ESTIMER LES EFFECTIFS DE COUPLES DE PERSONNES DE MÊME SEXE AU RECENSEMENT : EXPÉRIMENTATION D'UNE SOLUTION DE VALIDATION DU SEXE PAR LE PRÉNOM

Élisabeth ALGAVA, Sébastien HALLÉPÉE

Insee, Direction des Statistiques Démographiques et Sociales

elisabeth.algava@insee.fr
sebastien.hallepee@insee.fr

***Mots-clés** : Recensement, couple, contrôle de cohérence, imputation*

Résumé

Il est actuellement impossible d'établir à partir du recensement des statistiques fiables concernant le nombre de couples de personnes de même sexe en France. En effet, une proportion importante des couples de même sexe sont comptés comme tels du fait d'une erreur de codage sur le sexe d'un des conjoints. Cela conduit à surestimer le nombre de couples de personnes de même sexe. C'est ce qu'avait montré un important travail de redressement mené après la collecte de l'enquête Famille et logements en 2011 : d'une mesure non corrigée dans l'enquête annuelle de recensement (EAR) 2011 de 295 000 personnes en couple avec une personne du même sexe, on passe après redressement à 173 000 personnes en couples co-résidents. Plus de 40 % des situations ont donc été corrigées.

Or, dans le cadre des règlements européens sur le recensement, la France est engagée à fournir des données sur les couples de personnes de même sexe pour le recensement 2021. Repérer de façon plus fiable les couples de même sexe co-résidents 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, relativement rare actuellement (0,6 % des personnes en couples co-résidents en 2011). Il sera notamment possible de réaliser des études plus fines sur leurs caractéristiques démographiques, familiales et socio-professionnelles.

Cela justifie 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 dans le présent document est en cours de spécification, 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.

Dans un premier temps, nous présentons le recensement, la façon dont il est collecté et les obstacles rencontrés dans la construction d'une mesure fiable de la proportion de personnes en couples avec une personne du même sexe. La seconde partie est consacrée aux expériences et solutions mises en œuvre à l'étranger, notamment au Canada et aux États-Unis. Elles montrent que la validation par les prénoms fonctionne correctement même s'il existe d'autres façons d'améliorer la qualité de la

mesure des couples de même sexe, par les modifications apportées au questionnaire ou l'appariement avec des données administratives par exemple. Ces solutions sont plus directes et efficaces, mais nettement plus difficiles à mettre en œuvre (coût, délai, sécurisation des données).

La partie suivante décrit les choix opérés pour construire un dictionnaire et l'appliquer aux enquêtes annuelles de recensement. Elle commence par montrer l'apport de l'échantillon démographique permanent dans la validation de la procédure : il permet de tester la capacité de la méthode à repérer les erreurs de codage du sexe sur un échantillon pour lequel elles sont connues. Certaines différences dans les traitements post-collecte entre le recensement et l'échantillon extrait pour l'échantillon démographique permanent rendent cependant nécessaire d'adapter le traitement avant son application au recensement. Il s'agit notamment de traiter séparément la collecte papier et la collecte internet.

Enfin, la dernière partie présente les résultats obtenus en appliquant la solution retenue à l'échantillon démographique permanent et à l'EAR 2017. Cela permet de vérifier la cohérence de ces résultats, autant en termes d'effectifs que d'évolution.

Abstract

In France, same-sex couples are overestimated when using Census data. This is due to a large amount of opposite-sex couples being wrongly considered as same-sex couples following an error in the reported sex of one of the partners. In order to identify those couples who likely are same-sex couples compared to those who are most likely opposite-sex couples who mismarked the sex item for one of the partners, we propose a solution using first names. An index associates to a particular first name the proportion of reported females among the holders of that name. It thus indicates whether an error in the reported sex is likeky or not. We intend to demonstrate that this solution, when implemented in the French Census, should permit to produce high-quality estimates and new studies on same-sex couples.

Il est actuellement impossible d'établir à partir du recensement 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 est compté comme tel du fait d'une erreur de codage sur le sexe d'un des conjoints. Cela conduit à surestimer le nombre de couples de personnes de même sexe.
Lors de l'édition 1999 de l'enquête Famille, dénommée « Étude de l'histoire familiale », l'estimation du nombre de couples de même sexe cohabitants restait très incertaine, dans une enquête qui n'avait « pas été conçue pour compter les couples homosexuels » (Toulemon et al., 2005). Lors de l'édition suivante, l'enquête Famille et logements de 2011, il en allait autrement : un important travail de vérification des réponses par croisement des informations disponibles dans le recensement et l'enquête Famille Logement a été réalisé (Breuil-Genier et al., 2016). Il a permis d'estimer qu'au recensement de 2011, 0,6 % des couples cohabitants étaient des couples de même sexe et 0,36 % 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 personnes en couples de même sexe : d'une mesure non corrigée dans l'enquête annuelle de recensement (EAR) 2011 de 295 000 personnes en couple avec une personne du même sexe, on passe après redressement à 173 000 personnes en couples co-résidents. Plus de 40 % des situations ont donc été corrigées. Cette étape d'apurement a rendu possible la réalisation d'analyses statistiques nouvelles concernant les personnes en couples de même sexe en 2011 (Rault 2013, Rault 2017).
Les enquêtes Famille ont lieu environ tous les dix ans. La prochaine enquête, envisagée au début des années 2020, permettra en principe d'avoir un nouveau chiffrage du nombre de couple de même sexe. Il semble toutefois difficile de se contenter d'une estimation décennale, 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 à fournir des données au niveau européen.

En effet, dans le cadre des règlements européens sur le recensement[1], la France est engagée à fournir des données sur les couples de personnes de même sexe pour le recensement 2021. Repérer de façon plus fiable les couples de même sexe co-résidents 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, relativement rare actuellement (0,6 % des personnes en couples co-résident en 2011). Il sera notamment possible de réaliser des études plus fines sur leurs caractéristiques démographiques, familiales et socio-professionnelles.
Cela justifie 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 dans le présent document est en cours de spécification, 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.

Dans un premier temps, nous présentons le recensement, la façon dont il est collecté et les obstacles rencontrés dans la construction d'une mesure fiable de la proportion de personnes en couples avec une personne du même sexe. La seconde partie est consacrée aux expériences et solutions mises en œuvre à l'étranger, notamment au Canada et aux États-Unis. Elles montrent que la validation par les prénoms fonctionne correctement même s'il existe d'autres façons d'améliorer la qualité de la mesure des couples de même sexe, par les modifications apportées au questionnaire ou l'appariement avec des données administratives par exemple. Ces solutions sont plus directes et efficaces, mais nettement plus difficiles à mettre en œuvre (coût, délai, sécurisation des données).

La partie suivante décrit les choix opérés pour construire un dictionnaire et l'appliquer aux enquêtes annuelles de recensement. Elle commence par montrer l'apport de l'échantillon démographique permanent dans la validation de la procédure : il permet de tester la capacité de la méthode à repérer les erreurs de codage du sexe sur un échantillon pour lequel elles sont connues. Certaines différences dans les traitements post-collecte entre le recensement et l'échantillon extrait pour l'échantillon démographique permanent rendent cependant nécessaire d'adapter le traitement avant son application au recensement. Il s'agit notamment de traiter séparément la collecte papier et la collecte internet.

Enfin, la dernière partie présente les résultats obtenus en appliquant la solution retenue à l'échantillon démographique permanent et à l'EAR 2017. Cela permet de vérifier la cohérence de ces résultats, autant en termes d'effectifs que d'évolution.

1 Le recensement français et la mesure des couples de même sexe

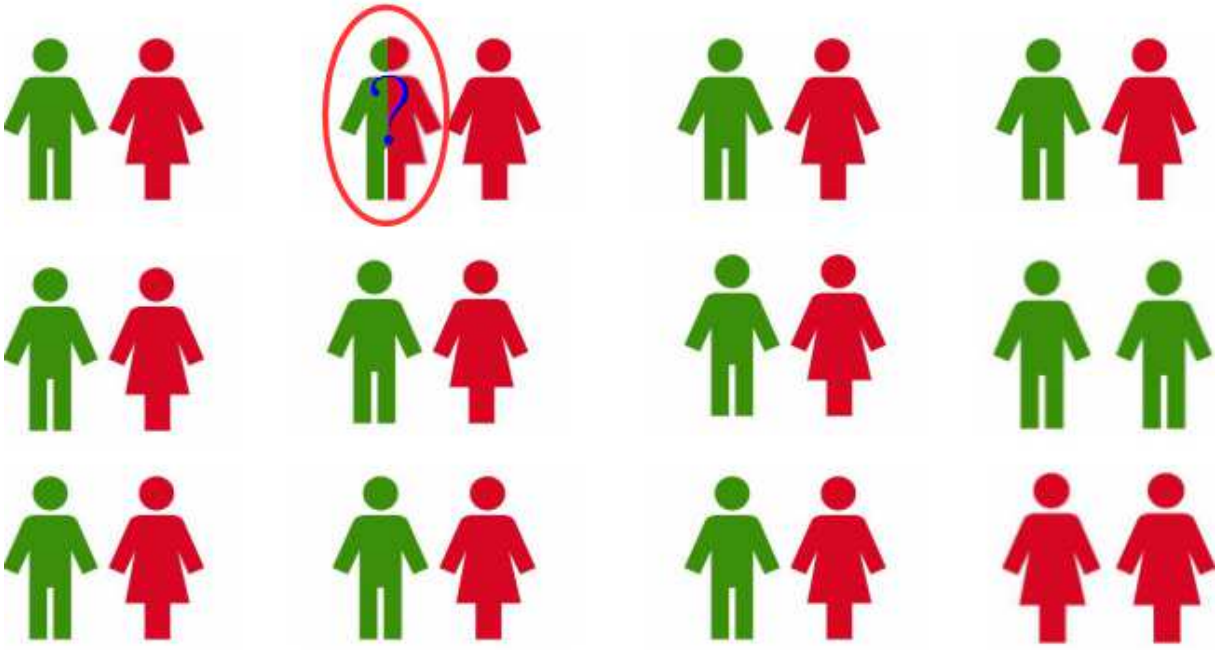
La principale difficulté pour mesurer les couples de personnes de même sexe (CMS[2]) est d'ordre méthodologique. Dans un couple de personnes de sexes différents, une erreur sur le sexe d'un seul des conjoints aboutit en général à compter ce couple comme étant de même sexe. Si cela ne

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

[2]Pour plus de simplicité, on parlera par la suite de couples de même sexe, abrégés en CMS et comparés aux couples de sexe opposé ou différent (CSO).

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 de sur-estimation n'est pas propre aux couples de même sexe. Il est présent dès qu'il s'agit d'estimer des populations rares, c'est-à-dire 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. En revanche, une erreur sur le sexe au niveau individuel, pour un des conjoints, va conduire à considérer les deux conjoints comme ayant un partenaire de même sexe (Schéma 1) : une erreur compte double, ce qui démultiplie les problèmes d'estimation de la part des CMS parmi l'ensemble des couples.

Schéma 1 : 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 opposé (CSO), 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 CSO à CMS. L'erreur compte donc « double ».
Le nombre de CMS augmenterait ainsi artificiellement de 50 % alors que le nombre de CSO ne serait réduit artificiellement que de 10 %. On retrouve le fait que les faibles erreurs de codage ont un impact beaucoup plus visible sur les populations rares.

Cette difficulté n'est donc ni spécifique au recensement ni même à la France : l'ensemble des enquêtes auprès des ménages sont concernées et plusieurs pays ont mis en place des solutions pour parer au problème, comme nous le soulignerons par la suite. Toutefois, afin de comprendre la solution proposée dans le présent document, il est nécessaire de la replacer au préalable dans le contexte du recensement français, de ses spécificités et de ses évolutions.

1.1 Enquête annuelle de recensement et recensement de la population

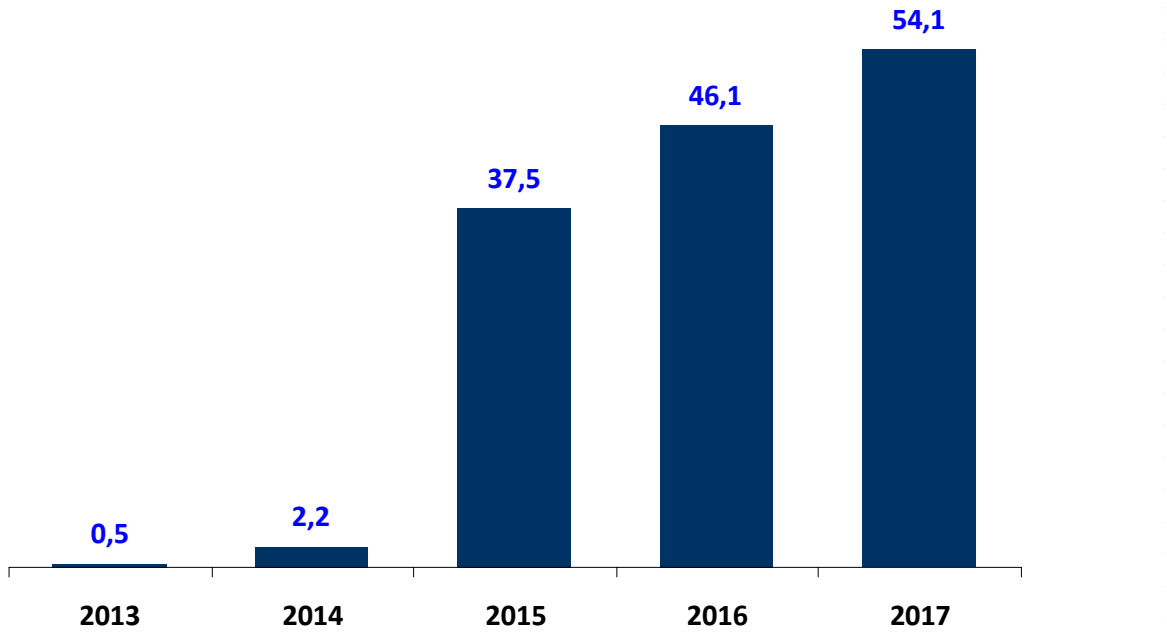
Depuis 2004, le recensement français est une enquête sur un échantillon. La collecte est annuelle. Chaque année, une « petite commune » (moins de 10 000 habitants) sur cinq est recensée de façon exhaustive. Dans les « grandes communes » (à partir de 10 000 habitants), un échantillon d'adresses représentant 8 % des logements environ est recensé. À l'issue de cinq années de collecte, l'ensemble des logements des petites communes ont été recensés et 40 % des logements des grandes

communes. 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. Dans le présent document, nous utilisons les collectes annuelles de façon individuelle. Ainsi, l'enquête annuelle de recensement ou EAR 2017 désigne la collecte annuelle 2017, c'est-à-dire l'ensemble des personnes et des logements recensés cette année-là.
Les résultats obtenus sur une seule collecte sont moins précis mais suffisent largement pour obtenir des estimations, notamment au niveau national. La collecte porte sur environ 8 millions d'individus chaque année et une pondération a été calculée afin d'obtenir des résultats représentatifs de l'ensemble de la population résidant en France. Elle est indispensable pour tenir compte des probabilités de sondage différenciées, principalement entre grandes et petites communes.
Cette utilisation séparée d'une collecte à des fins d'étude ne pose pas de difficultés. En revanche, les données fournies à Eurostat sur les couples de même sexe dans le cadre du règlement d'application n°1201/2009 de la commission européenne porteront sur le recensement 2021, donc sur les collectes annuelles 2019 à 2023. Le processus de correction des erreurs sera mis en œuvre dans les traitements de la collecte annuelle 2020 du recensement mais une procédure expérimentale permettra d'appliquer cette méthode et de l'intégrer a posteriori dans les fichiers de diffusion du RP pour les EAR 2017 à 2019.

1.2 Collecte sur papier et collecte par internet

Une des principales évolutions récentes du recensement est la mise en œuvre de la collecte par internet et sa généralisation rapide. Quasiment inexistante en 2013, elle concernait en 2017 plus de la moitié des individus recensés. Pour les personnes recensées par internet, les erreurs de codification des réponses liées à la reconnaissance optique et aux corrections manuelles disparaissent. Pour cette raison, la qualité des réponses sur internet est estimée un peu meilleure comparée à celle des réponses papier. Par ailleurs, ce mode de collecte peut paraître présenter plus de garanties de confidentialité pour les enquêtés : leur réponse n'est pas remise à l'agent recenseur, même si bien entendu celui-ci est tenu de ne pas divulguer les informations sur les personnes recensées. Cela peut contribuer à 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).

Graphique 1 : Part de personnes recensées par internet selon l'année de collecte



Champ : Ménages ordinaires, personnes vivant en couple
Source : Enquêtes annuelles de recensement 2013 à 2017

Les différences de traitement après la collecte selon qu'elle a eu lieu sur papier ou sur internet ont un impact plus ou moins important selon les variables. Cet impact est déterminant dans les résultats obtenus ici, et doit être discuté pour nos deux principales variables d'intérêt : le sexe et le prénom.

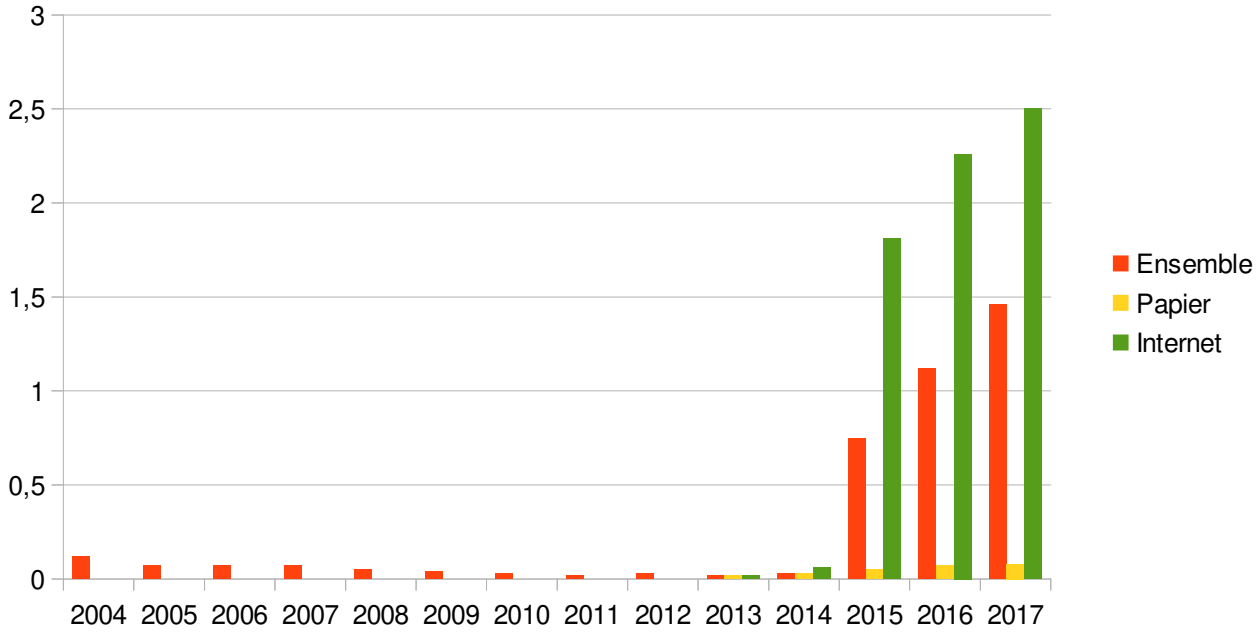
Sur le sexe, les enquêtés ont une case à cocher, homme ou femme, et les erreurs sont très rares au cours de la collecte papier. La différence entre les deux modes de collecte est très faible. Toutefois, la non-réponse sur papier est corrigée directement au moment de la saisie des questionnaires lorsque cela est possible. Les consignes de saisie indiquent ainsi explicitement que le prénom est pris en compte pour affecter un sexe en cas de non-réponse ou de réponse difficile à interpréter (les deux cases, homme et femme, sont cochées par exemple) :

> *« Cette variable, présente dans les BI, BIC et BIPLD, ne suit pas la règle de la modalité la plus forte lorsque les deux cases ont pour valeur 1.*
> *Si aucune case n'a pour valeur 1, ou bien si les deux cases ont pour valeur 1, il faut se reporter au prénom pour acquérir la variable. Si le prénom ne permet pas de saisir cette variable (prénom non renseigné, mixte ou inconnu), elle sera mise à blanc ».*
> https://www.insee.fr/fr/information/2526415

Cela explique l'extrême rareté de la non-réponse dans les fichiers issus de la saisie de la collecte papier : les corrections ont eu lieu en amont. Sur internet, ce processus de saisie et correction n'existe pas et la proportion de non-réponse est plus élevée, en légère croissance avec la généralisation de ce mode de collecte, en partie auprès de personnes qui sont un peu moins à l'aise avec l'outil informatique. Toutefois, à compter de l'EAR 2018, la réponse sur le sexe deviendra obligatoire sur internet (au même titre que la date de naissance, le nom et le prénom déclarés dans le tableau des habitants du logement, seules réponses requises impérativement). La refonte de la feuille de logement, expliquée ci-dessous, devrait aussi se traduire par la correction, en plus des cas de non-réponse, des cas d'incohérences entre le sexe renseigné sur le bulletin individuel et le sexe renseigné sur la feuille de logement. Ils seront en partie corrigés en amont des traitements

statistiques, ce qui devrait aussi diminuer la proportion d'erreurs de codage dans les cas de réponse papier.

Graphique 2 : Évolution de la proportion de non-réponses à la question du sexe, en %



Champ : Ménages ordinaires, personnes vivant en couple
Source : Enquêtes annuelles de recensement 2004 à 2017

Sur les prénoms, la reconnaissance optique de l'écriture manuscrite des personnes recensées est bien plus délicate et source d'erreurs, comparée à la saisie sur internet. Surtout, 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. Certes les personnes recensées par internet peuvent aussi faire des fautes de frappe ou répondre de façon inadéquate, mais les différences se sont avérées assez cruciales, nécessitant d'adapter le traitement proposé dans la suite du texte. Signalons dès maintenant que les prénoms ne figurent pas dans les fichiers de diffusion du recensement, et servent uniquement à la collecte. C'est pourquoi nous proposons de construire très en amont dans les chaînes de traitement des données du recensement un indicateur sur le caractère féminin/masculin du prénom, qui seul sera conservé ensuite.

1.3 Les situations conjugales dans le questionnaire du recensement

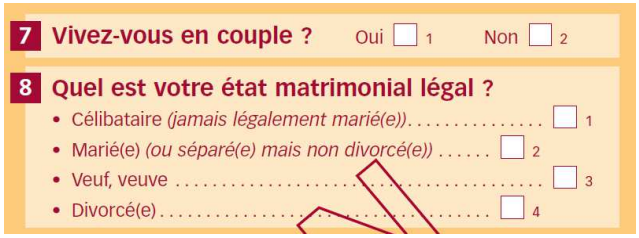
Pour chaque logement recensé, l'agent recenseur doit collecter une feuille de logement et un bulletin individuel par habitant du logement. La feuille de logement est le premier document renseigné, elle permet de construire les relations entre les habitants du logement. Par construction, cela limite l'analyse des relations de couple aux **unions cohabitantes**. En effet, la première unité du recensement est le logement. Il s'intéresse aux occupants habituels des logements et aux relations qu'ils entretiennent. Les circulations et relations entre personnes vivant dans différents logements ne sont approchées que dans la mesure où elles permettent de décider si la personne doit être comptée dans le logement enquêté ou dans un autre. La principale préoccupation est d'éviter les omissions ou doubles comptes pour garantir la qualité des dénombrements de la population affectée à chaque unité géographique (population légale, définie par voie réglementaire). **La définition du couple est ainsi conditionnée par le fait de déclarer vivre en couple et de partager un même**

logement. Cela limite l'appréhension des différentes formes de conjugalité[3]. Or les relations conjugales entre personnes de même sexe sont plus souvent non cohabitantes, « à distance », que les unions entre personnes de sexe différent (Toulemon et al., 2005 ; Rault, 2018). Néanmoins, compte tenu des contraintes méthodologiques, c'est bien aux seules unions cohabitantes que le présent document est consacré, qui restent la forme d'union très largement majoritaire également parmi les CMS (84 % en 2011, voir Buisson et Lapinte, 2013).

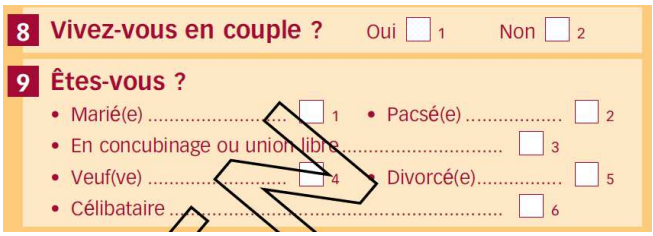
Une refonte du bulletin individuel a eu lieu en 2015, et celle de la feuille de logement en 2018. Ces deux refontes affectent la façon d'appréhender les relations conjugales dans le recensement.
Le changement dans le bulletin individuel de la question sur l'état matrimonial légal (marié, divorcé, veuf, célibataire) et son remplacement par une question sur les situations de fait (marié, pacsé, en union libre, divorcé, veuf, célibataire) améliore la qualité des réponses en rapprochant les modalités de réponse des situations concrètes des individus.

Extraits des fac-simile des bulletins individuels de recensement

Bulletin individuel 2004-2014



Bulletin individuel à partir de 2015



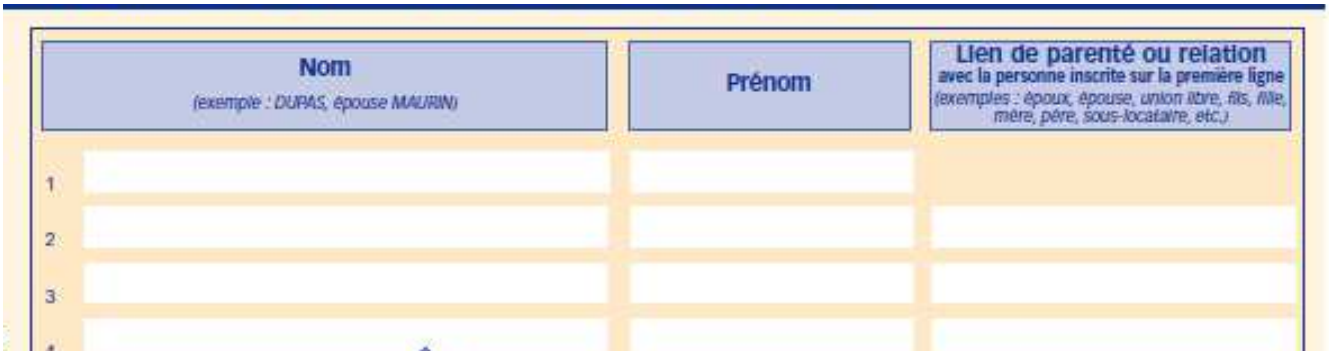
Par exemple, l'absence d'une modalité sur le PACS incitait certaines personnes à se déclarer mariées à la question sur le statut matrimonial légal, considérant cette modalité plus proche de leur mode de vie effectif (Buisson, 2017). Cette refonte a pu aussi renforcer l'idée que les unions entre personnes de même sexe doivent être déclarées et prises en compte au même titre que les autres dans le recensement. Elle ne modifie toutefois pas fondamentalement les modalités de collecte et de traitement des informations sur les relations entre les personnes du logement, au contraire de la refonte de la feuille de logement et de l'Analyse ménage-famille (AMF) à partir de la collecte 2018.
L'AMF permet de reconstituer des familles au sein des logements et d'établir des statistiques sur ces familles et leur composition, à partir des données collectées au recensement. Elle a aussi pour objectif de déterminer la position des différents habitants dans la famille. Cette analyse ne prend en compte les couples de même sexe que depuis 2015 et ses règles ont évolué pour ce faire.
Avant la refonte de 2018, la feuille de logement permettait la description des relations de chaque occupant du logement avec la première personne listée. Ces liens étaient manuscrits et n'étaient pas saisis. L'ensemble de traitements nommé « Analyse ménages-familles » était ensuite réalisé pour une partie seulement des logements, environ un sur quatre, afin de limiter les coûts manuels. Lorsque la composition du ménage était *a priori* (à l'aide d'un algorithme) estimée complexe, l'image scannée de la feuille de logement était visualisée par une personne en charge du codage des relations dans le logement et du type de famille. Le plus souvent, seule l'information des bulletins individuels était utilisée, lorsque la situation était évidente (une seule personne dans le logement), ou jugée suffisamment simple pour décider de la composition du ménage. Par exemple, si le logement comprenait uniquement deux habitants, ayant moins de 14 ans d'écart d'âge et déclarant tous deux vivre en couple dans leur bulletin individuel, alors le ménage était catégorisé comme composé d'une seule famille : un couple sans enfant[4].

[3]Certaines unions non cohabitantes peuvent être repérées lorsqu'une personne déclare vivre en couple sur son bulletin individuel sans conjoint dans le logement. Néanmoins, en l'absence de question sur le sexe du conjoint, il est impossible de savoir s'il s'agit d'unions non cohabitantes entre personnes de même sexe ou de sexe différent.

Cette situation change en 2018 : sur la nouvelle feuille de logement, chaque habitant habituel du logement a un numéro (01, 02 par exemple) et les enquêtés doivent donner le numéro de leur conjoint, quels que soient les rangs de l'enquêté et de son conjoint dans la liste des occupants du logement. Cette information (le numéro d'ordre du conjoint) sera systématiquement saisie. Cela permettra de décrire finement les liens conjugaux entre les habitants du logement deux à deux. La relation conjugale entre deux personnes sera mieux établie, puisqu'elle le sera à partir de ce qu'ont déclaré les personnes concernant leur situation de couple et non plus déduite des informations individuelles collectées pour chacun des conjoints sur le fait qu'il vit ou non en couple (sans préciser avec qui). L'analyse ménages-familles sera profondément transformée et généralisée à l'ensemble des logements, car elle ne reposera plus sur des traitements manuels.
En revanche, cela ne devrait pas directement améliorer le repérage des faux couples de même sexe puisque l'item « conjoint, conjointe » est unique, même sur internet, et ne distingue pas conjoints de même sexe et de sexe opposé. Cette refonte de la feuille de logement est l'aboutissement d'un projet de longue haleine, ce qui éloigne la perspective de nouvelles modifications substantielles intégrant une question directe sur la vie en couple avec un partenaire de même sexe. Elle crée toutefois 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 qui collectent des informations (descripteurs sociaux notamment) pour chacun des habitants. Cet appariement est réalisé sur le critère du sexe, de la date de naissance, et si ces deux premières variables sont insuffisantes pour réaliser l'appariement, du nom et du prénom. Pour réaliser cet appariement nécessaire au traitement du recensement, l'ensemble des prénoms et noms seront donc désormais exploitables (ce n'était pas le cas avant). C'est ce qui rend possible le traitement proposé dans le présent document. Cette nouvelle saisie a été organisée par anticipation dès la collecte 2016, permettant la réalisation des tests présentés dans la suite du document.

Extraits des fac-similé des feuilles de logement du recensement

Feuille de logement 2004-2017



[4]Plus précisément, jusque 2015, s'y ajoutait la condition que les conjoints soient de sexe opposé, ce qui conduisait à reclasser les conjoints de même sexe en célibataires. À partir de 2015, cette contrainte est levée, et la condition sur la vie de couple est élargie (soit les personnes déclarent vivre en couple, soit elles se disent mariées, pacsées ou en union libre à la question suivante).

Feuille de logement à partir de 2018

Numéro de la personne	Nom	Prénom	Sexe (Masculin/ Féminin)	Année de naissance (AAAA)	Pour chacune des personnes vivant dans ce logement, **renseignez le numéro de la personne** ayant l'un des liens de parenté suivants avec elle		
					Son conjoint (mariage, pacs, concubinage ou union libre)	**Sa mère** (biologique ou adoptive)	**Son père** (biologique ou adoptif)
1			M☐ F☐		Le conjoint de la **personne 1** est la **personne n°**	La mère de la **personne 1** est la **personne n°**	Le père de la **personne 1** est la **personne n°**
2			M☐ F☐		Le conjoint de la **personne 2** est la **personne n°**	La mère de la **personne 2** est la **personne n°**	Le père de la **personne 2** est la **personne n°**
3			M☐ F☐		Le conjoint de la **personne 3** est la **personne n°**	La mère de la **personne 3** est la **personne n°**	Le père de la **personne 3** est la **personne n°**
4			M☐ F☐		Le conjoint de la **personne 4** est la **personne n°**	La mère de la **personne 4** est la **personne n°**	Le père de la **personne 4** est la **personne n°**
					Le conjoint de la **personne 5**	La mère de la **personne 5**	Le père de la **personne 5**

1.4 L'enquête Familles et logements de 2011 adossée au recensement, une première estimation des couples de même de sexe

En 2011, l'Insee a réalisé une nouvelle édition de l'enquête Famille traditionnellement adossée au recensement tous les 10 ans environ : 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 les questionnaires du recensement. Selon les zones, ce questionnaire était à remplir par chaque femme adulte du logement (zones géographiques d'enquête des femmes) ou chaque homme adulte (zones d'enquête des hommes). Environ 360 000 questionnaires ont été collectés.

L'enquête a permis d'actualiser les analyses de la fécondité et d'avoir de nouvelles connaissances sur différents sujets : la multi-résidence, les familles recomposées, les personnes ayant contractualisé leur union par le pacte civil de solidarité (Pacs) mais aussi sur les couples de même sexe (Bodier, 2015).

La validation des couples de même sexe a fait l'objet d'un important travail. L'information sur le sexe des conjoints provenait des bulletins individuels de recensement, puisque l'EFL est une enquête adossée au recensement. Cette information était ensuite confrontée à celle disponible spécifiquement dans l'EFL, dont l'échantillon était réparti entre des zones d'agent recenseur collectant les données uniquement sur les femmes et d'autres zones collectant uniquement des données sur les hommes, zones réparties sur le territoire. L'enquête comportait ainsi deux versions du questionnaire, une pour les femmes et une pour les hommes et chaque agent recenseur avait soit des questionnaires femmes, soit des questionnaires hommes, facilitant ainsi la collecte, et indirectement la qualité des données sur le sexe. Cela revient en effet à coder de nouveau le sexe des personnes répondant à l'enquête, selon le type de questionnaire rempli. Si un logement comprenait d'après les bulletins individuels du recensement un couple d'hommes, et si les deux avaient rempli un questionnaire hommes d'EFL (dans une zone où seuls les hommes étaient enquêtés), cela validait la codification de leur sexe et le fait qu'il s'agisse bien d'un couple de personnes de même sexe. Le questionnaire de l'EFL comprenait aussi une question plus directe : « Votre conjoint(e)/ami(e) est 1) un homme 2) une femme ? ». Enfin, pour les cas qui restaient litigieux, les corrections ont été faites manuellement à l'aide des informations saisies lors de l'enquête et notamment les prénoms.

Ce travail a permis d'estimer le nombre de personnes en couple de même sexe à 205 000, dont 173 000 sont cohabitantes : 101 000 hommes et 72 000 femmes. Cette donnée nous sert de référence à la fois pour vérifier la façon d'appréhender les couples apparemment de même sexe dans le recensement et pour apprécier la cohérence de la mesure et des évolutions constatées.

1.5 Un indicateur transitoire : les couples apparemment de même sexe

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, il est utile de s'affranchir de l'analyse ménages-familles (et en premier lieu de la distinction entre ménage et famille), pour repartir des réponses « brutes » (avant redressement de la non-réponse) figurant sur les bulletins individuels de recensement. L'indicateur est ainsi construit à partir des données de tous les bulletins individuels dans un logement : si la moitié des personnes en couple dans le logement (ménage ordinaire) sont des femmes alors toutes les personnes en couple du logement sont classées en couple de personnes de sexe différent. Si le nombre de personnes en couple est égal à 1, on considère que la personne est en union non cohabitante avec une personne hors du logement. S'il y a deux personnes en couple, de même sexe, alors ces personnes sont classées en couples *apparemment* de même sexe. Les autres cas sont classés soit en couples de sexe opposé, soit en situations complexes et ne sont pas comptés comme des CMS. La méthode peut malgré tout conduire à des erreurs. Par exemple, deux personnes vivant ensemble et déclarant toutes deux « vivre en couple » parce qu'elles ont chacune un conjoint non cohabitant, vivant dans un autre logement, seront comptées à tort comme formant un couple ensemble. Mais ces erreurs sont assez rares, d'autant que la formulation du recensement « Vivez-vous en couple ? » les limite : elle est plus souvent interprétée par les répondants comme restreinte aux unions cohabitantes, à la différence de la question « Êtes-vous en couple ? » posée par exemple dans l'enquête Famille et Logements 2011 (Breuil-Genier et al., 2016). L'avantage de cet indicateur est qu'il peut être calculé en amont des traitements de l'analyse ménages-familles (par exemple sur l'EAR 2017), et sur différentes années indépendamment des évolutions de l'AMF (qui ne prenait pas en compte les couples de même sexe avant 2015).

À titre d'étalonnage, pour s'assurer que les imperfections que l'on ajoute en approximant la situation dans le ménage sont de moindre ampleur que celles liées aux erreurs de codage du sexe, la méthode de repérage des couples *apparemment* de même sexe a été appliquée aux données de l'enquête Famille et Logements de 2011 (EFL). Il est alors possible de comparer l'approximation à une mesure validée.

Si l'on se restreint aux personnes en CMS apparent d'après l'indicateur, c'est-à-dire ayant déclaré vivre en couple sur leur bulletin individuel de recensement, et pour lesquelles le seul conjoint potentiel dans le logement est de même sexe, 55 % (155 000 sur 283 000) sont *in fine* effectivement comptabilisées comme des personnes vivant en couple de même sexe dans le logement enquêté dans l'EFL adossée au recensement (donc le conjoint repéré dans le logement par l'indicateur est bien validé comme étant le conjoint et il est de même sexe). 40 % (114 000 sur 283 000) sont classées dans l'EFL comme des couples de sexe opposé cohabitant. Le conjoint repéré par l'indicateur est validé comme conjoint, mais il y a une erreur de codage sur le sexe d'un des conjoints qui a conduit à compter le couple par erreur parmi les CMS. Ces deux situations correspondent précisément à celles que l'on souhaite repérer dans l'indicateur des CMS apparents, afin de tester si l'on arrive, par la méthode proposée, à repérer les vrais CMS et les faux au sein de cet ensemble de CMS apparents.

La comparaison avec l'enquête Famille et logements permet aussi de vérifier que cet indicateur simplifié ne conduit pas à intégrer trop de personnes qui vivent en unions non cohabitantes en leur affectant à tort un conjoint dans le logement : ce sont quelques 14 000 individus comptés en CMS apparent qui ne sont pas en couple cohabitant d'après l'EFL. Cela ne représente que 5 % des personnes en CMS apparent d'après l'indicateur, ce qui paraît une erreur raisonnable. Dans l'autre sens, la comparaison avec l'EFL nous permet de vérifier que l'indicateur ne « rate » pas trop de situations de vrais CMS en considérant par exemple le ménage comme trop complexe. Le résultat est

là aussi satisfaisant puisque 90 % des personnes en vrai CMS après consolidation dans l'EFL sont repérées comme apparemment en CMS dans l'analyse préliminaire (155 000 / 173 000).
L'indicateur est donc correct pour estimer les CMS apparents, ce que l'on fera de façon plus solide suite à l'analyse ménages-familles à partir de sa refonte en 2018. 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 en sont omis.

Tableau 1 : Comparaison entre l'indicateur de CMS apparent, construit à partir des seules données des bulletins de recensement, et la situation de couple d'après les réponses consolidées à l'enquête Familles et Logements

		Situation d'après les réponses à EFL consolidées				
En milliers		CSO cohabitant	CMS cohabitant	CMS non cohabitant	Autres	Total
Situation apparente dans le logement, *d'après les bulletins individuels de recensement*	Ne vit pas en couple	298	13	21	16 719	17 052
	CSO	29 461	-	-	42	29 503
	CMS apparent	114	155	2	12	283
	Seule personne en couple	212	4	9	526	751
	Situation complexe	180	1	0	61	242
	Total	30 265	173	32	17 361	47 831

CSO = couple de sexe opposé
Champ : Personnes majeures, France métropolitaine
Source : Enquête Familles et Logements 2011, Insee, données pondérées.
Lecture : sur fond rose les CMS apparents, sur fond vert les « vrais » CMS validés lors de l'enquête EFL 2011, sur fond orange les situations que l'EFL a corrigées, sur fond bleu celles où la situation en CMS a été confirmée par l'EFL.

2 Les solutions testées à l'étranger

La difficulté de mesure des couples de même sexe n'est ni nouvelle ni spécifique à la France et différentes solutions ont été mises en place. 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[5]. Elle s'échelonne de 0,25 à 0,57 %. La part de ces mêmes « faux couples » dans le total des couples apparaissant comme de même sexe est comprise entre 27 et 55 %. 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.

2.1 Les solutions de redondance et recoupement d'informations

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 (erreur pour chacun des conjoints) est très faible.
Dans le recensement canadien depuis 2001, comme dans le recensement américain à compter de 2020, la relation conjugale est appréhendée grâce à quatre items :

[5]Il s'agit en principe de couples cohabitants, l'information portant généralement sur les personnes qui vivent dans le logement.

« **Quel est le lien entre cette personne et la Personne 1?**

- Époux ou épouse de sexe opposé de la Personne 1
- Partenaire en union libre de sexe opposé de la Personne 1
- Époux ou épouse de même sexe de la Personne 1
- Partenaire en union libre de même sexe de la Personne 1 »

Questionnaire téléchargé ici :
http://www23.statcan.gc.ca/imdb/p3Instr_f.plFunction=getInstrumentList&Item_Id=295241&UL=1V

Il est alors possible de recouper cette information avec le sexe des deux conjoints. Si deux conjoints ayant déclaré le 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, si les deux conjoints ont déclaré le même sexe et ont choisi cet item, la probabilité de deux erreurs est très faible et il s'agit selon toute vraisemblance d'un couple de personnes de même sexe. 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). De nombreux tests ont précédé cette mise en œuvre et ils 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).
Cette démarche de recoupement d'informations collectées de différentes manières est très similaire à celle adoptée pour l'enquête Familles et logements de 2011.
S'il peut paraître à première vue simple et efficace de dédoubler les modalités de la question du bulletin individuel de recensement sur la vie de couple : « Vivez-vous en couple ?[6] Oui avec une personne du même sexe / Oui avec une personne de sexe différent / Non », cette solution ne peut être envisagée à court ou moyen terme. En effet, les arbitrages sont difficiles entre différentes demandes d'ajout sur le bulletin individuel où la place est comptée puisqu'il doit impérativement conserver un format lisible sur deux pages. De plus, il vient d'être refondu comme déjà évoqué, et c'est un processus de longue haleine, qui nécessite la consultation et l'aval de nombreuses institutions, ainsi que des expérimentations préalables. Une prochaine enquête Famille et Logements adossée au recensement permettrait néanmoins de confronter à nouveau les résultats du recensement avec ceux d'une enquête posant les questions de façon plus directe.

2.2 Les solutions de validation par appariement à des données administratives

Cette solution 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 entre le sexe codé au recensement et celui du registre de sécurité sociale (Numident) lorsque les couples sont apparemment de même sexe au recensement.
Les proportions d'incohérences entre le sexe déclaré au recensement et celui figurant dans les données de sécurité sociale pour au moins un des conjoints sont très faibles au sein des couples de sexe opposé tandis qu'elles sont très élevées par les couples apparemment de même sexe, surtout s'ils sont mariés (72,7 %). Les écarts sont nettement moins importants s'agissant des couples non mariés : 6,4 % d'erreurs pour les couples apparemment de même sexe et 0,8 % pour ceux apparemment de sexe différent.

Cette démarche est donc en principe très efficace. En France, elle n'est utilisable à notre connaissance que pour les personnes qui font individuellement partie de l'échantillon démographique permanent. En effet, on dispose à des fins de gestion (variable non disponible dans les bases d'études) pour ces personnes du sexe enregistré dans le registre des personnes physiques géré par l'Insee (RNIPP), qui peut être comparé à celui collecté lors d'une des enquêtes annuelles de

[6]Dédoubler les modalités de la question sur la vie de couple paraît plus simple car pour celle sur les situations conjugales, il faudrait au moins dédoubler « marié(e) », « pacsé(e) » et « en concubinage ou en union libre ».

recensement. La présence (sauf exceptions) d'un seul des deux conjoints dans l'échantillon démographique permanent est toutefois une limite très forte à l'application de cette méthode pour améliorer la mesure des CMS. Néanmoins, c'est un outil très approprié pour confronter nos différentes solutions de correction et tester leur capacité à repérer de vraies erreurs de codage. Cela permet, comme expliqué ci-dessous, de valider la solution retenue par les prénoms.

2.3 Les solutions de « validation statistique » par le prénom

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 %, 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 comme masculin était porté par seulement 5 % d'hommes, ou moins, alors le sexe était corrigé. Autrement il était conservé. De façon symétrique, le sexe d'une personne codée comme femme était corrigé en homme si 95 % ou plus des porteurs de son prénom étaient des hommes. Ce seuil conduisait à corriger des incohérences entre le prénom et le sexe pour au moins un des conjoints dans 50 % des couples recensés comme de même sexe, 69 % en cas de couples mariés et 21 % au sein des unions libres.

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 % des personnes avaient un prénom considéré comme non ambigu et pouvaient donc faire l'objet d'une correction. Dans 96 % des cas, le sexe assigné sur la base du prénom était identique à celui figurant sur le registre de sécurité sociale.

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 en France.

3 Mode de correction envisagé en France

La correction du sexe doit donc passer par l'utilisation d'informations complémentaires qui permettent de remettre en question le codage de la déclaration du sexe proposée par le répondant. Le fait que le prénom soit à disposition de l'Insee pour mener des contrôles de qualité de la collecte depuis 2016 nous oriente vers une solution articulée principalement sur une « validation statistique » par le prénom. Pour mettre en place cette solution, il faut déterminer différents paramètres :
– sur quel champ appliquer la correction,
– comment construire le dictionnaire de prénoms et évaluer la fiabilité de la correspondance entre prénom déclaré et sexe associé,
– si l'on utilise d'autres variables complémentaires dans la correction,
– comment combiner l'ensemble de ces informations pour acter une correction de la variable déclarée par le répondant au recensement.

3.1 Tirer profit de l'échantillon démographique permanent pour cibler au mieux la correction

L'échantillon démographique permanent (EDP) est un panel sociodémographique 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 (Durier, 2018). Le principe général consiste à conserver pour les individus appartenant à l'échantillon (environ 4 % de la population) des informations collectées dans les cinq sources statistiques qui alimentent l'EDP. Ces cinq sources sont :

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

L'intérêt de l'EDP dans notre approche est de permettre la combinaison de deux informations :

- d'une part les données des enquêtes annuelles du recensement y sont intégrées. Elles comprennent les prénoms disponibles dans les bases de production afin de faciliter l'appariement avec les autres sources (Le prénom est en revanche absent de la base études à finalité statistique pour éviter une identification directe des personnes). Il est donc possible de calculer la variable d'intérêt associée à un dictionnaire indiquant la proportion de femmes ou d'hommes portant un prénom de la même façon qu'elle le sera en production dans les futures enquêtes annuelles de recensement ;
- d'autre part, dans le cadre des appariements, les informations sur les personnes EDP sont confrontées à celles du répertoire national d'identification des personnes physiques (RNIPP), répertoire qui sert à la gestion des numéros de sécurité sociale et est donc fiable. Il est donc possible de déceler les erreurs de codage du sexe au recensement lorsque celui-ci diffère de celui enregistré dans le RNIPP[7].

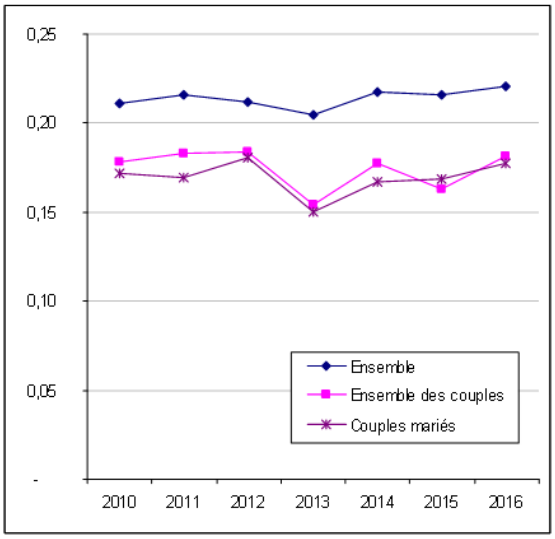
De cette façon, il est possible de tester la capacité de la procédure par les prénoms à repérer les erreurs de codage du sexe avéré. La détermination de la procédure optimale s'apparente au développement d'un test diagnostique en épidémiologie (encadré 1). C'est pourquoi les outils mobilisés sont

[7]Pour les besoins de la présente étude, on considère que le sexe codé dans le RNIPP est juste et qu'en cas de discordance, il y a une erreur dans le bulletin de recensement.

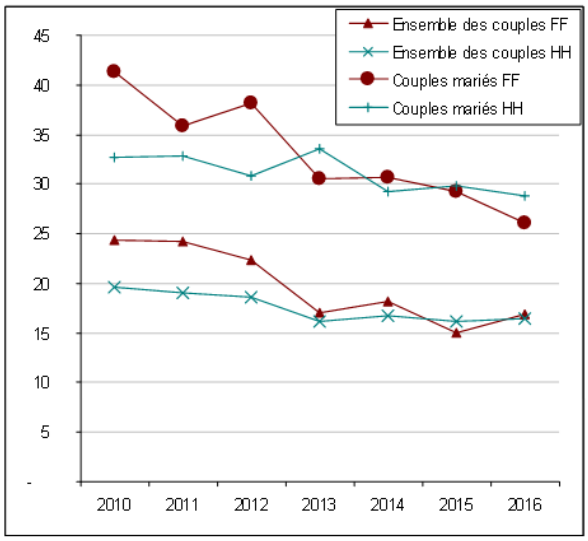
les mêmes : mesures de spécificité et de sensibilité, détermination du seuil optimal et comparaison de tests grâce à la courbe ROC (Robin, 2011).
L'équipe chargée de la production de l'échantillon démographique permanent a en effet accepté d'appliquer les différents dictionnaires testés aux prénoms, saisis afin de fiabiliser l'appariement entre sources via le numéro d'identification au répertoire (NIR). Les indicateurs en sortie de cette application des dictionnaires ont ensuite été appariés à la base études 2016 afin de tester leur capacité à repérer les erreurs de codage.
Dans la base études 2016, on compte 2,8 millions de personnes EDP recensées à au moins une EAR depuis 2010. Parmi elles, le taux d'erreur sur le sexe (discordance entre le RNIPP et le bulletin de recensement) est de 0,21 %, soit près de 6 000 erreurs. C'est donc un phénomène très rare. Le taux est légèrement plus faible pour les personnes en couple et il n'existe pas de tendance évidente à la hausse ou à la baisse selon les années. Le développement de la collecte internet de façon significative depuis 2014 pourrait faire diminuer le taux d'erreur car ce taux est un peu plus faible sur internet que sur papier. Mais l'écart est assez restreint. Le taux d'erreur est en revanche considérablement plus élevé pour les personnes apparemment en CMS, pour lesquelles il s'élève à 16 % en 2016, avec une tendance sensible à la baisse entre 2010 et 2016. La proportion d'erreurs de codage est plus importante parmi les personnes apparemment en CMS et qui se sont déclarées mariées au recensement.
II faut noter que dans un couple, on recherche le NIR uniquement pour la personne née un jour EDP, pour l'inclure dans l'échantillon démographique permanent et compléter les données statistiques la concernant. Les informations statistiques sur les habitants de son logement sont aussi incluses, mais sans identification – recherche de NIR – de ces personnes. On ne peut donc pas certifier le sexe des autres habitants du logement comme pour les personnes EDP elles-mêmes. Comme 0,17 % des personnes EDP en couple sont concernées par une erreur de codage, alors on peut estimer, en supposant que les erreurs de sexe entre deux conjoints sont indépendantes, que 0,34 % 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. Cette valeur est très proche de l'estimation réalisée à partir de l'enquête Famille et Logements.

Graphique 3 : Évolution des taux d'erreurs sur le codage du sexe

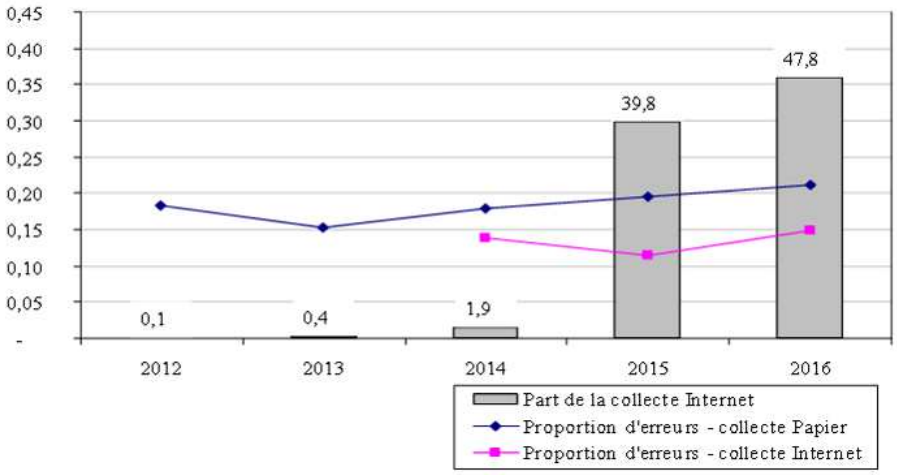
a) Ensemble des individus de l'EDP recensés



b) Personnes vivant apparemment en CMS



c) Selon l'année et le mode de collecte pour les personnes en couple



Source : EDP bases étude 2016, Insee.

- **La construction d'un dictionnaire de prénoms**

Un dictionnaire de prénoms associe à chaque prénom la proportion de femmes (respectivement d'hommes) le portant, pour apparier cette information aux prénoms des personnes enquêtées dans le recensement (ou une autre enquête) et comparer le sexe déclaré par les enquêtés au sexe le plus fréquemment associé à ce prénom. La finalité est de repérer les cas d'erreur de codage les plus probables (voir encadré 2 pour le mode de construction du dictionnaire). L'échantillon démographique permanent nous a permis de comparer les performances des différents dictionnaires pour choisir celui qui est le plus efficace pour repérer les erreurs de codage du sexe concernant les personnes EDP.

Cette première évaluation a conduit à retenir une combinaison des différents dictionnaires testés. Nous avons ainsi choisi d'utiliser deux sources pour construire le dictionnaire retenu. En tout premier lieu, l'état civil, qui donne lieu chaque année à la diffusion d'un fichier de prénoms sur le site insee.fr. Ce fichier comprend l'ensemble des prénoms donnés à des enfants nés en France depuis 1900, par sexe, avec quelques conditions de fréquence d'attribution de ces prénoms. C'est la source privilégiée

a priori pour constituer les dictionnaires du fait de son caractère exhaustif. Il est important que le dictionnaire repose sur un très grand nombre d'observations afin d'avoir des fréquences d'attribution suffisantes pour calculer une proportion d'hommes et de femmes parmi les porteurs d'un prénom. Ce fichier a été complété en ajoutant les occurrences des prénoms pour les personnes recensées en 2017 et nées à l'étranger, puisque l'état-civil ne concerne que les personnes nées en France[8]. Les personnes nées à l'étranger portent en effet plus fréquemment un prénom absent du dictionnaire construit avec l'état-civil seulement[9]. Un fichier intégrant les prénoms de toutes les personnes enquêtées à l'enquête annuelle de recensement 2017, qui couvre l'ensemble des personnes résidant en France en 2017 était disponible pour la collecte du recensement de cette année-là. Ces prénoms ont été saisis aussi bien pour les répondants par internet que 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.

Le dictionnaire retenu est par ailleurs une combinaison de plusieurs dictionnaires au sens où on commence par chercher une correspondance dans un dictionnaire le plus détaillé possible (même prénom, même année de naissance). En cas d'absence de correspondance, on cherche dans un dictionnaire moins détaillé une correspondance sur la première partie du prénom et sans condition sur l'année de naissance. L'indicatrice finalement affectée a un individu 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 les personnes portant un prénom dont la première partie est identique. Par simplicité on parlera pour désigner cette indicatrice de la proportion de femmes portant le même prénom.

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

On observe sur les graphiques 4 a et b la distribution des personnes selon la valeur de la variable du dictionnaire qui leur a été affectée. Si le sexe déclaré au recensement est masculin, il s'agit de la proportion de femmes portant le même prénom d'après le dictionnaire. Si le sexe déclaré est féminin, alors il s'agit de la proportion d'hommes. Plus la valeur est élevée plus il y a suspicion d'erreur de codage, et on cherche à déterminer un seuil de la valeur au-delà duquel on considère qu'il y a erreur. Par convention, si le prénom de la personne n'est pas trouvé, la proportion est fixée à 0 car aucune correction ne peut être faite.

La valeur ainsi obtenue est étroitement corrélée aux erreurs constatées en confrontant l'enquête annuelle de recensement (EAR) et l'information du RNIPP : en l'absence d'erreur presque toutes les valeurs sont nulles ou très proches de 0. En cas d'erreur, elles sont au contraire presque toutes proches de 1, et les valeurs à 0 correspondent pour la plupart à des échecs d'appariement du prénom avec le dictionnaire. Cela signifie que le dictionnaire construit est très efficace pour repérer les erreurs de codage.

Ce graphique montre également la rareté des valeurs intermédiaires (qui correspondent aux prénoms portés indifféremment par des hommes et des femmes).

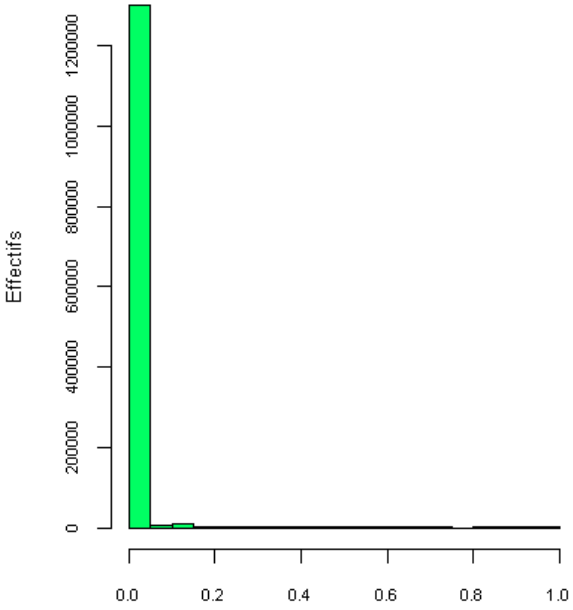
[8] Des entrées ont aussi été créées pour les « prénoms » qui n'apparaissaient pas dans le fichier des prénoms de l'état-civil et qui étaient présentes dans le recensement. Les cas sont assez rares, mais de nature variées : JEANLUC, M CHRISTINE, MARIA DE FATIMA, EPOUSE, etc.

[9] Parce que prénom n'était pas donné à des enfants nés en France, ou pas assez fréquemment.

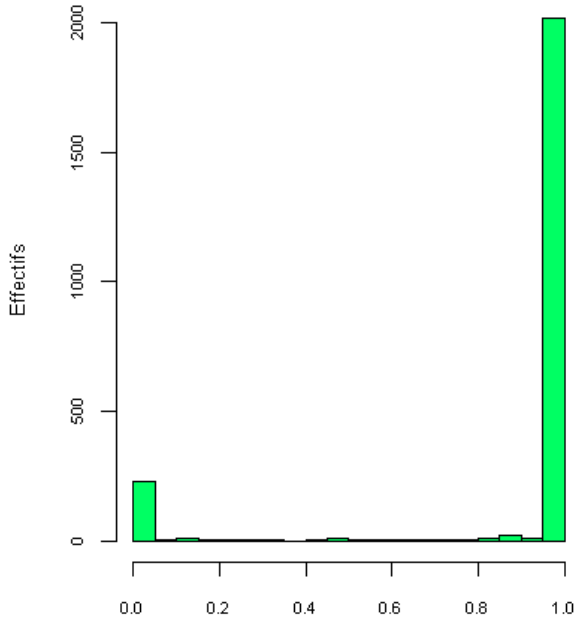
Graphique 4 : Répartition de l'indicatrice d'erreur selon la présence d'une erreur avérée

a) Ensemble des personnes EDP en couple

Pas d'erreur sur le sexe
(concordance entre l'EAR et le RNIPP)
n = 1 341 572



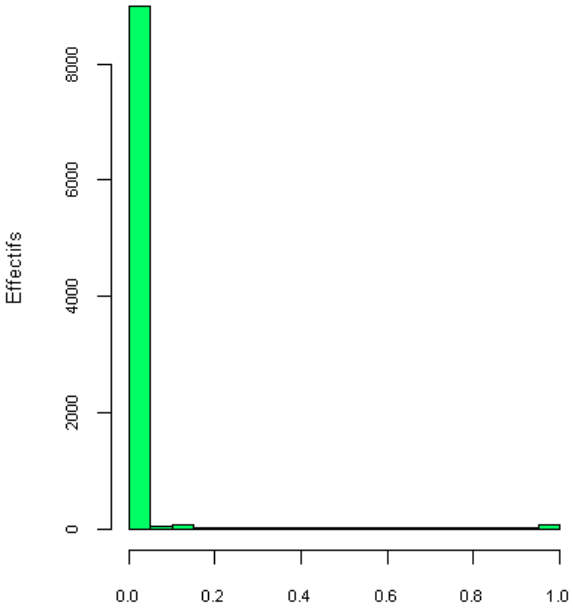
Erreur sur le sexe
(discordance entre l'EAR et le RNIPP)
n = 2 336



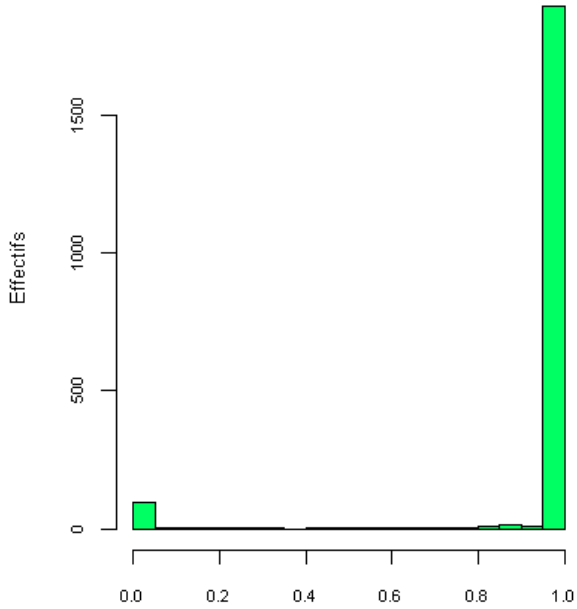
Proportion de personnes portant le même prénom et ayant un sexe différent

b) Personnes EDP apparemment en CMS

Pas d'erreur sur le sexe
(concordance entre l'EAR et le RNIPP)
n = 9 290



Erreur sur le sexe
(discordance entre l'EAR et le RNIPP)
n = 2 059



Proportion de personnes portant le même prénom et ayant un sexe différent

Source : EDP bases étude 2016, Insee.

- **Il est préférable de limiter les corrections apportées aux seuls CMS apparents**

Parmi les personnes en couple, la grande majorité de celles concernées par une erreur de codage se voient comptées comme des CMS apparents : sur 2 336 erreurs repérées parmi les 1 343 908 individus en couple, 2 059, soit 88 %, concernent un des 11 349 CMS apparents. Les 277 autres erreurs doivent être recherchées parmi 1,3 million de personnes, et pèsent donc peu. Proposer une correction du sexe déclaré grâce au prénom déclaré semble donc très prometteur dans le cas des CMS apparents. Cela semble en revanche plus risqué si on généralise la correction à l'ensemble de la population.

Effectivement, Le nombre de corrections « abusives » serait alors nettement plus élevé que l'effectif que l'on cherche à corriger, compris entre 10 000 et 20 000 personnes si on corrige toute la population dans l'EDP, alors qu'il est inférieur à 200 si on se limite aux CMS.

Tableau 2 : Répartition des erreurs de codage sur le sexe selon la situation conjugale apparente

	Erreur		
	OUI	NON	Total
CMS apparents	2 059	9 290	11 349
Autres	277	1 332 282	1 332 559
Total	2 336	1 341 572	1 343 908

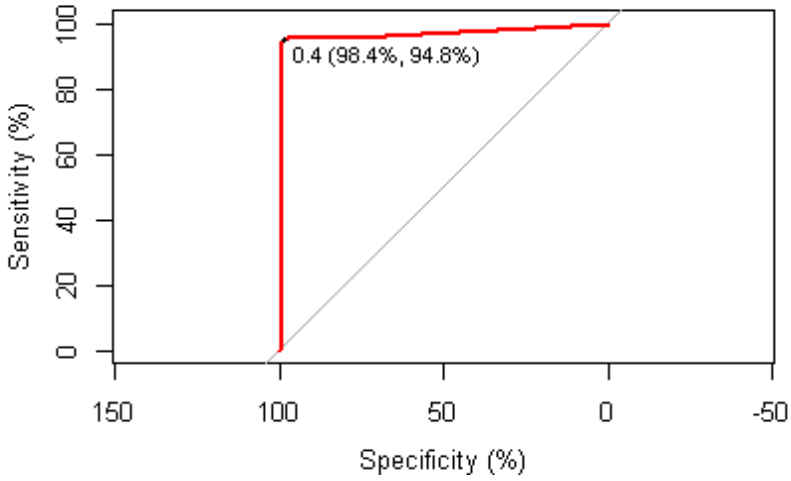
Source : EDP, base études 2010-2016, Insee.

Champ : Personnes de plus de 15 ans vivant en couple.

- **Le dictionnaire choisi est très efficace et le seuil optimal de correction est peu élevé**

Le dictionnaire retenu a de très bonnes performances quand on l'applique à l'EDP en se limitant aux personnes en CMS apparent. L'analyse de la courbe ROC conduit à retenir un seuil de 41 %. Cela signifie que pour un homme, si plus de 41 % des personnes portant le même prénom sont des femmes, une correction est apportée. Un tel seuil peut paraître étonnant car le test préconise une correction alors même que la majorité des porteurs du prénom sont des hommes. Retenir le seuil préconisé paraît néanmoins justifié dans la mesure où la correction ne sera effectuée que sur les CMS apparents, c'est-à-dire dans un contexte où le risque d'erreur de codage sur le sexe est connu pour être très élevé par rapport à la situation en population générale. De surcroît, le test est dans tous les cas de très bonne qualité car il y a très peu de cas où la valeur du dictionnaire est ambiguë, comprise entre 20 et 80 %. La fixation du seuil a donc un impact modéré dans un contexte aussi favorable. Avec le dictionnaire retenu et le seuil de 0,41, cela permet d'atteindre une sensibilité de 98 % et une spécificité de 95 %. Compte tenu des effectifs faibles, le nombre de corrections à tort serait d'environ 150 dans l'EDP, pour près de 2 000 vraies erreurs repérées.

Graphique 5 : La courbe ROC pour le dictionnaire choisi appliqué aux CMS apparents.



Source : Base études 2016 de l'échantillon démographique permanent, EAR 2017, fichiers des prénoms de l'Etat-civil.

Tableau 3 : Performances du dictionnaire pour repérer les erreurs de codage sur le sexe

	Personnes vivant en couple	*Personnes vivant apparemment en CMS*
Effectif	1 343 908	11 349
Erreurs avérées	2 336	2 059
Meilleur seuil	0,41	0,41
Spécificité	0,99	0,98
Sensibilité	0,89	0,95
Corrections abusives	16950	147
Corrections valides	2088	1 951

Source : EDP, base études 2010-2016, Insee.
Champ : Personnes de plus de 15 ans vivant en couple.

3.2 La confrontation à l'EAR 2017 et l'introduction du mode de collecte

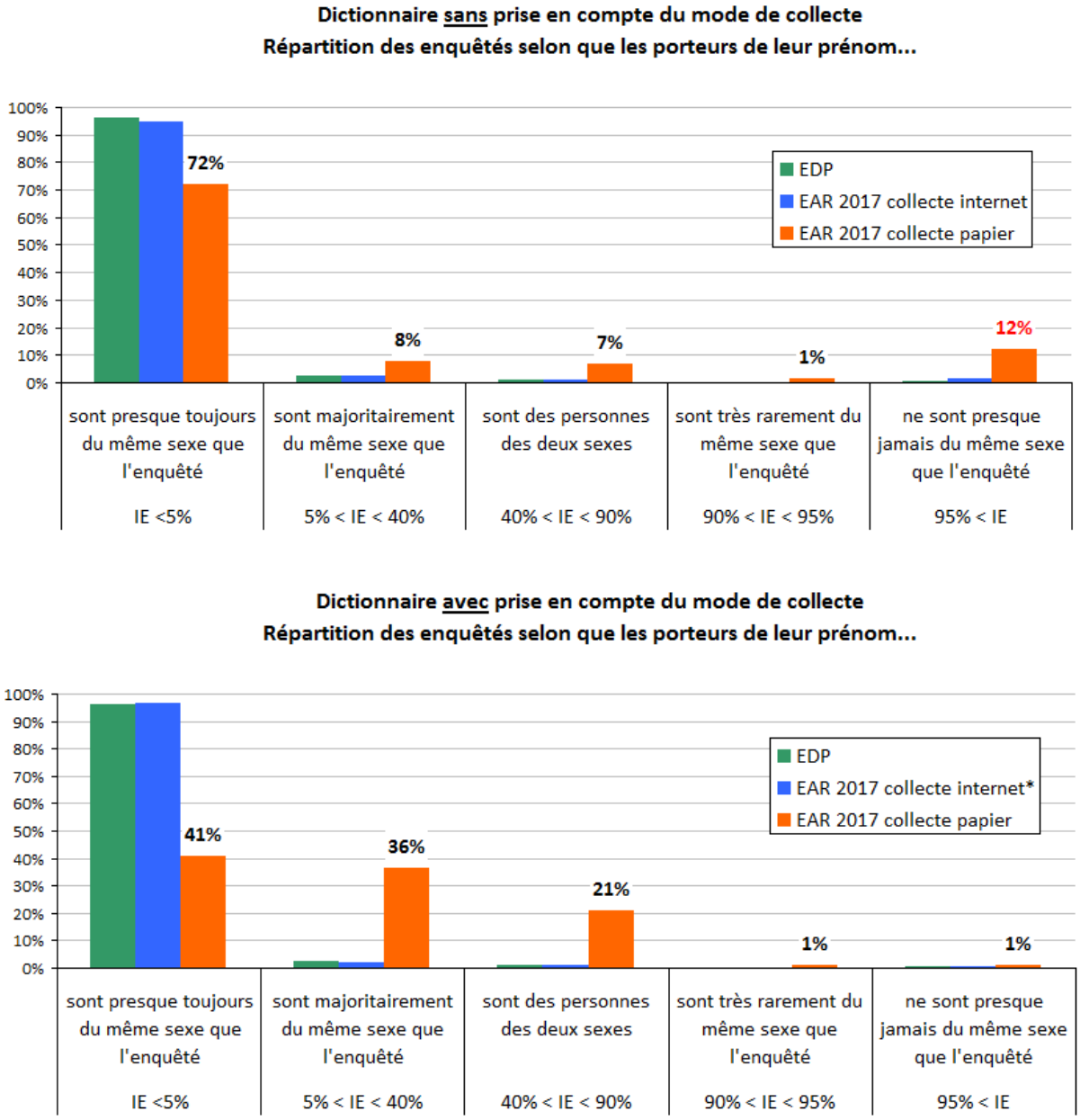
L'utilisation de l'échantillon démographique permanent permet donc d'évaluer les performances du dictionnaire choisi et des modalités de la correction sur un échantillon dans lequel les erreurs de codage sont connues. Comme les données de l'EDP utilisées sont une extraction de celle du recensement, cela semble a priori une source idéale pour préfigurer les performances de la correction une fois appliquée à l'ensemble du recensement. Néanmoins, nous avons aussi testé le comportement des différents dictionnaires, et particulièrement de celui retenu, sur l'**enquête annuelle de recensement 2017**, dont les prénoms étaient disponibles transitoirement à des fins de gestion et de contrôle de la qualité de la collecte durant l'année 2017.
Sur cette enquête, il a été possible de comparer les différents dictionnaires sur le taux d'échecs (proportion de personnes dont le prénom ne correspond à aucune entrée du dictionnaire) et la distribution de l'indicatrice mesurant le risque d'erreur selon les différentes situations, et notamment selon que la personne est ou non en CMS apparent.

- **Un dictionnaire adapté pour éviter des corrections excessives de la collecte papier de l'EAR**

Il apparaît suite à ces tests que **la transposition des résultats obtenus sur l'EDP ne peut se faire de façon automatique et qu'il est nécessaire de différencier le mode de construction du dictionnaire selon le mode de collecte** (Internet/Papier). En effet le niveau de qualité du prénom disponible dans

l'EAR est très différent entre la collecte papier et la collecte par Internet du fait du processus de saisie du prénom où le niveau de qualité demandé au prestataire n'est pas très élevé (encadré 3). Sans prise en compte du mode de collecte nous étions amenés à considérer qu'environ 10 % 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. Or en fait, il s'est avéré que les erreurs portaient bien plus probablement sur l'acquisition du prénom que sur le sexe. Une fois le type de collecte intégré à la construction des dictionnaires, la proportion de situations où lors de la collecte papier, plus de 95 % portent un sexe différent de celui porté par l'enquêté devient très faible. L'application du dictionnaire ne génère plus d'erreurs supplémentaires en forçant des corrections sur des prénoms traités incorrectement lors de la saisie. Le revers est la proportion élevée de cas ambigus, ce qui correspond à la réalité compte tenu de la qualité de la saisie. Cela nécessitera une prise en compte spécifique dans les traitements ultérieurs. Comme la qualité est meilleure dans l'échantillon démographique permanent, même en cas de collecte papier, il ne sera pas possible de tester sur l'EDP la façon de traiter la moindre qualité de la collecte papier du recensement.

Graphique 6 : Répartition des enquêtés selon la proportion de porteurs de leur prénom qui sont du même sexe qu'eux, avec et sans prise en compte du mode de collecte dans la construction du dictionnaire



* : Hors FLNE et hors ménages où le sexe d'un des deux conjoints a été imputé.
Sources : Bases études 2016 de l'échantillon démographique permanent et enquête annuelle de recensement 2017, Insee.

S'agissant de la collecte internet en revanche, les résultats sont bien plus proches de l'EDP. La différence qui subsistait était liée à la correction de la non-réponse sur le sexe : le sexe est imputé pour la collecte internet à l'aide d'algorithmes. Hors non réponse au sexe, les répartitions sont très proches entre collecte internet et EDP. La non-réponse sur le sexe va disparaître à compter de 2018 (le nouveau module feuille de logement sur internet requiert impérativement une réponse de l'internaute sur son sexe), rapprochant encore les deux sources et leurs traitements.
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.

- **Pour la collecte internet, la méthode validée sur l'EDP est transposée**

Pour la collecte internet de l'enquête annuelle de recensement, il est proposé d'utiliser le dictionnaire adapté à internet et de retenir un seuil de 0,51 similaire à celui obtenu sur la collecte internet de l'EDP. Pour une personne en CMS apparent, une correction sera apportée au sexe déclaré dès que la proportion de personnes portant un sexe opposé au sien est de 51 %. Cette méthode permet de proposer une correction pour la très grande majorité des erreurs de codage avérés : 99 % des erreurs font effectivement l'objet d'une correction. En revanche, la méthode a une légère tendance à proposer des corrections à tort : 5 % des cas sans erreur de codage du sexe sont corrigés. Appliquées aux seuls CMS apparents, ces corrections abusives n'ont néanmoins que très peu d'influence sur le taux de couple de même sexe calculé après correction.

Tableau 4 : Comportement du dictionnaire appliqué à l'EDP selon le mode de la collecte pour les personnes en CMS apparent

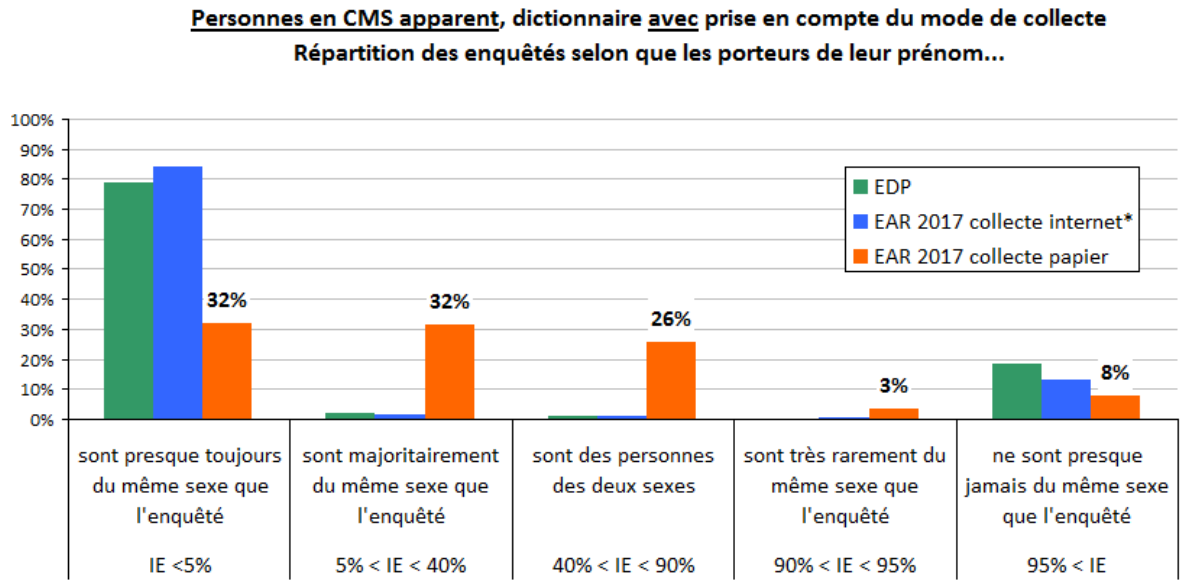
	Ensemble	***Papier***	***Internet***
Effectif	11 349	9 605	1 744
Erreurs avérées	2 059	1 852	207
Seuil	0,41	0,41	0,51
Spécificité	0,98	0,98	0,99
Sensibilité	0,95	0,95	0,95
Correction abusive	147	134	9
Correction valide	1951	1754	197

Source : EDP, base études 2010-2016, Insee.
Champ : Personnes de plus de 15 ans vivant en CMS apparent.

- **Pour la collecte papier, la correspondance entre sexe et prénom est plus incertaine**

Pour la collecte papier de l'EAR, il est beaucoup plus délicat de s'appuyer sur la collecte papier de l'EDP, dont la saisie est de meilleure qualité. Le dictionnaire a été construit de façon à prendre en compte le mode de collecte, permettant d'éviter des corrections abusives. Néanmoins, cela ne permet pas d'éviter qu'une proportion assez élevée de prénoms soient associés à des valeurs intermédiaires de proportions. Le critère du prénom fonctionne de ce fait moins bien sur papier. Pour la collecte internet, le résultat n'est pas très sensible au seuil, il l'est beaucoup plus sur papier. Avec un seuil de 95 % par exemple, 8 % des individus en CMS apparents collectés sur papier seraient corrigés, 13 % sur internet. En passant à un seuil de correction de 40 %, les taux de correction sur papier sont presque multipliés par 5 et passent à 37. Ils sont presque inchangés sur internet, passant à 14. Cela vient des répartitions très différentes des probabilités d'erreur, avec beaucoup plus de valeurs intermédiaires pour la collecte papier, du fait de la saisie des prénoms.

Graphique 7 : Comparaison du dictionnaire appliqué aux personnes en CMS apparent dans l'EDP et dans l'EAR collecte internet et collecte papier



* : hors FLNE et hors ménages où le sexe d'un des deux conjoints a été imputé.
Champ : Personnes en CMS apparent
Sources : Bases études 2016 de l'échantillon démographique permanent, EAR 2017, Insee.

L'incertitude est donc assez forte sur la proportion estimée de CMS, car celle-ci est très dépendante du seuil fixé. En corrigeant effectivement les sexes des personnes en CMS apparent, on obtient avec un seuil de 50 % une estimation de la proportion de « vrais CMS » parmi les couples cohabitants de 0,79. Avec un seuil de 95 %, l'estimation passe à 0,94. L'estimation pour la partie internet de l'EAR varie très peu (de 0,92 à 0,94), tandis que pour la partie papier le changement de seuil conduit à augmenter l'estimation de moitié (de 0,63 à 0,94).
Cela montre bien que sur la collecte papier les valeurs intermédiaires de l'indicateur doivent être considérées avec beaucoup de circonspection. Elles indiquent une forte incertitude sans parvenir à la lever. Une stratégie s'appuyant sur des variables annexes est donc proposée.

Tableau 5: Estimation des « vrais CMS » selon le seuil

	Papier	*Internet*	*Ensemble*
Proportion d'individus en CMS apparent avant correction	1,09	1,23	1,17
Proportion estimée d'individus en « vrai » CMS, *seuil de 50%*	0,63	0,92	0,79
Proportion estimée d'individus en « vrai » CMS, *seuil de 95%*	0,94	0,94	0,94

Champ : Personnes en couple, population des ménages, EAR 2017, effectifs pondérés, hors FLNE et hors ménages où le sexe d'un des deux conjoints a été imputé, Insee.

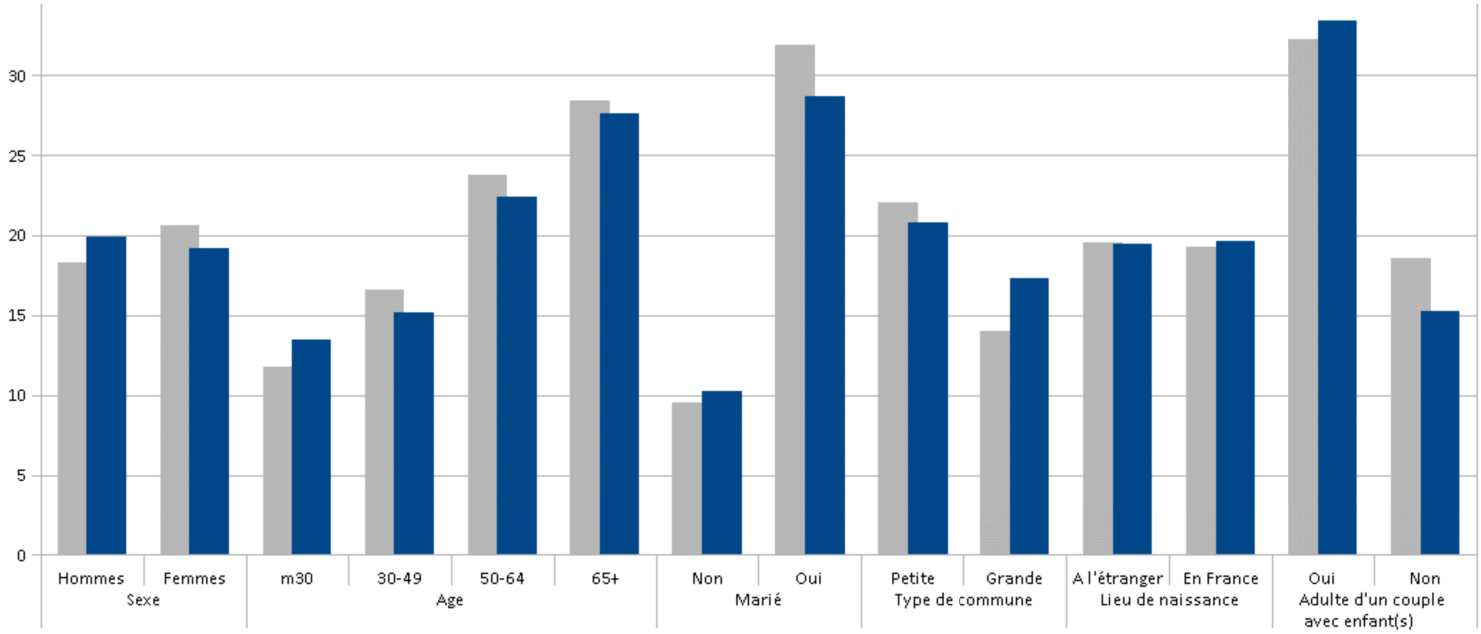
- **Une stratégie adaptée pour la collecte papier**

Afin de mieux asseoir la décision de correction du sexe et dans la mesure où l'information tirée du prénom est insuffisante, il est proposé de s'appuyer sur d'autres variables qui peuvent aider à estimer si la personne est vraiment en CMS ou s'il y a eu erreur de codage du sexe.
Pour trouver quelles variables sont les plus déterminantes, l'EDP est là encore d'une grande utilité.
Plusieurs variables ont été testées. Leur corrélation avec la proportion d'erreurs de codage est calculée sur l'ensemble des individus EDP en CMS apparent recensés entre 2010 et 2016 sur papier, mais aussi en se restreignant à 2015 et 2016. En effet, en 2010, le mode internet n'existait pas et l'ensemble de la collecte de l'EAR était faite sur papier. En revanche, depuis que l'internet est un

mode de collecte significatif, les répondants papier présentent des caractéristiques particulières (plus âgés par exemple que la moyenne). Les répondants papiers de 2015-2016 ressemblent sans doute davantage aux répondants papier 2017 et futurs que ceux d'avant 2015 (c'est-à-dire presque tous les recensés de ces années-là).

Trois groupes, qui diffèrent assez significativement par la proportion d'erreurs de codage, sont proposés : les personnes non mariées, pour lesquelles les taux d'erreur sont les plus faibles (10 %) ; les personnes mariées et âgées de moins de 50 ans (23 % d'erreurs) ; les personnes mariées et âgées de 50 ans ou plus (32 %)[10].

Graphique 8 : Proportion d'erreurs de codage sur le sexe parmi les personnes en CMS apparent selon leurs caractéristiques



Gris : recensés sur papier 2010-2016, bleu : recensés sur papier 2015-2016
Source : Bases études EDP 2016, Insee.
Champ : Personnes en couple et apparemment en CMS.

Tableau 6 : Part de chaque groupe et proportion d'erreurs de codage sur le sexe au sein des CMS apparent

			Part	*Proportion d'erreurs (en %)*
Non marié	Adulte d'un couple avec enfant(s)		9	27
	Autres situations		40	6
Marié	Adulte d'un couple avec enfant(s)		15	37
	Autres situations	Moins de 50 ans	8	7
		50 ans et plus	28	31
Ensemble			100	20

Source : Base études EDP 2016, Insee.
Champ : Personnes en couple, en CMS apparent, recensés sur papier en 2015 ou 2016.

[10] A partir de 2018, il sera possible d'intégrer à la correction le fait d'être parent d'enfants présents dans le logement, qui augmente considérablement la probabilité qu'il s'agisse d'une erreur de codage du sexe, surtout si la personne vit en union libre. La nouvelle analyse ménage-famille sera en effet étendue à l'ensemble des personnes recensées alors qu'elle est restreinte jusqu'en 2017 à l'exploitation complémentaire.

3.3 Description de la méthode de correction proposée pour l'appliquer aux EAR

Les enseignements de l'EFL combinés aux différents scénarios testés sur les données de l'EDP et des spécificités de l'EAR prise dans son ensemble ont conduit à retenir les choix de correction suivants :

Le champ d'application de la correction est restreint aux personnes en CMS apparent. Lorsque la méthode sera appliquée en régime courant, on disposera des résultats de l'analyse ménage famille (AMF) du RP pour mener la correction. On se limitera donc plutôt en régime courant aux personnes en CMS d'après l'AMF.

Le dictionnaire des prénoms sera construit à partir des données de l'état civil combinées aux données de la dernière EAR (utilisées essentiellement pour les personnes recensées sur papier et/ou nées à l'étranger). Dans le dictionnaire, les entrées pour un prénom donné seront décomposées selon l'année de naissance, une règle de simplification du prénom (prénom entier/première partie du prénom) et le mode de collecte (Internet/Papier).

Pour les individus qui ont répondu par internet, on proposera de corriger le sexe dès lors que le dictionnaire indique que les personnes portant le même prénom ont un sexe différent du prénom déclaré dans plus de 51 % des cas, seuil optimal sur la courbe ROC pour la collecte internet de l'EDP. Dans le cas contraire, on ne proposera pas de correction.

Pour les individus qui ont répondu sur papier, on considère que le prénom de l'EAR n'est pas forcément le prénom déclaré. On est donc plus prudent avant de mener une correction.

- Si le prénom est porté par plus de 90 % de personnes de sexe opposé à individus, on corrige le sexe.
- S'il est porté par moins de 20 % de personnes de sexe opposé à l'individu, on ne corrige pas le sexe.
- S'il est porté par entre 20 % et 90 %[11] de personnes de sexe opposé à l'individu, on considère que le prénom est ambigu et qu'il est nécessaire d'utiliser des informations complémentaires pour affiner la correction : on corrige de façon aléatoire en fonction du groupe (non marié avec enfant(s), non marié sans enfant, marié avec enfant(s), marié sans enfant moins de 50 ans, marié sans enfant 50 ans et plus). 39 % des individus en CMS apparent de la collecte papier ont une valeur intermédiaire et 1 % une valeur manquante. Au total, ce sont 17 % des personnes en CMS apparents, qui se voient qualifier de vrais ou faux CMS de façon aléatoire en fonction de leur groupe.

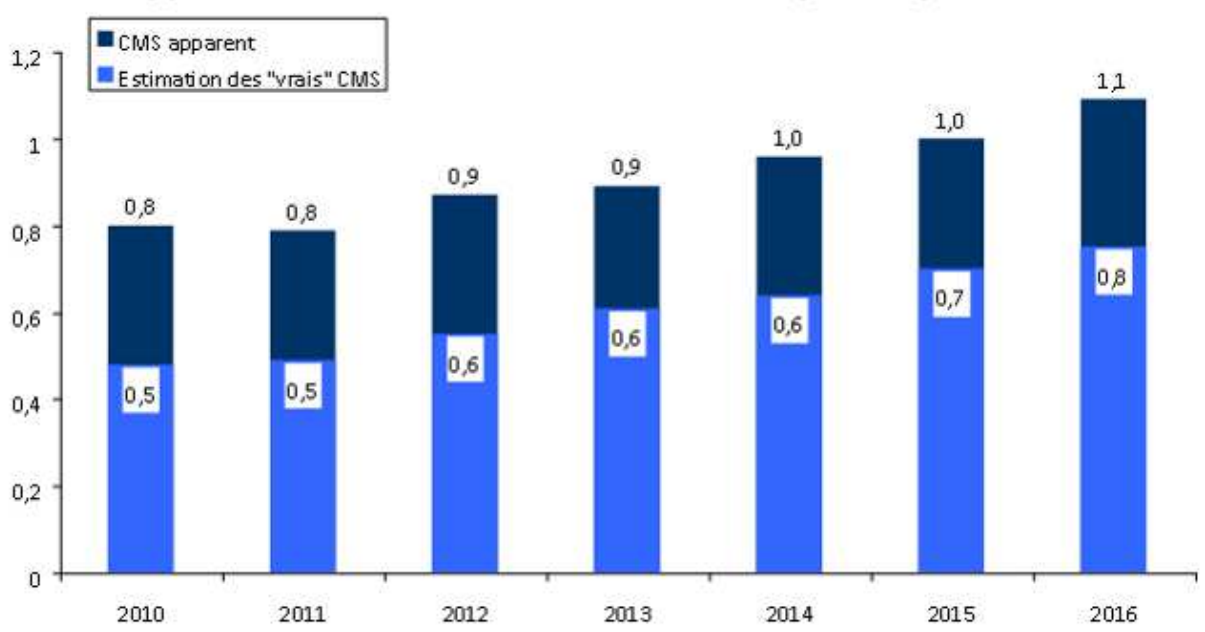
[11] L'intervalle est asymétrique parce qu'en cas d'incohérence entre sexe et prénom, nous considérons que l'association d'un prénom à un sexe est davantage bruitée par une mauvaise qualité de l'acquisition du prénom par rapport à une mauvaise qualité du codage du sexe. On accorde donc toujours du crédit au codage du sexe tant que le prénom saisi pour l'enquêté entre en contradiction pour moins de 20 % des cas. En revanche, on accorde du crédit au prénom seulement à partir du moment où le prénom correspond au sexe codé pour moins de 10 % des cas.

4 L'application du dictionnaire retenu à la comptabilité des CMS

4.1 Dans l'EDP, une estimation de la proportion de couples en CMS, 2010-2016

Dans l'EDP, il a été possible d'appliquer le dictionnaire, avec un seuil de 0,41, aux individus EDP, et non à leur conjoint. Toutefois, en supposent que la fréquence des corrections serait identique parmi les conjoints, il est possible d'obtenir une estimation des vrais CMS, après correction.
Les effectifs et proportions ainsi obtenues sont cohérentes avec celles relevées dans l'enquête Familles et Logements en 2011. Elles montrent aussi que la proportion de personnes en CMS parmi celles en couple cohabitant semble avoir augmenté régulièrement depuis 2010, passant de 0,50 % à 0,75 %.

Graphique 9 : Évolution de la proportion de CMS apparents et de « vrais CMS » estimés



Champ : Personnes de plus de 15 ans ayant déclaré vivre en couple dans leur bulletin individuel.
Source : EDP base études 2010-2016, données pondérées, Insee.

Tableau 7 : Évolution des CMS apparents et de l'estimation dans l'EDP

	Effectifs		En % des personnes en couple	
	CMS apparents	Estimation "vrais" CMS	CMS apparents	Estimation "vrais" CMS
2010	247 000	149 000	0,80	0,48
2011	245 000	147 000	0,79	0,49
2012	269 000	166 000	0,87	0,55
2013	276 000	187 000	0,89	0,61
2014	300 000	198 000	0,96	0,64
2015	311 000	213 000	1,00	0,70
2016	337 000	231 000	1,09	0,75

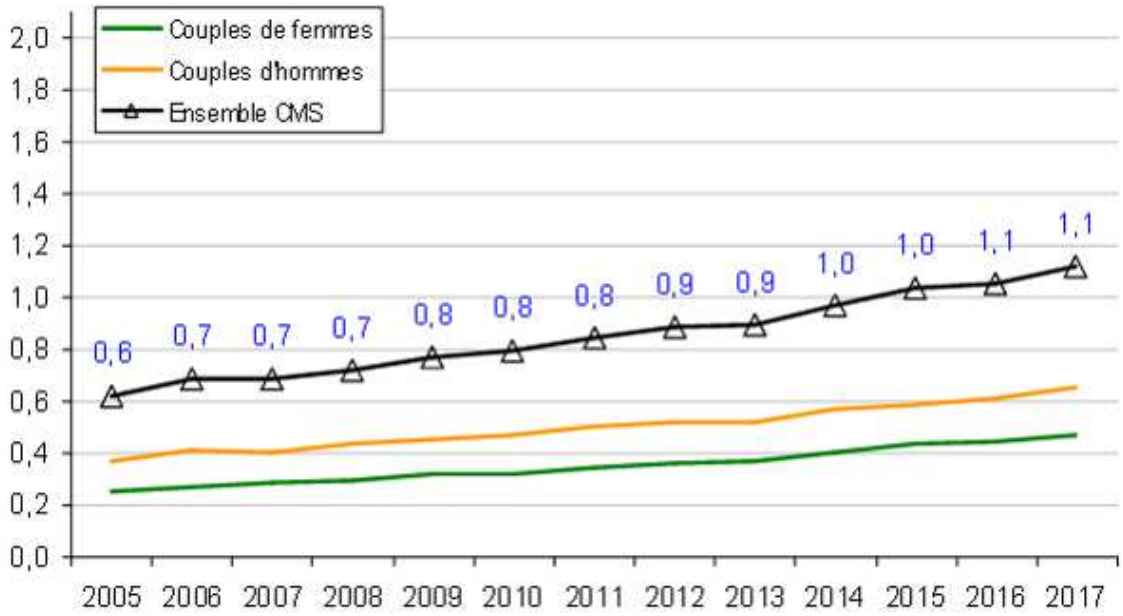
Champ : Personnes de plus de 15 ans ayant déclaré vivre en couple dans leur bulletin individuel.
Source : EDP base études 2010-2016, données pondérées, Insee.

4.2 Dans les enquêtes annuelles de recensements

Près de 47 000 personnes ayant répondu à l'enquête annuelle de recensement 2017 sont en CMS apparent, avant pondération. Après pondération, les résultats obtenus sont sensiblement identiques à ceux de l'EDP, montrant que la hausse des CMS apparents s'observe aussi sur une période plus longue, entre 2005 et 2017. Le rythme de croissance est similaire pour les couples de femmes et d'hommes.
Il n'est en revanche pas possible d'estimer les « vrais » CMS et leur évolution sur les années 2005-2016 à partir des seules données des enquêtes annuelles car les prénoms n'ont pas été conservés.
Pour l'EAR 2017, il a été possible de confronter les prénoms aux différents dictionnaires. L'avantage par rapport à l'EDP est que l'on dispose des prénoms des deux conjoints dans les couples (cohabitants). Par ailleurs, cela permet de tester la méthode retenue dans des conditions très proches de celles des futures EAR (nonobstant quelques évolutions : refonte de la feuille de logement, obligation de réponse sur le sexe).

Grâce à l'utilisation d'un dictionnaire de prénoms, on peut donc estimer qu'en 2017, environ 250 000 personnes, soit 0,8 % des personnes en couples cohabitants ont un conjoint de même sexe. Cette proportion aurait augmenté sensiblement depuis la dernière estimation disponible, celle de l'enquête Famille et Logements en 2011 (0,6 %).

Graphique 10 : Proportion de personnes en CMS apparent parmi celles déclarant vivre en couple



Source : enquêtes annuelles de recensement 2005 à 2017, Insee.
Champ : Personnes déclarant vivre en couple.

Tableau 8 : Estimation des effectifs de personnes en couples de même sexe en 2017

	Effectifs		En % des personnes en couple	
	CMS apparents	Estimation "vrais" CMS	CMS apparents	Estimation "vrais" CMS
Couples d'hommes	203 000	142 000	0,67	0,47
Couples de femmes	148 000	105 000	0,49	0,35
Ensemble CMS	351 000	247 000	1,16	0,82

Source : EAR 2017, Insee.

Bibliographie

[1] Banens Maks, Le Penven Eric, « Les erreurs de sexe dans le recensement et leurs effets sur l'estimation des couples de même sexe », *Population*, 2016/1 (Vol. 71), p. 135-148.

[2] Bodier M. et al. (coord), Couples et familles, Insee Références, 2015.

[3] Breuil-Genier Pascale, Buisson Guillemette, Robert-Bobée Isabelle, Trabut Loïc, « Enquête Famille et logements adossée au recensement de 2011 : comment s'adapter à la nouvelle méthodologie des enquêtes annuelles et quels apports ? », Economie et Statistiques, 2016, n°483-484-485.

[4] Buisson Guillemette, Lapinte Aude, Le couple dans tous ses états. Non-cohabitation, conjoints de même sexe, pacs... », Insee Première n°1432, Insee, 2013.

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

[6] Cortina Clara, Festy Patrick, 2014, « Identification of same-sex couples and families in censuses, registers and surveys », Families and Societies Working paper series 8, 27 p.

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

[8] Festy Patrick, 2007, « Enumerating same-sex couples in censuses and population registers », Demographic Research, 17(12), p. 339-368.

[9] Godinot Alain, Durr Jean-Michel. Avant-Propos. La rénovation du *Recensement de la population*. In: *Economie et statistique*, n°483-485, 2016. Le *Recensement* rénové : avancées méthodologiques et apports à la connaissance. pp. 7-14.

[10] Godinot Alain, « La rénovation du recensement de la population », revue Courrier des statistiques n°105-106, juin 2003, Insee.

[11] Howard Hogan, Martin O'Connell and Sarah Feliz, Same Sex Households in the United States Census:Measurement Issues and Substantive Results, U.S. Bureau of the Census

[12] Kreider Rose M., Bates Nancy, and Yerís Mayol-García, Improving Measurement of Same-Sex Couple Households in Census Bureau Surveys: Results from Recent Tests, SEHSD Working Paper 2017-28.

[13] Kreider, Rose M., and Daphne A. Lofquist. 2015. "Matching Survey Data with Administrative Records to Evaluate Reports of Same-Sex Married Couple Households." SEHSD Working Paper, 2014-36. U.S. Census Bureau: Washington, DC, available online at: https://www.census.gov/library/working-papers/2015/demo/SEHSD-WP2014-36.html

[14] Lathe Heather, Ménard France-Pascale, Martel Laurent, Hallman Stacey, "Les couples de même sexe au Canada en 2016", Le recensement en bref, Statistiques Canada, No 98-200-X2016007, 2017.

[15] O'Connell Martin, Feliz Sarah, Same-Sex Couple Household Statistics from the 2010 Census**,** SEHSD Working Paper, 2011-26.

[16] Rault Wilfried, 2016b, « 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.

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

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

[19] Robin Xavier, Turck Natacha, Hainard Alexandre, Tiberti Natalia, Lisacek Frédérique, Sanchez Jean-Charles and Müller Markus (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, p. 77. DOI: 10.1186/1471-2105-12-77.

[20] Toulemon Laurent, Vitrac Julie, Cassan Francine, 2005, « Le difficile comptage des couples homosexuels d'après l'enquête EHF », in Lefèvre Cécile, Fillon Alexandra (dir.), Histoires de familles, histoires familiales. Les résultats de l'enquête Famille de 1999, Ined, Cahier n° 156, p. 589-602.

Encadré 1 : Sensibilité, spécificité et détermination du seuil optimal

Dans l'échantillon démographique permanent, il est possible d'évaluer les dictionnaires et de les comparer de façon similaire à ce qu'on utiliserait pour un test diagnostique en épidémiologie.

Schéma des indicateurs pour un test diagnostique en épidémiologie.

	Le test indique la nécessité d'une correction ?	
Erreur avérée de codage sur le sexe ?	Oui	Non
Oui	a	b
Non	c	d

La sensibilité est la proportion d'erreurs effectivement repérée (a / a+b). La spécificité mesure la capacité d'un test à donner un résultat négatif lorsque l'hypothèse n'est pas vérifiée (d / d+c). Pour comparer différents tests, un des indicateurs possibles est la somme de la sensibilité et de la spécificité. Plus cette somme est élevée, plus le test est de bonne qualité.
Ce critère permet de choisir pour un dictionnaire donné le seuil à partir duquel on décide de corriger le sexe : le meilleur seuil est celui qui permet d'atteindre les sensibilités et spécificités les plus élevées. Dans cette optique, on représente la sensibilité et de la spécificité aux différents seuils sur la courbe ROC (Robin, 2011) : pour chacun des seuils possibles, un point figure la spécificité et la sensibilité. Le meilleur point est ensuite choisi. Différents critères de choix peuvent être retenus. Une approche prudente, averse à la correction fera privilégier un seuil de correction très élevé (on ne corrige que si 99 % des personnes enquêtées sont d'un sexe opposé à celui de la personne enquêtée). Dans ce cas, priorité est donnée à une spécificité élevée. À l'inverse, une volonté de repérer toutes les erreurs (priorité à la sensibilité), quitte à faire des corrections à tort, fera abaisser le seuil. Ici, le critère retenu est de choisir le seuil qui maximise la somme de la spécificité et de la sensibilité.

Une fois le seuil optimal choisi pour chaque dictionnaire, il est possible de comparer leurs performances, c'est-à-dire la sensibilité et la spécificité associées pour chaque dictionnaire à son seuil optimal.

Encadré 2 : La construction des dictionnaires et leur comparaison

- **Sources et critères de construction**

Un dictionnaire de prénoms associe chaque prénom à la proportion de femmes (respectivement d'hommes) le portant, pour apparier cette information aux prénoms des personnes enquêtées dans le recensement et comparer le sexe déclaré par les enquêtés au sexe le plus fréquemment associé à ce prénom. La finalité est de repérer les cas d'erreur de codage les plus probables. Il existe de nombreuses façons de construire le dictionnaire et différents tests ont été menés afin de choisir le meilleur.

Deux sources étaient disponibles que nous avons combinées. La source privilégiée a priori pour constituer les dictionnaires a été le fichier des prénoms de l'état civil, qui couvre l'ensemble des prénoms donnés à des enfants nés en France depuis 1900. Y manquent donc potentiellement des prénoms de personnes résidant en France et nées à l'étranger. C'est pourquoi ces dictionnaires ont été complétés en recourant à l'enquête annuelle de recensement 2017, qui couvre l'ensemble des personnes résidant en France en 2017).

Par exemple, l'entrée OCEANE – 1990 du dictionnaire 1 est construite en ajoutant aux 883 personnes prénommées Océane nées en 1990 d'après l'Etat-civil (toutes des femmes), les 4 personnes nées à l'étranger et recensées en 2017 (un homme et trois femmes). Au total, l'entrée comprend donc 887 personnes, et la proportion de femmes est de 886 / 887, soit 100 %.
Lorsqu'un prénom présent à l'EAR 2017 était complètement absent de l'état-civil, l'entrée a été ajoutée. Par exemple, « EPOUSE » ou « JLUC » sont des « prénoms » qui apparaissent au recensement (réponse en clair) mais évidemment jamais à l'état-civil. Or ils apportent une information sur le sexe de leurs porteurs.

Une entrée n'est maintenue dans un dictionnaire qui si elle repose sur au moins 10 personnes. Par exemple, seulement 9 personnes prénommées Océane et nées en 1971 sont trouvées en ajoutant état-cil et personnes recensées nées à l'étranger en 1971. Cette entrée n'apparaît pas dans le dictionnaire et une Océane née en 1971 sera en échec d'appariement pour le dictionnaire 1. En revanche, dans le dictionnaire 2, qui fait abstraction de l'année de naissance, cette même personne pourra être appariée. Plus les critères sont précis, plus le taux d'échecs est donc élevé. Cela nous a conduit à renoncer à certains critères comme le pays de naissance, trop précis.
Les dictionnaires présentés ici diffèrent également selon la façon de traiter le prénom, soit en le prenant en compte dans son intégralité soit sa première partie. Dans le premier cas (dictionnaires 1 et 2), des entrées « Marie Cécile » et « Marie Héloïse » figurent dans le dictionnaire si les occurrences sont en nombre suffisant et les personnes ainsi prénommées ne seront appariées qu'avec cette occurrence. Dans le second cas (dictionnaires 3 et 4), seule la première partie « Marie » est prise en compte. L'appariement, même pour « Marie Cécile » et « Marie Héloïse », se fait avec l'entrée « Marie », qui regroupe toutes les occurrences de Marie en première (et souvent seule) partie du prénom.

Enfin, le dictionnaire 5 est une combinaison construite *a posteriori*, par étapes : s'il n'y a pas d'appariement dans le dictionnaire 1, on prend le 3 (première partie du prénom) puis le 2 (prénom mais sans année de naissance) et enfin le 4 (première partie du prénom, sans année de naissance). L'idée de cet ordre est de chercher un appariement dans le dictionnaire le plus précis, qui a des chances de mieux correspondre à la situation (donc celui basé sur le prénom complet et l'année de naissance). S'il n'y a pas de correspondant dans ce dictionnaire, on regarde dans le dictionnaire basé sur l'année de naissance et le prénom simplifié, et ainsi de suite.

Tableau 1 : Caractéristiques des dictionnaires construits, hors mode de collecte

Numéro	Prénom	Année de naissance	Nombre d'entrées	Taux d'échecs	
				Ensemble	Personnes nées à l'étranger
1	Ensemble	Oui	247 773	14,6	34,3
2	Ensemble	Non	34 549	8,1	17,2
3	Première partie	Oui	239 862	6,7	20,1
4	Première partie	Non	28 580	1,3	4,9
5	Combinaison*			1,3	4,9

*** : « combinaison » signale que le dictionnaire est construit en cherchant d'abord dans le dictionnaire le plus détaillé puis en cas d'échec dans un qui l'est moins**

- **Prise en compte du mode de collecte**

Compte tenu de la nécessité d'éviter des corrections excessives pour la collecte papier (voir encadré 3), les dictionnaires finalement retenus sont construits séparément pour chaque mode de collecte.

Pour la partie internet, l'état-civil sert de point de départ et on y ajoute les occurrences de personnes de l'EAR 2017 nées à l'étranger lorsqu'elles ont été recensées par internet. Pour reprendre l'exemple, précédent, l'entrée OCEANE – 1990 du dictionnaire 1 internet est construite en ajoutant aux 883 personnes prénommées Océane nées en 1990 d'après l'état civil (toutes des femmes), l'unique femme née à l'étranger et recensée par internet. La proportion par sexe n'est en pratique pas affectée par cet ajout.

Pour la partie papier, priorité est accordée aux données issues de la collecte papier, puis en cas de défaut d'appariement, on prend le dictionnaire construit pour la collecte internet. Dans le cas des Océane de 1990, 3 personnes ont été recensées sur papier, ce qui est insuffisant pour que l'entrée apparaisse. C'est donc l'entrée du dictionnaire précédent qui sera prise en compte.

Pour le dictionnaire 5, la combinaison se fait en cherchant un appariement par ordre de priorité dans les dictionnaires suivants : **1 papier > 3 papier > 2 papier > 4 papier > 1 internet > 3 internet > 2 internet > 4 internet.**

Pour la partie papier, donner priorité aux dictionnaires basés sur la collecte papier permet de prendre en compte les prénoms affectés trop fréquemment lors de l'acquisition (comme PEGGY, voir encadré 3). Une fois ces problèmes spécifiques réglés, l'usage des dictionnaires appuyés sur l'état civil et la collecte internet ne pose plus de difficultés.

Tableau 2 : Taux d'échec des dictionnaires construits, avec mode de collecte

Numéro	Prénom	Taux d'échecs-EAR 2017 INTERNET		Taux d'échecs-EAR 2017 PAPIER	
		Ensemble	Personnes nées à l'étranger	Ensemble	Personnes nées à l'étranger
1	Ensemble	7,1	36,2	22,5	33,3
2	Ensemble	3,1	15,5	13,8	18,9
3	Première partie	5,6	31,1	7,2	12,5
4	Première partie	2,0	10,2	1,2	1,9
5	Combinaison	2,0	10,2	1,2	1,9

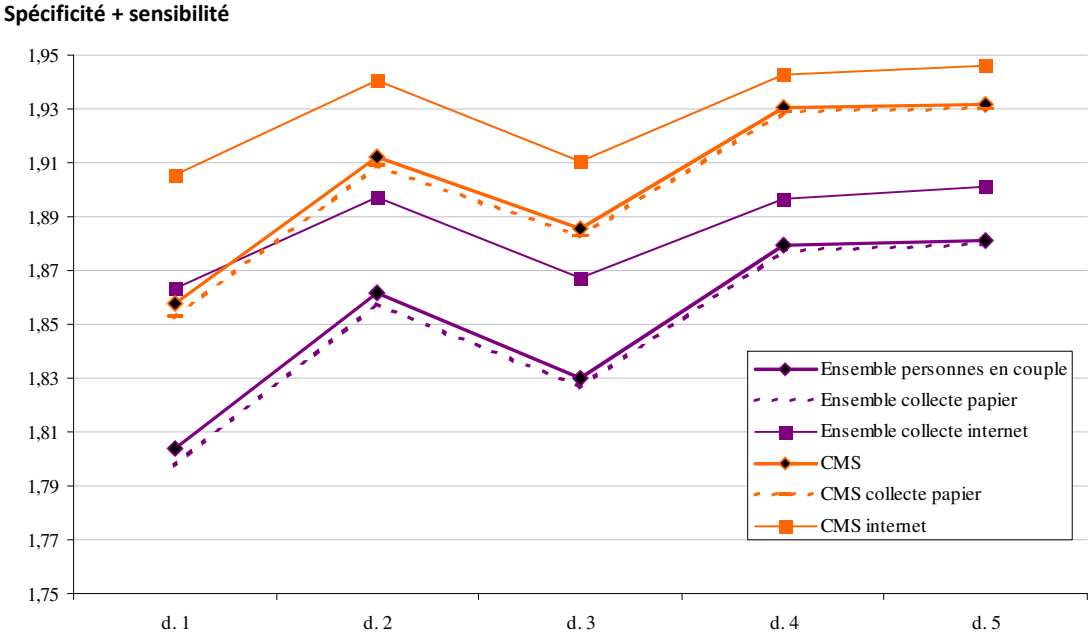
Les dictionnaires construits en tenant compte du mode de collecte comprennent plus d'échecs d'appariement côté internet. En effet, on s'interdit de prendre en compte les occurrences de

prénoms collectées via papier et qui pouvaient permettre de calculer une proportion de femmes sur les répondants à l'EAR 2017 lorsque le prénom est absent de l'état civil. On a vu toutefois que cette façon de faire est risquée compte tenu des spécificités des prénoms collectés sur papier.

- **Choix du dictionnaire**

Le critère principal pour choisir le dictionnaire appliqué a été sa capacité à repérer les erreurs de codage du sexe avérées dans l'EDP. Après avoir déterminé le seuil optimal de correction, on compare la somme des spécificités et sensibilités de chaque dictionnaire.
On constate en premier lieu que l'application est de bonne qualité, en particulier quand on se restreint aux CMS apparents : l'indicateur se rapproche de 2 qui est le maximum (sensibilité et spécificité à 100 %). Par ailleurs, les écarts sont faibles : on choisit de retenir le dictionnaire 5 car il semble très légèrement supérieur aux dictionnaires 2 et 4. Les dictionnaires 1 et 3 présentent de moins bonnes performances du fait de la prise en compte de l'année de naissance, qui augmente les échecs d'appariement.

Graphique : Comparaison de la performance des 5 dictionnaires appliqués à différentes populations de l'EDP



Source : EDP, base Étude 2010-2016, Insee

Encadré 3 : Pourquoi prendre en compte le mode de collecte ?

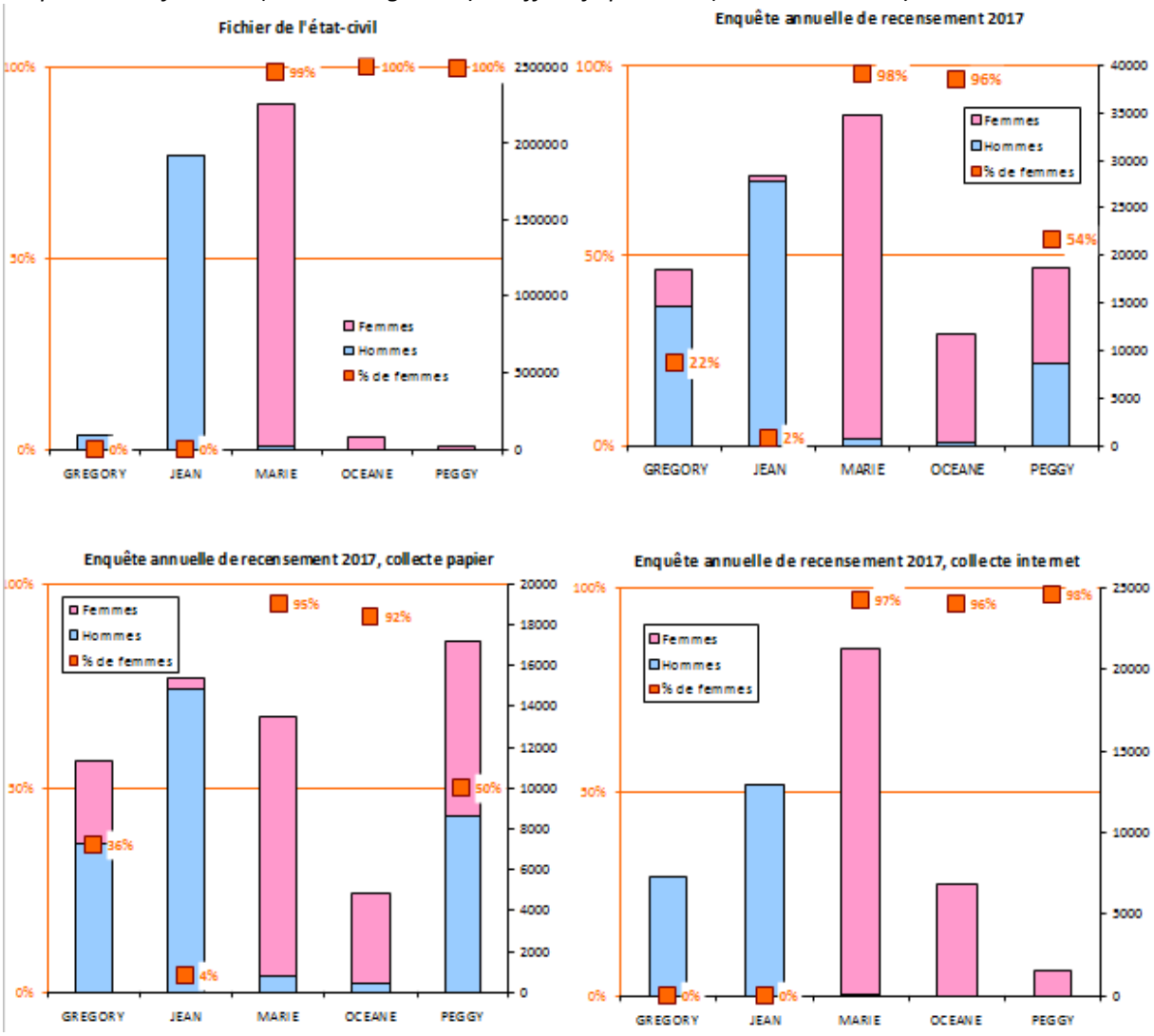
Les différences de comportement d'un même dictionnaire appliqué à l'enquête annuelle de recensement et à l'échantillon démographique permanent sont à première vue surprenantes, car l'EDP est un extrait du recensement. Mais les écarts sont liés au mode de collecte : c'est le comportement des dictionnaires confrontés à la collecte papier de l'EAR 2017 qui se distingue, tandis qu'entre l'EDP et la collecte internet de l'EAR 2017, les résultats sont similaires. La différence entre EAR et EDP s'explique par un écart de qualité important dans la saisie des noms et prénoms en cas de collecte du recensement sur des questionnaires papier. Le seuil d'exigence fixé au prestataire de saisie pour les individus de l'EDP est de 1 % d'erreurs, il est de 50 % d'erreurs pour les autres individus recensés. Cette différence s'explique par les finalités de la saisie[12]. Le prestataire code le prénom le plus souvent à partir d'une saisie optique, d'une reconnaissance de caractère et de l'affectation du prénom le plus proche dans un dictionnaire. Pour l'EDP ce processus est fréquemment suivi d'une vérification manuelle des informations telles qu'elles figurent effectivement sur les documents de collecte papier (visionnage de l'image des bulletins). Cela permet de corriger les cas litigieux plus précisément en cas de doute.
Du côté de la collecte internet, le traitement des prénoms déclarés est similaire entre individus EDP et autres individus recensés. Il peut y avoir de la non-réponse partielle sur le prénom ou des réponses inadaptées (un surnom par exemple). Les prénoms collectés sont plus variés et peuvent être affectés par des fautes d'orthographe (alors qu'un prénom choisi dans un dictionnaire ne l'est pas en principe).
L'exemple du prénom PEGGY montre que des incohérences entre sexe et prénom peuvent se former, au cours de la collecte et surtout de la saisie, et pourquoi il faut adapter le dictionnaire en conséquence. Pour le prénom PEGGY, la proportion de femmes est de 100 % dans l'état-civil, et 54 % à l'enquête annuelle de recensement 2017. Cette dernière proportion recouvre une forte différence selon le mode de collecte : 98 % des PEGGY recensé(e)s sur internet sont des femmes, 50 % des PEGGY recensé(e)s sur papier. Par ailleurs, l'effectif recensé de PEGGY paraît trop élevé sur papier (l'effectif des PEGGY est plus élevé que celui des MARIE par exemple, ce qui n'est pas vrai pour la collecte internet) : il semble qu'un ou plusieurs autres prénoms, masculins, et dont la graphie visuellement ressemble à PEGGY, doivent être codés en PEGGY lors de la saisie optique. Dans ce cas, forcer l'appariement avec l'entrée PEGGY d'un dictionnaire construit avec le fichier des prénoms de l'état-civil, conduirait à de fortes proportions de corrections sur le sexe : tous les PEGGY déclarés hommes et recensés sur papier se verraient corriger en femmes car l'état-civil indique une proportion de 100 % de femmes. Or il s'agit bien plus vraisemblablement d'erreurs de codage du prénom. Pour parer à cette difficulté, et prendre en compte au plus près le processus de codage du prénom, il a été choisi d'apparier en priorité les individus recensés sur papier avec un dictionnaire créé à partir de la collecte papier de l'EAR. Ainsi, sur papier, le prénom MARIE porté par un homme fera suspecter une erreur de codage du sexe (proportion de femmes de 95% d'après l'EAR papier), tandis que le sexe d'un répondant codé comme homme et ayant pour prénom PEGGY ne sera pas corrigé (proportion de femmes de 50 % seulement, pas assez élevée pour présumer d'une erreur sur le sexe). En revanche, si la personne a répondu sur internet, son sexe sera corrigé.
Cela permet de s'adapter au processus qui a conduit à avoir ce prénom dans la base et qui crée ou non du bruit observable sur les données passées. En revanche, une évolution des traitements effectués par le prestataire responsable de la saisie des prénoms peut conduire à des erreurs (si par exemple il ajoute à son dictionnaire les prénoms masculins souvent confondus avec PEGGY). La

[12]Pour les individus EDP, l'enjeu est d'avoir suffisamment d'information pour coder le NIR, identifiant des individus au répertoire national d'identification des personnes physiques – RNIPP. C'est la condition nécessaire aux appariements dans l'EDP entre les sources de données statistiques l'alimentant (dont les données statistiques du recensement de la population). Pour les autres individus, il s'agit « seulement » de pouvoir discriminer les cas litigieux d'appariement dans un même logement entre d'une part les individus (bulletins individuels) et d'autre part les informations les concernant dans la feuille de logement. Si deux personnes de même sexe sont nées la même année, on recourt au prénom et au nom pour apparier les individus listés sur la feuille de logement et les bulletins individuels remplis. Une valeur approchée suffit généralement à trancher.

méthode présentée nécessite donc d'être actualisée chaque année et de consulter régulièrement le prestataire pour connaître les modifications qu'il met en œuvre.

Graphique : Répartition par sexe des Grégory, Jean, Marie, Océane et Peggy dans le fichier de l'état-civil et dans l'enquête annuelle de recensement 2017 selon le mode de collecte

Proportion de femmes (échelle de gauche) et effectifs par sexe (échelle de droite)



Note : On a choisi ici des exemples contrastés : deux prénoms très courants (MARIE et JEAN) et deux prénoms (PEGGY et GREGORY) pour lesquels les incohérences sont particulièrement fréquentes entre le sexe du recensement et le sexe le plus souvent associé au prénom dans l'état-civil.