Nations Unies ECE/CES/GE.41/2019/9

Conseil économique et social

Distr. générale
3 July 2019

Original : français

Commission économique pour l'Europe

Conférence des statisticiens européens

Groupe d'experts des recensements de la population et des habitations

Vingt et unième réunion
Genève, 18-20 septembre 2019
Point 2 de l'ordre du jour provisoire
Résultats des essais menés, en ce qui concerne les méthodes, les techniques, la participation et d'autres aspects

Utiliser les prénoms pour améliorer la mesure des couples de personnes de même sexe dans le recensement

Note de l'Institut national de la statistique et des études économiques (INSEE), France*

Summary

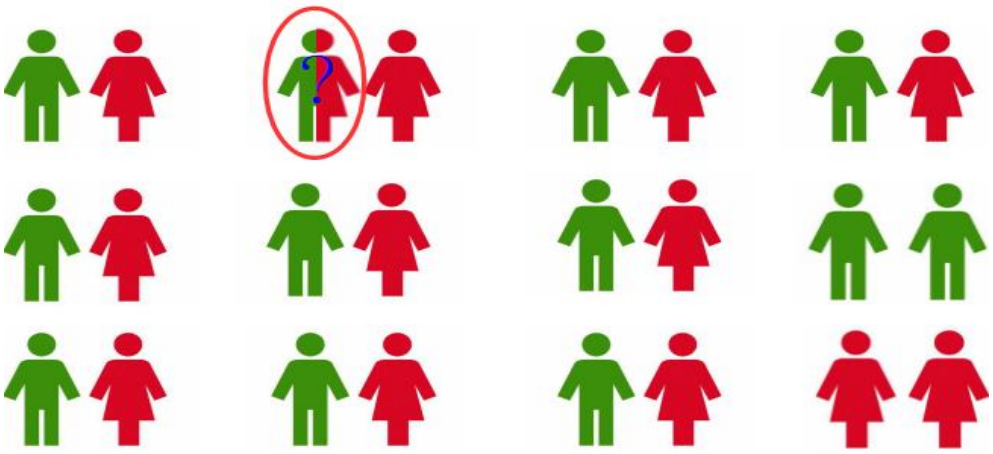
Il est difficile d'établir à partir du recensement français des statistiques fiables concernant le nombre de couples de personnes de même sexe en France. En effet, un nombre important de couples de même sexe, plus de 40pour cent d'après une étude de 2011, est compté comme tel du fait d'une erreur de codage sur le sexe d'un des conjoints. Cela conduit à une surestimation. La procédure de correction proposée consiste à calculer dans quelle proportion un prénom est plutôt masculin ou féminin et à utiliser cette information pour redresser la variable de sexe pour les personnes qui, d'après les données du recensement, vivent au sein d'un couple de personnes du même sexe. Cette méthode semble efficace et donne des résultats cohérents avec ceux d'autres sources.

* Établie par Elisabeth Algava et Sébastien Hallépée.

I. Introduction

1. Il est actuellement impossible d'établir à partir du recensement des statistiques fiables concernant le nombre de couples de personnes de même sexe (CMS) qui cohabitent en France, du fait d'une difficulté méthodologique. En effet, dans un couple de personnes de sexes différents, une erreur de déclaration ou de codage sur le sexe d'un seul des conjoints aboutit en général à compter ce couple comme étant de même sexe. Si cela ne concerne qu'une toute petite proportion des personnes en couple de sexe différent, cela suffit à surestimer très fortement la proportion de personnes en CMS. Ce risque n'est pas propre aux couples de même sexe. Il est présent dès qu'il s'agit d'estimer des populations rares, concernant un petit effectif de personnes enquêtées. Qu'il s'agisse de mesurer le nombre de veufs de moins de 30 ans, ou de personnes mariées à 18 ans, il faut tenir compte des erreurs sur l'âge ou le statut matrimonial, erreurs dont la fréquence peut dépasser celle des mariés ou veufs précoces. Mais la particularité est qu'ici une erreur sur le sexe au niveau individuel, pour un seul des conjoints, va conduire à considérer les deux conjoints comme ayant un partenaire de même sexe (Figure I) : une erreur compte double.

Figure I
Impact d'une erreur de codage sur la variable sexe



Note de lecture : Cette population fictive comporte 12 couples et 24 personnes. 9 sont des couples de sexe différent (CSD), 2 sont des CMS et il subsiste un doute sur le codage du sexe d'un des individus du dernier couple.

S'il s'agit d'une erreur de codage, 1 erreur sur 24 va faire basculer 1 couple sur 12 de CSD à CMS. L'erreur compte donc « double ». Le nombre de CMS augmenterait ainsi artificiellement de 50 pour cent alors que le nombre de CSD ne serait réduit artificiellement que de 10 pour cent. On retrouve le fait que les faibles erreurs de codage ont un impact beaucoup plus visible sur les populations rares.

2. Cette difficulté a pu être contournée en 2011, lorsque l'Insee a réalisé l'enquête sur les familles et les logements (EFL). Pour un échantillon des logements recensés, un questionnaire complémentaire de 4 pages était déposé en même temps que ceux du recensement. Environ 360 000 questionnaires ont été collectés. L'EFL a permis de nouvelles études sur de nombreux sujets (Bodier et al., 2015 ; Imbert et al., 2018) dont celui des couples de même sexe. En effet, un important travail de vérification des réponses a été réalisé, par croisement des informations disponibles dans le recensement et l'EFL (Breuil-Genier et al., 2016). Il a permis d'estimer le nombre de personnes en couple de même sexe à 205 000, dont 173 000 cohabitent. 0,57 pour cent des couples cohabitant étaient des couples de même sexe et 0,36 pour cent des « faux » couples de même sexe (Buisson et Lapinte, 2013 ; Banens et Le Penven, 2016). L'importance de la correction est donc assez considérable sur les CMS : de 295 000 personnes concernées avant redressement à 173 000 après. Plus de 40 pour cent des situations ont donc été corrigées. Cette étape de correction a permis de nouvelles analyses concernant les personnes en couples de même sexe en 2011 (Rault, 2016, 2017, 2018).

3. La prochaine enquête Famille, qui permettra en principe d'avoir un nouveau chiffrage contrôlé du nombre de couples de même sexe, n'est pas envisagée avant 2023. Disposer de données plus rapidement est important, compte tenu des évolutions récentes sur la législation (notamment la loi de mai 2013 ouvrant le mariage aux couples de personnes de même sexe), et des engagements de la France à fournir des données sur les couples de personnes de même sexe pour le recensement européen de 2021.[1] Les repérer de façon plus fiable dans le recensement permettra d'apporter une réponse de qualité à l'institut européen de statistiques. En tirant profit du recensement, et donc d'une collecte annuelle d'information auprès de plusieurs millions de personnes et de logements (Godinot, 2016), cette amélioration rendra aussi possible la réalisation d'analyses nouvelles sur cette population, ses caractéristiques démographiques, familiales et socio-professionnelles. Certes au recensement, l'analyse se limite aux unions cohabitantes, alors que les relations conjugales entre personnes de même sexe sont plus souvent « à distance » que les unions entre personnes de sexe différent (Toulemon et al., 2005 ; Rault, 2018). Toutefois l'intérêt reste important car les unions cohabitantes sont la forme d'union très largement majoritaire même parmi les CMS (84 pour cent en 2011, Buisson et al., 2013).

4. Cela justifie donc la mise en œuvre d'une solution permettant de distinguer au sein des couples apparemment de même sexe ceux qui le sont réellement et ceux qui sont comptés comme tels suite à une erreur dans le codage du sexe. Pour ce faire, il est envisagé d'ajouter dans les chaînes de traitements du recensement une nouvelle variable individuelle calculée, indiquant dans quelle proportion le prénom déclaré est plutôt masculin ou féminin. Cette variable serait ensuite utilisée pour redresser la variable de sexe pour les personnes qui, d'après les données du recensement, vivent au sein d'un couple de personnes du même sexe. La mise en œuvre de la procédure proposée est envisagée dans la chaîne de traitement au plus tôt pour l'enquête annuelle de recensement de 2020. Une mise en œuvre expérimentale, en dehors des traitements standards, est néanmoins prévue pour les enquêtes annuelles de recensement 2017 à 2019.

5. Après avoir présenté les expériences et solutions mises en œuvre à l'étranger et le recensement français, nous exposons la procédure proposée et une première application

II. Les solutions mises en œuvre à l'étranger

6. La difficulté de mesure des couples de même sexe n'est pas spécifique à la France et différentes solutions ont été mises en place dans d'autres pays. Banens et Penven (2016) présentent ainsi des estimations dans les recensements américains, canadiens et britanniques, de la proportion de « faux » couples de même sexe dans le total des couples. Elle s'échelonne de 0,25 à 0,57 pour cent. La part de ces mêmes « faux couples » dans le total des couples apparaissant comme de même sexe est comprise entre 27 et 55 pour cent. Les ordres de grandeur sont donc très similaires à ceux mesurés pour la France, avec les mêmes difficultés : peu d'erreurs sur l'ensemble mais avec des conséquences très dommageables pour estimer l'effectif de CMS. Les pays confrontés à cette difficulté ont expérimenté différentes stratégies pour la contourner.

A. Redondance et recoupement d'informations

7. Un premier ensemble de solutions sont celles qui consistent à modifier le questionnaire ou le protocole d'une enquête ou d'un recensement, afin d'avoir des informations supplémentaires de validation. Le principe général est de s'appuyer sur le fait que les erreurs sont rares et la probabilité qu'il y en ait deux qui se cumulent est très faible.

8. Dans le recensement canadien depuis 2001, comme dans le recensement américain à compter de 2020, parmi tous les items qui permettent de qualifier la relation entre deux personnes, quatre décrivent une relation conjugale : Époux ou épouse de sexe opposé, Partenaire en union libre de sexe opposé, Époux ou épouse de même sexe, Partenaire en

[1] Règlement d'application n°1201/2009 de la commission européenne.

union libre de même sexe. Il est alors possible de recouper cette information avec le sexe des deux conjoints. Si deux conjoints de même sexe ont choisi l'item « époux de sexe opposé », il est probable qu'il y ait une erreur de codification du sexe de l'un des deux conjoints ; à l'inverse, s'ils ont choisi l'item « époux de même sexe », la probabilité de deux erreurs est très faible et il s'agit d'un « vrai » CMS. Pour les répondants par internet au recensement américain de 2020, il est prévu de surcroît un contrôle (une fenêtre) lorsqu'il y a incohérence entre le choix de l'item et les sexes déclarés (une femme se déclare épouse de sexe opposé d'une autre femme par exemple). Les tests préalables montrent que la réduction des incohérences est très appréciable avec la nouvelle question et les vérifications automatiques en cas de réponse sur internet (Kreider, 2017).

9. Cette démarche de recoupement d'informations collectées de différentes manières est très similaire à celle adoptée pour l'enquête Familles et Logements de 2011.

B. Appariement à des données administratives

10. Une expérience de validation du sexe déclaré et codé par appariement à des données administratives a été mise en œuvre aux États-Unis (Kreider, 2015). En appariant les données du recensement avec le registre de la sécurité sociale, les auteurs relèvent des incohérences bien plus fréquentes pour au moins un des conjoints entre le sexe codé au recensement et celui du registre de sécurité sociale (Numident) lorsque les couples sont apparemment de même sexe au recensement : 72,7 pour cent pour les couples mariés et 6,4 pour cent pour ceux non mariés.

C. « Validation statistique » par le prénom

11. Lors de l'exploitation du recensement de 2010, le Census Bureau américain a utilisé un index des prénoms (O'Connell 2011). Cet index était construit à partir des réponses au recensement lui-même et indiquait la proportion d'hommes portant le prénom (« maleness »), entre 0 et 1 000. Un seuil de 50 pour 1 000, soit 5 pour cent, a été retenu par les auteurs pour effectuer les corrections, seuil qu'ils jugeaient « conservateur ». Autrement dit, si d'après l'index le prénom porté par un enquêté codé par exemple comme masculin était porté par seulement 5 pour cent d'hommes, ou moins, alors le sexe était corrigé. Sinon il était conservé. Ce seuil conduisait à corriger des incohérences entre le prénom et le sexe pour au moins un des conjoints dans 50 pour cent des couples recensés comme de même sexe, plus fréquemment s'ils étaient mariés (69 pour cent) qu'en union libre (21 pour cent).

12. Les résultats obtenus ont pu être confrontés ultérieurement avec les registres de sécurité sociale, afin de vérifier si les corrections faites correspondaient vraiment à des erreurs de codage du sexe (Kreider, 2015). Sur ce test, 85 pour cent des personnes avaient un prénom considéré comme non ambigu et pouvaient donc faire l'objet d'une correction. Dans 96 pour cent des cas, le sexe assigné sur la base du prénom était identique à celui figurant sur le registre de sécurité sociale.

13. La méthode de validation statistique par le prénom semble donc suffisamment fiable, du moins dans son application au recensement américain de 2010. Ces résultats invitent à tester les possibilités d'utiliser cette méthode sur le recensement français.

III. Le recensement français et ses évolutions récentes

14. Depuis 2004, le recensement français est une enquête sur un échantillon. La collecte est annuelle et consiste à remplir un questionnaire sur papier ou sur internet. Pour publier les résultats sur les populations légales commune par commune, les données de cinq collectes annuelles sont utilisées. On parle par exemple de résultats du recensement de la population 2014 pour les données utilisant les collectes annuelles 2012 à 2016. L'enquête annuelle de recensement (EAR) 2017 désigne en revanche la collecte annuelle 2017, c'est-à-dire l'ensemble des personnes et des logements recensés cette année-là. Ce sont les enquêtes annuelles prises individuellement que nous utilisons par la suite.

15. Pour chaque logement recensé, l'agent recenseur doit collecter une feuille de logement, qui décrit notamment les liens entre les habitants du logement, et un bulletin individuel pour chacun d'eux, décrivant leurs caractéristiques socio-démographiques. Une refonte du bulletin individuel a eu lieu en 2015 : la question sur l'état matrimonial légal (marié, divorcé, veuf, célibataire) a été remplacée par une question sur les situations de fait (marié, pacsé, en union libre, divorcé, veuf, célibataire). Cela améliore la qualité des réponses car les personnes s'y retrouvent mieux, notamment celles ayant signé un PACS qui se déclaraient auparavant parfois mariées, considérant cette modalité plus proche de leur mode de vie réel (Buisson, 2017).

16. Une refonte de la feuille de logement a eu lieu à partir de la collecte 2018. Elle permettra de mieux établir la relation conjugale entre deux personnes à partir de ce qu'elles ont déclaré et non plus en la déduisant du fait que chacune a déclaré vivre en couple (sans préciser avec qui). Cette refonte crée indirectement les conditions d'une amélioration de la mesure des CMS. En effet, sa mise en œuvre nécessite un appariement systématique entre d'une part les individus déclarés sur la liste des habitants du logement avec leurs liens deux à deux (feuille de logement) et d'autre part les bulletins individuels de chacun des habitants. Cet appariement est réalisé sur le critère du sexe, de l'année de naissance, et si ces deux premières variables sont insuffisantes pour réaliser l'appariement, du nom et du prénom. L'ensemble des prénoms et noms seront donc exploités au cours des traitements (puis supprimés dans les fichiers de diffusion) et la nouvelle saisie a été organisée par anticipation dès la collecte 2016, permettant la procédure proposée dans la suite du document.

17. Une des principales évolutions récentes du recensement est la collecte par internet. Quasiment inexistante en 2013, elle concernait en 2017 plus de la moitié des individus recensés. Pour eux, les erreurs de codification des réponses liées à la reconnaissance optique et aux corrections manuelles des bulletins papier disparaissent. C'est surtout pour les prénoms que la différence de qualité entre collectes internet et papier importe ici. D'autant que, afin de limiter les coûts de façon proportionnée à l'usage qui est fait de ces prénoms, les critères de qualité concernant la saisie des prénoms collectés sur papier sont peu élevés. Cette différence s'est avérée assez cruciale, nécessitant d'adapter le traitement proposé.

18. En dehors de cet aspect méthodologique, la collecte par internet peut paraître présenter plus de garanties de confidentialité et améliorer la sincérité des déclarations, notamment au sein des couples de même sexe. La généralisation du recueil par voie électronique pour le recensement américain de 2020 est d'ailleurs considérée comme une des voies d'amélioration de la mesure des couples de même sexe par le Census bureau (Kreider, 2017).

IV. La recherche d'une solution optimale

19. Pour repérer les erreurs de codage sur le sexe et ainsi distinguer les faux CMS au sein des CMS apparents, l'exemple américain suggère qu'il est assez efficace d'utiliser le prénom. Utilisant le fait que le prénom soit saisi pour mener des contrôles de qualité de la collecte du recensement depuis 2016 nous testons l'efficacité d'une telle « validation statistique » par le prénom.

A. Un indicateur transitoire : les couples apparemment de même sexe

20. Pour estimer les effectifs de couples apparemment de même sexe et leur évolution, et pouvoir tester dans quelle proportion certains seraient considérés comme de vrais couples de même sexe et d'autres catégorisés en erreurs de codage sur le sexe, on utilise un indicateur simplifié : si exactement deux personnes disent vivre en couple dans le logement et qu'elles sont de même sexe, alors elles sont classées en couple apparemment de même sexe. Cet indicateur est imparfait : deux personnes vivant ensemble et déclarant toutes deux « vivre en couple » parce qu'elles ont chacune un conjoint vivant dans un autre logement, seront comptées à tort comme formant un couple ensemble. Mais la comparaison de cet

indicateur simplifié et de la mesure validée sur les données de l'enquête Famille et Logements de 2011 (EFL) montre que l'indicateur est correct pour estimer les CMS apparents. Comme attendu, il englobe trop de couples, dont une bonne partie sont des couples de sexe opposé comptés comme CMS suite à une erreur de codage. Mais en revanche peu de « vrais » CMS sont omis. La refonte de la feuille de logement permettra de s'appuyer sur une meilleure mesure en 2018.

B. L'échantillon démographique permanent : un outil idéal pour tester la capacité de la procédure à repérer des erreurs avérées de codage du sexe

21. L'EDP est un panel sociodémographique de grande taille mis en place en France, pour étudier les parcours démographiques, familiaux, professionnels et géographiques (Durier, 2018). Le principe général consiste à conserver pour les individus appartenant à l'échantillon (environ 4 pour cent de la population) des informations collectées dans les cinq sources statistiques qui alimentent l'EDP (bulletins d'état-civil, recensements, fichier électoral, panel « tous salariés » sur les rémunérations perçues et, depuis 2011, données fiscales et sociales sur les revenus).

22. Le premier intérêt de l'EDP dans notre approche est de permettre la confrontation du « vrai sexe », celui enregistré au répertoire national d'identification des personnes physiques (RNIPP), réputé exact car utilisé pour gérer les numéros de sécurité sociale, avec celui déclaré dans les enquêtes annuelles de recensement. Cela permet d'estimer la proportion d'erreurs de codage sur le sexe au recensement. Dans la base études 2016, sur 1,3 millions de personnes « EDP » (nées un jour EDP), recensées au moins une fois entre 2010 et 2016 et vivant en couple, le taux d'erreur sur le sexe est de 0,17 pour cent. C'est donc un phénomène très rare. Le taux d'erreur est en revanche considérablement plus élevé (18 pour cent) pour les personnes apparemment en CMS.

23. Dans l'EDP, le RNIPP n'est interrogé que pour la personne « EDP » afin de l'inclure dans l'échantillon et compléter les données statistiques la concernant. Les informations statistiques sur les habitants de son logement sont aussi incluses, mais sans identification – interrogation du RNIPP. On ne peut donc pas certifier le sexe des autres habitants du logement, conjoints compris. Comme 0,17 pour cent des personnes EDP en couple sont concernées par une erreur de codage, on peut estimer, en supposant l'indépendance des erreurs entre deux conjoints, que 0,31 pour cent des couples seraient affectés par une erreur de codage du sexe de l'un des conjoints, qui conduit à les compter par erreur comme CMS, et 0,03pour cent par deux erreurs de codage (Tableau 1).

Tableau 1
Conséquence anticipée des erreurs de codage du sexe au niveau des couples

Situation réelle		*Situation apparente*		
H1 : 0,6 % de « vrais CMS parmi les couples		*H2 : Le sexe codé est erroné pour 0,17 % des personnes recensées*		
		Couples avec une erreur (0,31 %)	*Couples avec deux erreurs (0,03 %)*	*Couples sans erreur (99,66 %)*
Sur 1 000 000 de couples	6 000 CMS	➔ 19 CSD apparents	2 CMS	5 980 CMS
	994 000 CSD	3 078 CMS apparents	287 CSD	990 634 CSD

Source : Enquête Famille et Logements 2011, base étude 2016 de l'EDP, Insee.
Champ : Couples cohabitants.
Lecture : Selon l'Enquête Famille et Logements, 0,6 % des couples sont des « vrais » CMS, 99,4% des « vrais » CSD (H1). Avec une erreur de codage affectant 0,17 % des personnes (H2), recensées, 0,03% des couples sont affectés par deux erreurs de codage (0,17 x 0,17) tandis que 0,31% des couples sont affectés par une erreur sur un des conjoints (0,17 + 0,17 – 0,03). Ainsi, 0,31 % des CSD passent à une situation apparente de CMS, soit 3 078 pour un million de couples. Inversement 0,31% des CMS passent à une situation apparente de CSD soit 19. Enfin, pour 0,03% ont les deux membres du couple affectés d'une erreur de codage sur le sexe (soit 2 CMS et 287 CSD). Même si aucun des membres du couple n'a un sexe correspondant à la réalité, la situation apparente de leur couple correspond à la réalité : le couple reste un CSD. Avec 0,17% d'erreurs, on peut donc attendre que les CMS passent d'une situation

réelle où ils représentent 0,6 % des couples à une situation apparente dans l'EAR où ils représentent 0,9 % des couples (5980 + 3078 + 2 = 9060 sur un million).

C. Le dictionnaire choisi

24. Second intérêt de l'EDP, le prénom déclaré au recensement est dans la base de production afin de faciliter l'appariement (il est en revanche absent de la base études utilisée pour établir des statistiques, afin d'éviter une identification directe des personnes). Il est donc possible de construire dictionnaire et indicateur d'erreur de la même façon que dans la chaîne de production des futures enquêtes annuelles de recensement.

25. Un dictionnaire de prénoms associe à chaque prénom la proportion de femmes (respectivement d'hommes) le portant. Il est ensuite apparié aux prénoms des personnes recensées afin de comparer le sexe codé pour chaque recensé au sexe le plus fréquemment associé à son prénom. La finalité est de repérer les erreurs de codage les plus probables. L'EDP nous a permis de comparer les performances de différents dictionnaires pour choisir celui qui est le plus efficace pour repérer les erreurs de codage du sexe concernant les personnes EDP.

26. Une combinaison des différents dictionnaires testés a été retenue, utilisant deux sources. La source privilégiée, exhaustive pour les personnes nées en France, est le fichier des prénoms donnés à l'état civil depuis 1900, par sexe[2]. Il faut un très grand nombre d'observations pour pouvoir calculer une proportion d'hommes et de femmes parmi les porteurs de chaque prénom. Ce fichier a été complété en ajoutant les occurrences des prénoms des personnes recensées en 2017[3] et nées à l'étranger, puisqu'elles ne sont pas couvertes par l'état-civil. Elles portent plus fréquemment un prénom absent du dictionnaire construit avec l'état-civil. Le dictionnaire retenu est aussi une combinaison au sens où on cherche d'abord une correspondance dans un dictionnaire le plus détaillé possible (même prénom, même année de naissance). En cas d'échec, un dictionnaire moins détaillé est utilisé : même première partie du prénom, sans condition sur l'année de naissance. La proportion finalement affectée à une personne peut donc être la proportion de femmes parmi les personnes nées la même année et portant le même prénom ou la proportion de femmes parmi toutes les personnes portant un prénom dont la première partie est identique. Par simplicité on parlera pour désigner cet indicateur de la proportion de femmes portant le même prénom. On déduit ensuite de cette proportion un indicateur d'erreur : si le sexe codé au recensement est masculin, il s'agit de la proportion de femmes portant le même prénom d'après le dictionnaire. Si le sexe codé est féminin, alors il s'agit de la proportion d'hommes. Par convention, en l'absence de correspondance et donc de proportion l'indicateur est fixé à 0 : aucune correction ne peut être faite. Plus la valeur est élevée plus il y a suspicion d'erreur de codage.

D. Sur les personnes EDP, le repérage des erreurs de codage du sexe par les prénoms est très efficace

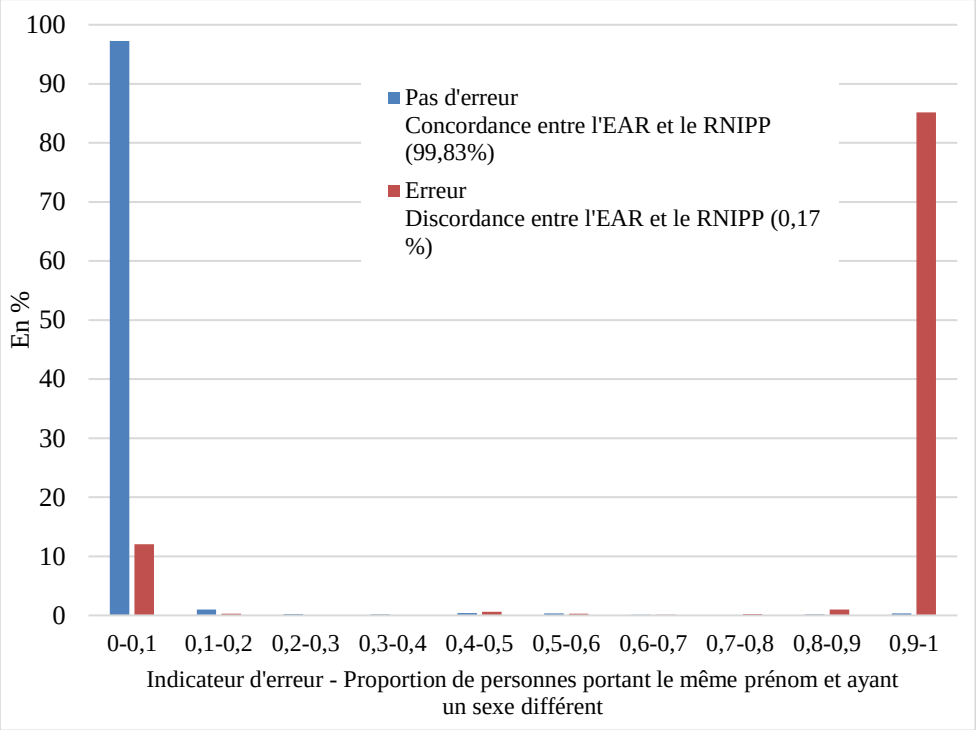
27. Cet indicateur est très efficace pour repérer les erreurs de codage. En l'absence d'erreur avérée presque toutes les valeurs sont nulles ou très proches de 0 (graphique 1). En cas d'erreur avérée (le sexe codé à l'EAR est différent de celui enregistré au RNIPP), elles sont au contraire presque toutes proches de 1. Les valeurs intermédiaires sont rares : peu de personnes portent des prénoms épicènes, comme Dominique.

[2] Les prénoms trop rarement donnés sont retirés.

[3] Ces prénoms disponibles pour les répondants par internet, ont été saisis pour les répondants par questionnaire papier, à des fins de gestion du recensement. Les prénoms ont ensuite été détruits, comme prévu, une fois les traitements réalisés courant 2017.

Graphique I

Répartition de l'indicateur d'erreur selon la présence d'une erreur avérée



Source : Base étude 2016 de l'EDP, Insee, données pondérées, ensemble des personnes en couple.

Lecture : Parmi les personnes de plus de 15 ans recensés entre 2010 et 2016, qui ont déclaré vivre en couple lors de ce recensement, et pour lesquels le sexe déclaré au recensement a été confirmé dans le RNIPP, la quasi totalité porte un prénom cohérent avec le sexe déclaré au recensement : les valeurs de l'indicateur d'erreur construit avec les prénoms sont concentrées au voisinage de 0. Au contraire, pour la grande majorité des personnes pour lesquelles le sexe déclaré au recensement a été invalidé par le RNIPP, l'indicateur construit sur la base du prénom signale la forte probabilité d'erreur sur le sexe déclaré au recensement, avec des valeurs concentrées au voisinage de 1. Les valeurs autour de 0 correspondent pour l'essentiel à des individus en échec d'appariement avec le dictionnaire des prénoms (le prénom porté est trop rare par exemple).

28. Pour déterminer le seuil optimal à partir duquel on décide de corriger, les outils mobilisés sont les mêmes que pour le développement d'un test diagnostique en épidémiologie : mesures de spécificité (bien repérer les erreurs) et de sensibilité (ne pas corriger, à tort, des réponses justes), détermination du seuil optimal et comparaison de tests grâce à la courbe ROC (Robin, 2011). Cela conduit à retenir un seuil de 41 pour cent : pour une personne codée comme masculine, si plus de 41 pour cent des personnes portant le même prénom sont des femmes, une correction est apportée. Il peut paraître étonnant de corriger alors même que la majorité des porteurs du prénom sont des hommes mais il y a en réalité très peu de cas où la valeur du dictionnaire est intermédiaire : la presque totalité des valeurs sont très proches de 0 ou au contraire de 100 pour cent. La fixation du seuil a un impact modéré dans un contexte aussi favorable. Avec le dictionnaire retenu et le seuil de 41 pour cent, cela permet d'atteindre une sensibilité de 98 pour cent et une spécificité de 95 pour cent. En se limitant aux CMS apparents, les corrections à tort seraient environ 150 dans l'EDP, pour près de 2 000 vraies erreurs repérées.

Tableau 2
Performances du dictionnaire pour repérer les erreurs de codage sur le sexe

	Ensemble des personnes en couple	*Personnes en CMS apparent*		
		Ensemble	*Collecte papier*	*Collecte internet*
Effectif	1 343 908	11 349	9 605	1 744
Erreurs avérées	2 336	2 059	1 852	207
Meilleur seuil (en %)	41	41	41	51
Spécificité (en %)	99	98	98	99
Sensibilité (en %)	89	95	95	95
Corrections abusives	16 950	147	134	9
Corrections valides	2 088	1 951	1 754	197

Source : Base étude 2016 de l'EDP, Insee, données non pondérées.
Champ : Personnes de plus de 15 ans vivant en couple, recensées entre 2010 et 2016.

V. Les résultats inattendus de la transposition de la procédure de l'EDP vers l'EAR, et la nécessité d'adapter la procédure selon le mode de collecte

A. Une transposition directe au recensement aboutit à des corrections excessives

29. L'utilisation de l'EDP a donc permis d'évaluer la procédure de correction sur un échantillon dans lequel les erreurs de codage sont connues. Comme les données de l'EDP sont extraites du recensement, cela semblerait préfigurer idéalement les performances de la correction une fois appliquée à l'ensemble du recensement. Malheureusement, la transposition directe aboutissait à des résultats inattendus : 12 pour cent des enquêtés de la collecte papier de l'EAR 2017 (quelle que soit leur situation de couple) portaient un prénom presque toujours attribué à l'autre sexe (Tableau 3), proportion bien trop élevée comparée aux 0,2 pour cent d'erreurs de codage sur le sexe qu'on cherche à repérer. Or, il est apparu que les erreurs portaient davantage sur la saisie du prénom que sur le sexe : pour des raisons de coût, le niveau de qualité exigé du prestataire lors de la saisie du prénom pour les personnes recensées sur papier est faible, exception faite des bulletins destinés à être intégrés dans l'EDP.

Tableau 3
Comparaison du dictionnaire appliqué à l'EDP et à l'EAR 2017 selon le mode de collecte

	Ensemble des couples					*CMS*		
		EAR 2017 – avant prise en compte du mode de collecte		*EAR 2017 – après prise en compte du mode de collecte*			*EAR 2017 – après prise en compte du mode de collecte*	
Les porteurs du prénom…	*EDP*	*papier*	*internet*	*papier*	*internet*	*EDP*	*papier*	*internet*
…sont presque toujours du même sexe que l'enquêté (IE<5%)	93	72	95	41	97	79	32	84
…sont majoritairement du même sexe que l'enquêté (IE compris entre 5 et 40%)	3	8	3	36	2	2	32	1
…sont des personnes des deux sexes (IE compris entre 40 et 90%)	1	7	1	21	1	1	26	1
…sont très rarement du même sexe que l'enquêté (IE compris entre 90 et 95%)	0	1	0	1	0	0	3	0
…ne sont presque jamais du même sexe que l'enquêté (IE>95%)	0	12	2	1	0	18	8	13
Total	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

IE : indicateur d'erreur.

Lecture : 32 pour cent des personnes en CMS apparent enquêtées sur papier portent (d'après la saisie de leur bulletin individuel de recensement) un prénom qui est presque toujours porté par des personnes du même sexe qu'elles : le dictionnaire indique une proportion inférieure à 5 pour cent de porteurs du sexe opposé.

Champ : Personnes en CMS apparent, pour l'EAR 2017, hors FLNE et hors ménages où le sexe d'un des deux conjoints a été imputé.

Sources : Base étude 2016 de l'EDP, EAR 2017, Insee.

B. Des dictionnaires différenciés selon le mode de collecte

30. Pour contourner la difficulté, la solution retenue a été d'intégrer le mode de collecte dans la construction des dictionnaires. Ceci fait, la proportion de situations où plus de 95 pour cent portent un sexe différent de celui déclaré par l'enquêté devient très faible même pour la collecte papier. Le revers est la proportion élevée de valeurs intermédiaires, qui nécessite une prise en compte spécifique dans les traitements. Comme la qualité est meilleure dans l'échantillon démographique permanent, même en cas de collecte papier, il n'est pas possible de tester sur l'EDP cette prise en compte de la moindre qualité de la collecte papier du recensement.

31. S'agissant de la collecte internet en revanche, les résultats sont bien plus proches de l'EDP. La qualité de repérage des erreurs de codage sur le sexe dans l'EDP peut donc être considérée comme représentative de la qualité du traitement sur le versant internet de l'EAR. Il est donc proposé d'utiliser le dictionnaire adapté à internet et de retenir un seuil de 51 pour cent, seuil maximisant sur la collecte internet de l'EDP la sensibilité (à 95 pour cent, donc 5 pour cent des cas sans erreur sont corrigés à tort) – et la spécificité (à 99 pour cent : 99 pour cent des erreurs avérées sont repérées). Pour une personne en CMS apparent, une correction sera apportée au sexe déclaré dès que la proportion de personnes portant le même prénom ayant un sexe opposé au sien est supérieure ou égale à 51 pour cent:

C. Pour la collecte papier, la correspondance entre sexe et prénom est plus incertaine

32. Pour la collecte papier de l'EAR, le dictionnaire a été construit de façon à éviter des corrections abusives. Néanmoins, l'indicateur d'erreur prend fréquemment des valeurs intermédiaires : le critère du prénom fonctionne moins bien et le résultat est très sensible au seuil retenu contrairement à la collecte internet. Avec un seuil de 95 pour cent par exemple, 8 pour cent des individus en CMS apparents collectés sur papier seraient corrigés, 13 pour

cent sur internet. Avec un seuil de 40 pour cent, les taux de correction passeraient à 37 pour cent sur papier, et à 14 pour cent, presque inchangés, sur internet. Pour la collecte papier, les valeurs intermédiaires indiquent que le prénom est sujet à des saisies erronées (certains prénoms le sont davantage que d'autres). On s'appuie alors sur d'autres variables pour estimer si la personne est probablement en « vrai » CMS ou non. Ces variables sont choisies grâce à l'EDP qui permet de mesurer la proportion d'erreurs effectives sur le sexe pour les personnes en CMS apparent recensées sur papier.

33. Finalement, si la proportion de personnes ayant un sexe opposé à l'enquêté parmi celles portant le même prénom est supérieure à 90 pour cent, on corrige le sexe ; si elle est inférieure à 20 pour cent, on ne corrige pas. Si elle se situe entre 20 pour cent et 90 pour cent, le prénom est considéré comme ambigu et des variables complémentaires sont utilisées. Pour l'EAR 2017, trois groupes ont été retenus : les personnes non mariées, pour lesquelles les taux d'erreur sont les plus faibles (10 pour cent) ; les personnes mariées et âgées de moins de 50 ans (23 pour cent d'erreurs) ; les personnes mariées et âgées de 50 ans ou plus (32 pour cent). À partir de l'EAR 2018, tirant profit de la refonte de la feuille de logement, les groupes sont constitués en prenant en compte le type de famille (couple sans enfant, famille avec enfant, recomposée ou non) et le traitement est amélioré en regardant la situation du couple (et non les deux conjoints séparément).

VI. Application : combien de CMS en France, et quelle évolution ?

34. Il est possible d'estimer les CMS apparents dans les EAR depuis 2005 et dans l'EDP depuis 2010. Ils sont en croissance régulière. La rupture de série en 2018 s'explique par une diminution forte des erreurs de codage sur le sexe suite à la refonte de la feuille de logement.

35. Pour les « vrais » CMS, les différentes estimations (procédure des prénoms pour les EAR 2017 et 2018, extrapolation des erreurs individuelles aux couples pour les années 2010 à 2018 dans l'EDP) semblent cohérentes entre elles et avec la donnée de l'EFL. La proportion de personnes en CMS parmi celles en couple cohabitant semble avoir augmenté régulièrement depuis 2010, passant de 0,5 pour cent à 0,8 pour cent. Avec une meilleure caractérisation de la situation familiale, les proportions estimées dans l'EFL 2011 et l'EAR 2018 sont légèrement plus élevées mais signalent une croissance, de 0,6 pour cent en 2011 à 0,9 pour cent en 2018.

Tableau 4
Proportion de CMS apparents et estimation des « vrais » CMS dans différentes sources de 2005 à 2018

	CMS apparents			*Estimation « Vrais »CMS*		
%	*EAR*	*EDP*	*EFL*	*EAR*	*EDP*	*EFL*
2005	0,6					
2006	0,7					
2007	0,7					
2008	0,7					
2009	0,8					
2010	0,8	0,8			0,5	
2011	0,8	0,8	0,9		0,5	0,6
2012	0,9	0,9			0,6	
2013	0,9	0,9			0,6	
2014	1,0	1,0			0,6	
2015	1,0	1,0			0,7	
2016	1,1	1,1			0,7	
2017	1,2	1,2		0,8	0,8	
2018	1,0 / 1,0*	1,0		0,8 / 0,9*	0,8	

Source : EAR 2005 à 2018, Base études 2016 de l'EDP et données complémentaires 2017 et 2018, Insee.

Champ : Personnes vivant en couple. Pour les EAR, hors FLNE et réponse manquante pour le sexe d'un des conjoints

* Appuyé sur de nouvelles informations suite à la refonte de la feuille de *logement :* le champ est celui des personnes vivant en couple cohabitant d'après la nouvelle analyse ménage famille. Les feuilles de logement non enquêté sont réintégrées, ainsi que les réponses manquantes sur le sexe (devenues inexistantes sur papier et très rares sur internet).

VII. Conclusion

36. Au terme de cette analyse et en dépit de quelques limites, l'utilisation de la méthode s'appuyant sur les prénoms dans le cas français donne des résultats plutôt satisfaisants. La première étude en cours sur les caractéristiques des personnes vivant en couple avec une personne du même sexe en 2018, réalisée grâce à cette méthode permettra d'affiner le diagnostic. L'objectif est ensuite de l'intégrer dans les chaînes de traitement du recensement afin que la qualité des variables de sexe figurant dans les fichiers de diffusion puisse être améliorée.

Références

Algava E., Hallepee S., 2018, « Estimer les effectifs de couples de personnes de même sexe au recensement : expérimentation d'une solution de validation du sexe par le prénom », document de travail F1807, Insee

Banens M., Le Penven E., 2016, « Les erreurs de sexe dans le recensement et leurs effets sur l'estimation des couples de même sexe », *Population* 71(1), p. 135–148.

Bodier M. et al. (coord.), 2015, *Couples et familles*, Paris, Insee Références, 190 p.

Breuil-Genier P., Buisson G., Robert-Bobée I., Trabut L., 2016, « Enquête *Famille et Logements* adossée au Recensement de 2011 : comment s'adapter à la nouvelle méthodologie des enquêtes annuelles et quels apports ? », *Économie et statistique*, n°483–485, p. 205–226.

Buisson G., Lapinte A., Le couple dans tous ses états. Non-cohabitation, conjoints de même sexe, pacs... », Insee Première n°1432, Insee, 2013.

Buisson G., La situation matrimoniale dans le recensement : impact de la refonte du questionnaire de 2015, document de travail F1707, Insee, 2017.

Durier S., 2018, *L'échantillon démographique permanent a 50 ans : retours sur un dispositif statistique original*, Présentation aux Journées de méthodologie statistique, Paris, juin.

Godinot A., Durr J.-M., 2016, La rénovation du *Recensement de la population*. In: *Économie et statistique*, n°483–485, p. 7–14.

Imbert C., Lelievre E., Lessault D. (dir.), 2018, *La famille à distance : mobilités, territoires et liens familiaux*, Ined, collections Questions de population, 376 p.

Kreider R., Bates N., Yeris M.-G., 2017, Improving Measurement of Same-Sex Couple Households in Census Bureau Surveys: Results from Recent Tests, SEHSD Working Paper 2017-28.

Kreider R., Lofquist.D, 2014, "Matching Survey Data with Administrative Records to Evaluate Reports of Same-Sex Married Couple Households." SEHSD Working Paper, 2014-36.

Lathe H., Ménard F.-P., Martel L., Hallman S., "Les couples de même sexe au Canada en 2016", Le recensement en bref, Statistiques Canada, No 98-200-X2016007, 2017.

O'Connelll M., Feliz S., 2011, Same-Sex Couple Household Statistics from the 2010 Census**,** SEHSD Working Paper, 2011-26.

Rault W., 2016, « Les mobilités sociales et géographiques des gays et des lesbiennes. Une approche à partir des femmes et des hommes en couple », Sociologie, 7(4), p. 337–360.

Rault W., 2018, « La distance, une composante plus fréquente des relations conjugales et familiales des gays et des lesbiennes ? » in Imbert Christophe, Lelievre Eva, Lessault David (dir.), La famille à distance, Paris, Ined, Questions de populations n° 2.

Rault W., "Secteurs d'activités et professions des gays et des lesbiennes en couple : des positions moins genrées", *Population*, 2017/3 (Vol. 72), p. 399–434.

Toulemon L., Vitrac J., Cassan F., 2005, « Le difficile comptage des couples homosexuels d'après l'enquête EHF », in Lefevre Cécile, Fillon Alexandra (dir.), Histoires de familles, histoires familiales. Les résultats de l'enquête Famille de 1999, Ined, Cahier n° 156, p. 589–602.