ de l'école primaire, les résultats sont plus proches : 2 150 enfants nés entre 2005 et 2008 dans les fichiers de l'Éducation nationale de l'année scolaire 2014-2015, 2 280 enfants dans le recensement âgés de 6 à 9 ans et 2 189 âgés de 7 à 10 ans.

Cette analyse pointue menée sur une commune spécifique ne permet donc pas de conforter l'hypothèse d'une omission d'enfants dans le recensement : en effet, les estimations pour cette commune restent cohérentes entre recensement et Éducation nationale. Au-delà, elle montre combien il est difficile de comparer deux sources de données différentes.

31 À noter : jusqu'à l'EAR 2017, une personne qui était déclarée dans la feuille de logement mais sans bulletin individuel associé n'était pas comptabilisée dans le décompte de population, les informations recueillies étant jugées trop pauvres. Depuis l'EAR 2018 et le changement du questionnaire logement, toutes les personnes déclarées dans la feuille de logement avec leurs nom, prénom, sexe et date de naissance sont comptabilisées même si aucun bulletin individuel n'a été rempli.