Direction des Statistiques Démographiques et Sociales

F1807

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 et Sébastien Hallépée

Document de travail

Institut National de la Statistique et des Études Économiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Série des Documents de Travail

de la Direction des Statistiques Démographiques Et Sociales

F1807

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* ET SÉBASTIEN HALLÉPÉE**

Document de travail

septembre 2018

* Insee, elisabeth.algava@insee.fr
** Insee, sebastien.hallepee@insee.fr

Les auteurs tiennent à remercier Adrien Pons, qui a préparé la mise en œuvre de la procédure de correction dans les futures enquêtes de recensement, Sébastien Durier qui nous a facilité l'accès et l'utilisation de l'échantillon démographique permanent, ainsi qu'Isabelle Robert-Bobée et Marie Reynaud pour leur aide et conseils.

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. Ces couples sont en effet peu nombreux et une erreur de codage sur le sexe, même de faible ampleur, conduit à une erreur relative importante sur le total. Or, dans le cadre des règlements européens sur le recensement, la France est engagée à fournir des données sur ces personnes pour le recensement européen de 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 et rendra aussi possible la réalisation d'analyses nouvelles sur cette population.

Pour 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, il est envisagé d'ajouter dans les chaînes de traitements du recensement une nouvelle variable individuelle calculée. Elle indiquerait dans quelle proportion le prénom déclaré est plutôt masculin ou féminin et 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.

Après avoir présenté 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, le document présente les expériences et solutions mises en œuvre à l'étranger. La partie suivante décrit la procédure envisagée. En utilisant l'échantillon démographique permanent, elle commence par démontrer la très grande efficacité de la méthode des prénoms pour repérer les erreurs de codage du sexe, sur un échantillon (l'EDP) dans lequel les erreurs 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.

Mots-clés : Recensement, couple, couple de même sexe, contrôle de cohérence, imputation

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 likely 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.

Key-words : Census, couple, same-sex partnership, data consistency, data imputation

Sommaire

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 en surestimer le nombre, comme l'a montré le travail réalisé à partir de l'enquête Famille et logements, associée au recensement de 2011.

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, dotée d'un protocole spécifique sur les couples de même sexe (cf 1.4), 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 issue de 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 contrôlé du nombre de couples 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 européen de 2021. Les repérer de façon plus fiable dans le recensement permettra d'apporter une réponse de qualité à l'institut européen de statistiques. En tirant profit du recensement, et donc d'une collecte annuelle d'information auprès de plusieurs millions de personnes et de logements (Godinot, 2016), cette amélioration rendra aussi possible la réalisation d'analyses nouvelles sur cette population, relativement rare actuellement. 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 plus difficiles à mettre en œuvre (coût, délai, sécurisation des données).

La partie suivante décrit la procédure de correction envisagée. 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 ces erreurs 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 la procédure avant son application au recensement. Il s'agit principalement de traiter séparément la collecte papier et la collecte internet.

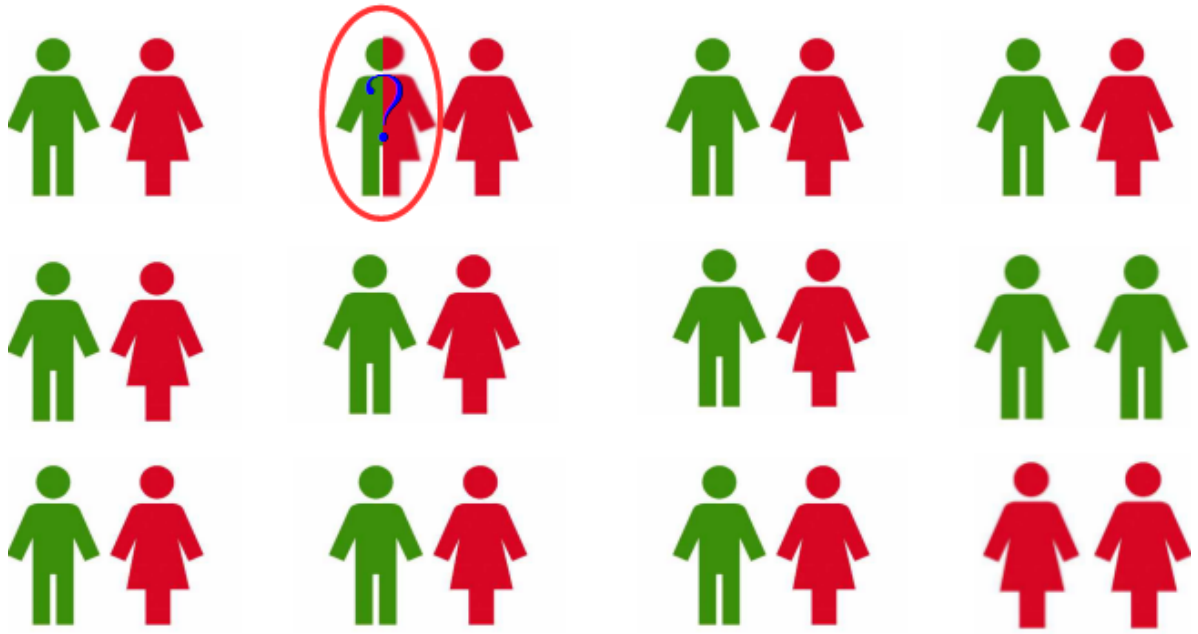
Enfin, la dernière partie présente les résultats obtenus en appliquant la solution retenue à l'échantillon démographique permanent et à l'EAR 2017. Cela permet de vérifier la cohérence de ces résultats, autant en termes d'effectifs que d'évolution.

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

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 de déclaration ou de codage sur le sexe d'un seul des conjoints aboutit en général à compter ce couple comme étant de même sexe. Si cela ne concerne qu'une toute petite proportion des personnes en couple de sexe différent, cela suffit à surestimer très fortement la proportion de personnes en CMS. Ce risque 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. Mais la particularité est qu'ici une erreur sur le sexe au niveau individuel, pour un seul des conjoints, va conduire à considérer les deux conjoints comme ayant un partenaire de même sexe (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 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.

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).

1.1 Enquête annuelle de recensement et recensement de la population

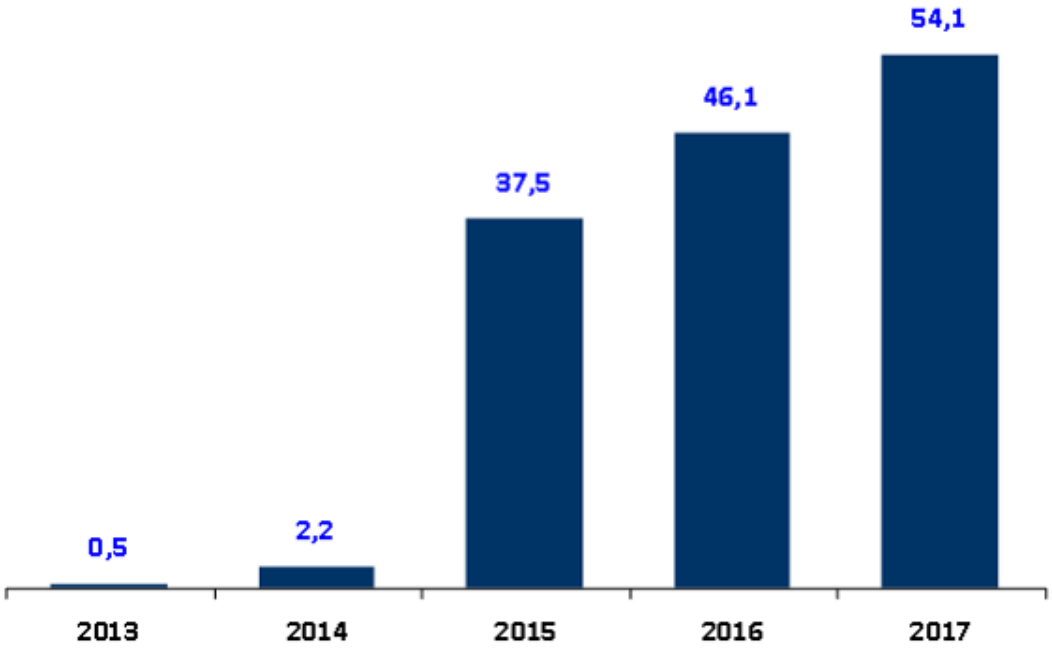
Depuis 2004, le recensement français est une enquête sur un échantillon. La collecte est annuelle et effectuée au moyen de questionnaires papier ou internet, directement remplis par les personnes concernées. Chaque année, une « petite commune » (moins de 10 000 habitants) sur cinq est recensée de façon 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[3]. 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 des bulletins papier 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).

3 Plus exactement, des contraintes de dates de disponibilité des données conduiront à prendre en compte en plus la collecte annuelle 2018 pour établir les statistiques transmises à Eurostat.

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, Insee

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.

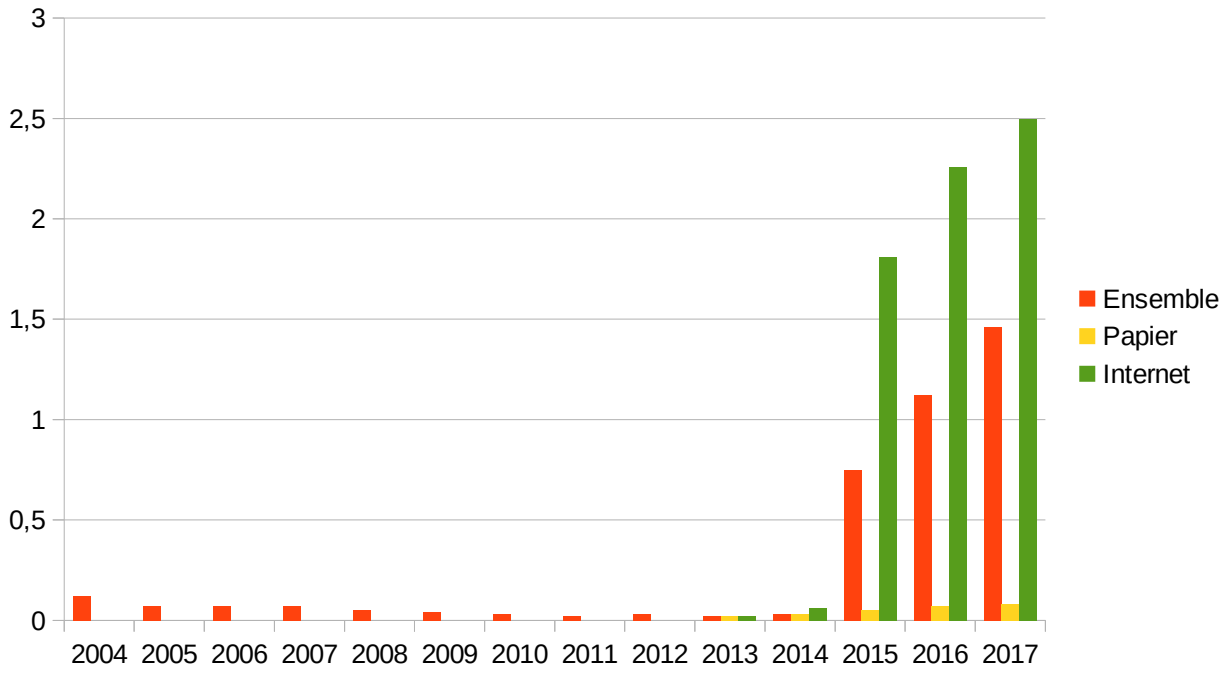
Sur le sexe, les enquêtés ont une case à cocher, homme ou femme, et les erreurs sont très rares au cours de la collecte papier. La différence entre les deux modes de collecte est très faible. Toutefois, la non-réponse sur papier est corrigée directement au moment de la saisie des questionnaires lorsque 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, Insee

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 pour des raisons de confidentialité, 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

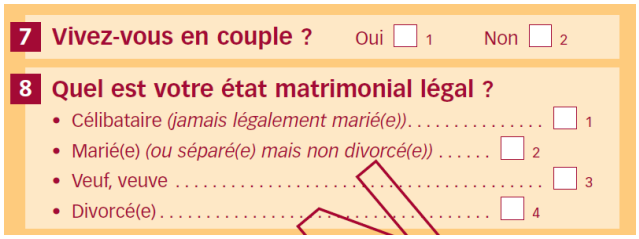
Pour chaque logement recensé, l'agent recenseur doit collecter une feuille de logement et un bulletin individuel par habitant du logement. La feuille de logement est le premier document renseigné, elle permet de décrire les relations entre les habitants du logement. Par construction, cela limite l'analyse 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 le même logement. Cela limite l'appréhension des différentes formes de conjugalité[4]. 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 même parmi les CMS (84 % en 2011, voir Buisson et Lapinte, 2013).

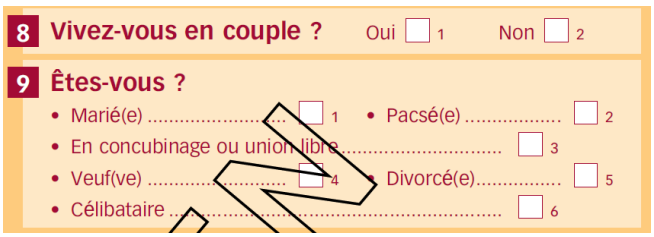
Une refonte du bulletin individuel a eu lieu en 2015, et celle de la feuille de logement en 2018. Ces deux refontes affectent la façon d'appréhender les relations conjugales dans le recensement.
Le changement dans le bulletin individuel de la question sur l'état matrimonial légal (marié, divorcé, veuf, 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 (conjoint, enfant, etc.) des différents habitants dans la famille. L'AMF a peu évolué de 2004 à sa refonte en 2018, si ce n'est pour prendre en compte les couples de même sexe à partir de 2015.
Avant la refonte, 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[5].

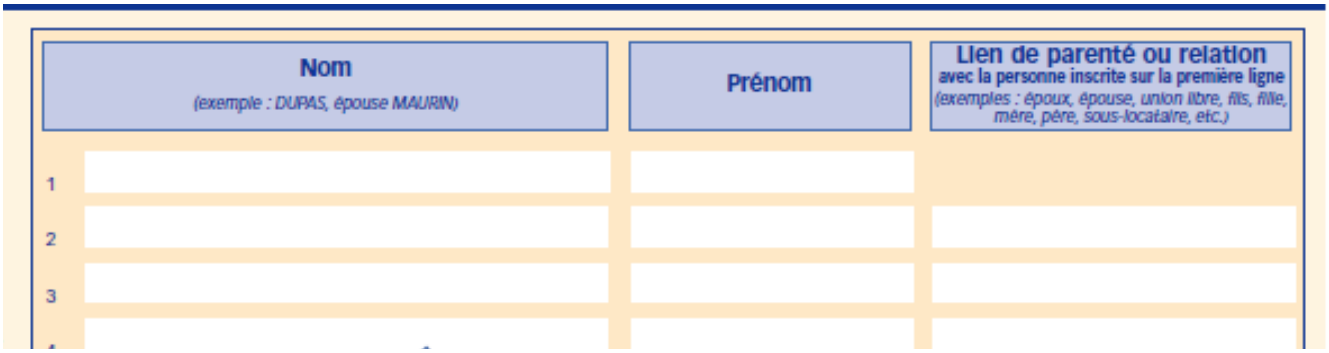
4 Certaines unions non cohabitantes peuvent être repérées lorsqu'une personne déclare vivre en couple sur son bulletin individuel sans conjoint dans le logement. Néanmoins, en l'absence de question sur le sexe du conjoint, il est impossible de savoir s'il s'agit d'unions non cohabitantes entre personnes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

5 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

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 l'année 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



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 et al., 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

En milliers		***Situation d'après les réponses à EFL consolidées***				
		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 *(c)*	155 *(b)*	2	12	283 *(a)*
	Seule personne en couple	212	4	9	526	751
	Situation complexe	180	1	0	61	242
	Total	30 265	173 *(d)*	32	17 361	47 831

CSO = couple de sexe opposé, CMS : couple de même sexe

Champ : Personnes majeures, France métropolitaine

Source : Enquête Familles et Logements 2011, Insee, données pondérées.

Lecture : En 2011, d'après les informations des bulletins individuels de recensement, 283 000 personnes vivaient apparemment en CMS (a). Parmi eux, 155 000 ont été confirmés comme « vrais » CMS lors de l'enquête EFL 2011 (b), tandis que 114 000 étaient requalifiés en CSO (c). Au total, il y avait 173 000 personnes en CMS cohabitant d'après l'EFL (d), dont 155 000 identifiées avec l'indicateur de CMS apparent.

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[6]. 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 :

« **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 ?[7] 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

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

7 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 ».

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

Une expérience de validation du sexe déclaré et codé par appariement à des données administratives a été mise en œuvre aux États-Unis (Kreider, 2015). En appariant les données du recensement avec le registre de la sécurité sociale, les auteurs relèvent des incohérences bien plus fréquentes 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[8].

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, plus fréquemment s'ils sont mariés (69%) qu'en union libre (21%).
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.

8 Nous reprenons cette démarche dans la suite du document (cf 3.1) en l'appliquant aux 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 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. L'EDP est en revanche 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 par la suite, de valider la solution retenue par les prénoms.

3 Le mode de correction envisagé en France

Pour repérer les erreurs de codage sur le sexe et ainsi distinguer les faux CMS au sein des CMS apparents, l'exemple américain suggère qu'il est assez efficace d'utiliser le prénom. Le fait que le prénom soit à disposition de l'Insee pour mener des contrôles de qualité de la collecte du recensement depuis 2016 nous oriente vers cette solution articulée principalement sur une « validation statistique » par le prénom. Pour la mettre en place , il faut en vérifier l'efficacité dans le cas français et 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 sexe déclarée par le répondant au recensement.

L'échantillon démographique permanent (EDP) est a priori la source idéale pour faire les vérifications et choisir les meilleurs paramètres, car il offre la possibilité de tester la capacité de la procédure à repérer individuellement les erreurs de codage du sexe sur un échantillon pour lequel ces erreurs sont connues (partie 3.1).
L'idée générale était initialement d'appliquer la procédure ainsi validée et définie à partir de l'EDP sur les enquêtes annuelles de recensement, en premier lieu l'EAR 2017, en supposant qu'elle pourrait être transposée avec la même efficacité. Toutefois, les résultats obtenus lors de cette transposition nous ont alertés sur la nécessité de prendre en compte le mode de collecte et d'adapter la procédure aux spécificités de la collecte papier de l'EAR (partie 3.2)

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

- **Le « vrai » sexe est connu dans l'échantillon démographique permanent**

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 qui alimentent la base de l'EDP 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, informant sur les inscriptions sur liste électorale ;
- le panel " tous salariés " depuis 1967, qui contient des informations sur les rémunérations perçues ;
- depuis 2011 les données sur les revenus, les aides perçues, les impôts et les pensions versées par le ménage de l'individu grâce au fichier démographique sur les logements et les individus (Fidéli) et au fichier sur les revenus localisés sociaux et fiscaux (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 bien entendu le sexe déclaré lors du recensement, mais aussi le prénom, disponible dans la base de production afin de faciliter l'appariement avec les autres sources (Le prénom est en revanche ab-

sent de la base études à finalité statistique pour éviter une identification directe des personnes). Il est donc possible de construire un dictionnaire qui indique pour chaque prénom, la proportion de femmes ou d'hommes le portant, de la même façon que cette proportion sera calculée dans la chaîne de production des 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. On dispose ainsi du « vrai » sexe, celui qui figure au RNIPP.

En comparant le sexe déclaré au recensement et celui enregistré dans le RNIPP, il est donc possible de déceler les erreurs de codage du sexe au recensement : on considère que le sexe codé dans le RNIPP est le vrai et qu'en cas de discordance, il y a une erreur dans le bulletin de recensement.

- **D'après l'EDP, il y a 0,2 % d'erreurs de codage sur le sexe déclaré au recensement**

Dans la base études 2016, on compte 2,8 millions de personnes dites « EDP », c'est-à-dire nées un jour 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 (graphique 3a). C'est donc un phénomène très rare. Le taux est légèrement plus faible pour les personnes en couple (0,17 %). 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 (graphique 3c). 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 (graphique 3b). La proportion d'erreurs de codage est plus importante parmi les personnes apparemment en CMS et qui se sont déclarées mariées au recensement.

Il faut noter que dans un couple, on recherche le NIR uniquement pour la personne 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 – (on notera cette proportion d'erreur de codage du sexe p_e) – alors on peut estimer, en supposant que les erreurs de sexe entre deux conjoints sont indépendantes, que 0,31 % des couples seraient affectés par une erreur de codage du sexe de l'un des conjoints qui conduit, la plupart du temps, à les compter par erreur comme CMS. Plus précisément, l'erreur de codage concernera un seul membre du couple pour une proportion ($2*p_e - p_e^2$) des couples, soit 0,31 % dans le cas présent et concernera les deux membres du couple pour p_e^2 des couples, soit 0,03 % dans le cas présent. D'après l'enquête Familles et Logements, les couples de sexes opposés représentent 99,4 % des couples cohabitants. Si l'on applique les taux d'erreurs observés dans l'EDP à la structure des couples (répartition entre CMS et CSO) mesurée dans l'enquête Famille et Logements, on observe que la plupart des erreurs transforment des couples de sexes opposés en couple de même sexe. La proportion de CMS apparents à laquelle on s'attend dans l'EAR, 0,9 %, est très proche de ce qui était observé dans l'enquête Famille et Logements.

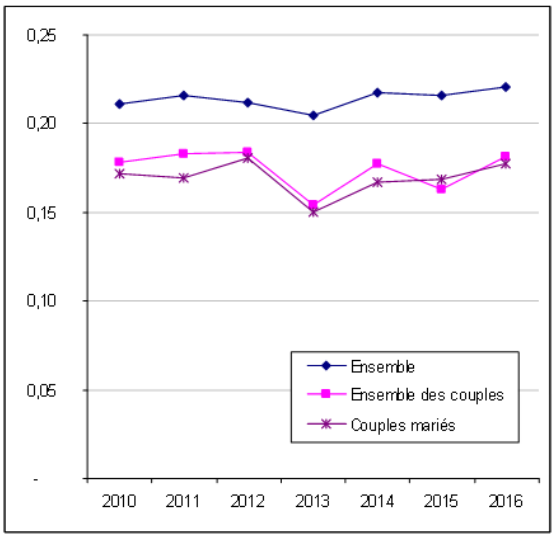
Tableau 2 : Conséquence anticipée des erreurs de codage du sexe au niveau des couples

En %	EFL	Affecté au recensement par...			=> situation apparente à l'EAR
		1 erreur de codage du sexe	2 erreurs de codage du sexe	Aucune erreur de codage du sexe	
CMS	0,6	*0,0019*	**0,0002**	**0,598**	**0,906**
CSO	99,4	**0,3078**	*0,0287*	*99,0634*	*99,094*
Ensembles des couples	100	0,3097	0,0289	99,6614	100

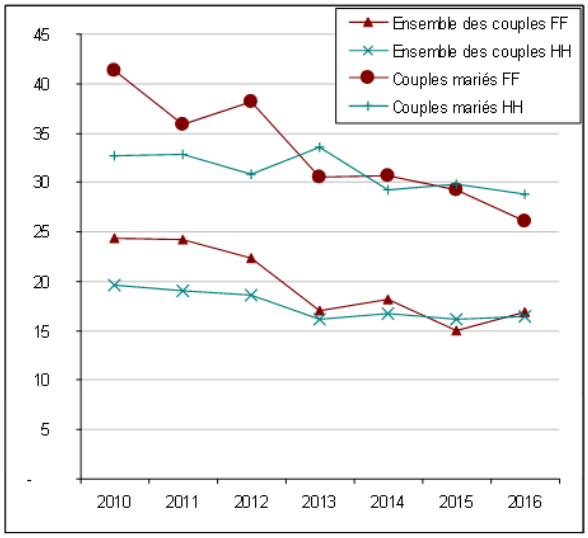
Source : Enquête Famille et Logements 2011, base étude 2016 de l'EDP, Insee
Champ : Couples cohabitants
Lecture : Selon l'Enquête Famille et Logements, 99,4 % des couples sont des CSO. Avec une erreur de codage affectant 0,17 % des personnes, lorsqu'on cherche à les recenser, 0,3078 % ont un seul des membres du couple affecté d'une erreur de codage du sexe. Ainsi, 0,3078 % des couples passent d'une situation réelle de CSO à une situation apparente de CMS. 0,0287 % des couples ont les deux membres du couple affectés d'une erreur de codage sur le sexe. Même si aucun des membres du couple n'a un sexe correspondant à la réalité, la situation apparente de leur couple correspond à la réalité : le couple reste un CSO.
Compte tenu des erreurs affectant CSO et CMS, on peut donc attendre que les CMS passent d'une situation réelle où ils représentent 0,6 % des couples à une situation apparente dans l'EAR où ils représentent 0,9 % des couples (somme des chiffres en gras sur fond grisé, à savoir 0,3078 + 0,0002 + 0,598 = 0,906 %). 99,1 % des couples seront apparemment des CSO (somme des chiffres en italiques soit 0,0019 + 0,0287 + 99,0634 = 99,094 %).

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

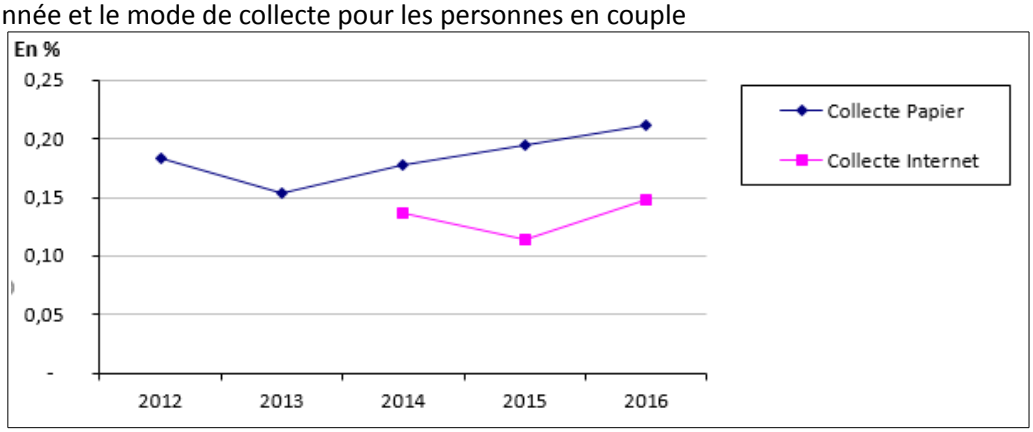
a) Ensemble des individus de l'EDP recensés



b) Personnes vivant apparemment en CMS



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



Source : Base étude 2016 de l'EDP, 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. Il est ensuite apparié aux prénoms des personnes enquêtées dans le recensement (ou une autre enquête) afin de comparer le sexe déclaré par les enquêtés au sexe le plus fréquemment associé à ce prénom. La finalité est de repérer les erreurs de codage les plus probables (voir encadré 1 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 pour les personnes nées en France. Il est important que le dictionnaire repose sur un très grand nombre d'observations de chaque prénom 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. 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 (Voir annexe 1 pour quelques exemples).

L'indicateur finalement affecté à 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 cet indicateur de la proportion de femmes portant le même prénom.

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

Encadré 1 : Quel dictionnaire de prénoms choisir pour tester la procédure dans l'EDP ?

Sources et critères de construction

Un dictionnaire de prénoms associe chaque prénom à la proportion de femmes (respectivement d'hommes) le portant. Il est ensuite apparié avec les prénoms des personnes enquêtées dans le recensement afin de 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 (Annexe 1).
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 ajoutant les prénoms des personnes ayant répondu à l'enquête annuelle de recensement 2017, qui couvre l'ensemble des personnes résidant en France en 2017.
Lorsqu'un prénom présent à l'EAR 2017 était complètement absent de l'état-civil, l'entrée a été aussi 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.
Les dictionnaires présentés ici diffèrent selon la façon de traiter le prénom, soit en le prenant en compte dans son intégralité soit sa première partie.
Les dictionnaires peuvent aussi varier selon qu'on prend en compte ou pas l'année de naissance.

Le dictionnaire A correspond aux prénoms de l'état civil et de l'EAR pour les nés à l'étranger par année de naissance : il associe à chaque prénom par année de naissance la proportion de femmes (respectivement d'hommes) portant ce prénom. Selon la date de naissance, un prénom peut être plus généralement masculin ou féminin (par exemple CAMILLE était autrefois plutôt un prénom d'homme et est actuellement plutôt féminin).
Le dictionnaire B se distingue du dictionnaire A par le fait que l'année de naissance n'est pas prise en compte.
Dans ces deux dictionnaires, le prénom est traité dans son intégralité.
Par exemple, l'entrée « MARIE HELENE » du dictionnaire B est construite en ajoutant aux 37 623 personnes prénommées « MARIE HELENE » d'après l'état-civil (toutes des femmes), les 368 personnes prénommées « MARIE HELENE » nées à l'étranger et recensées en 2017 (17 hommes et 351 femmes). Au total, l'entrée comprend donc 37 991 personnes, et la proportion de femmes est de 37 974 / 37 991, soit pratiquement 100 %.
Ainsi, des entrées « MARIE HELENE » figurent dans les dictionnaires A et B si les occurrences sont en nombre suffisant et les personnes ainsi prénommées ne seront appariées qu'avec cette occurrence.

Les dictionnaires C et D sont analogues aux dictionnaires A et B sauf pour le traitement du prénom. Seule la première partie est prise en compte. L'appariement, pour « MARIE HELENE » se fait avec l'entrée « MARIE», qui regroupe toutes les occurrences de MARIE en première partie du prénom.

Enfin, le dictionnaire E est une combinaison construite *a posteriori*, par étapes : s'il n'y a pas d'appariement dans le dictionnaire A (année de naissance et intégralité du prénom), on prend le C (année de naissance, première partie du prénom) puis le B (intégralité du prénom mais sans l'année de naissance) et enfin le D (première partie du prénom, sans l'année de naissance). L'idée de cet ordre est de chercher d'abord 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.

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 (voir encadré 2). On constate en premier lieu que tous les dictionnaires sont 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 entre dictionnaires sont faibles : on choisit de retenir le dictionnaire E car il semble très légèrement supérieur aux dictionnaires B et D. Les dictionnaires A et C présentent de moins bonnes performances du fait de la prise en compte de l'année de naissance, qui augmente les échecs d'appariement.

Tableau : Caractéristiques des dictionnaires construits, première étape hors mode de collecte

		Taux d'échecs (EAR)		Performances (EDP)
	Nombre d'entrées	Ensemble	Personnes nées à l'étranger	Spécificité + Sensibilité
A Ensemble du prénom / année de naissance	247 773	14,6	34,3	1,858
B Ensemble	34 549	8,1	17,2	1,912
C Première partie du prénom / année de naissance	239 862	6,7	20,1	1,885
D Première partie du prénom	28 580	1,3	4,9	1,930
E Combinaison		1,3	4,9	1,932

* : « 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

Lecture : Lorsque l'on applique à l'EAR 2017 le dictionnaire A, 14,6 % des personnes ont une valeur manquante à l'indicatrice d'erreur : aucune entrée du dictionnaire ne correspond à leur prénom (ensemble du prénom) et leur année de naissance. Appliqué à la base études de l'EDP, le dictionnaire A permet de repérer les erreurs avérées de codage du sexe avec une performance de 1,858, mesurée par la somme de sa spécificité et de sa sensibilité.

Source : EAR 2017, base étude 2016 de l'EDP, Insee.

Encadré 2 : 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 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, sans priorité à l'une ou l'autre.
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.

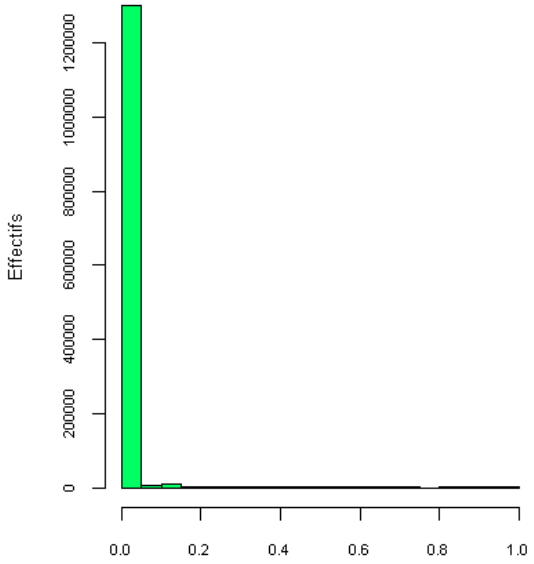
- **Sur les personnes EDP, le repérage des erreurs de codage du sexe par les prénoms est très efficace**

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 le sexe fourni dans l'enquête annuelle de recensement pour les personnes EDP (EAR) et l'information du RNIPP pour ces personnes : 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).

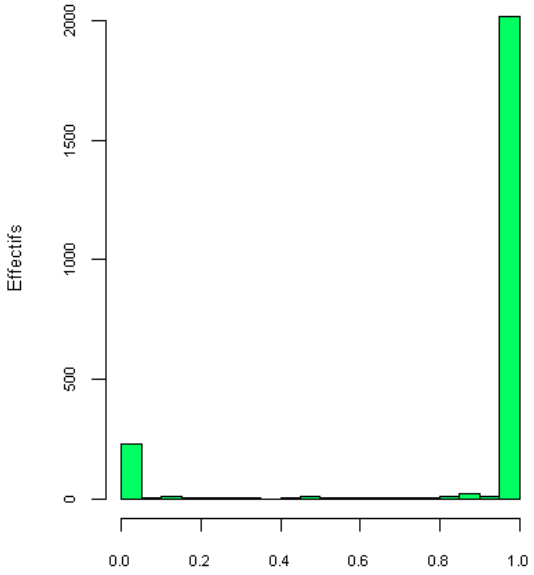
Graphique 4 : Répartition de l'indicateur d'erreur selon la présence d'une erreur avérée

a) Ensemble des personnes EDP en couple (1 343 908 personnes)

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



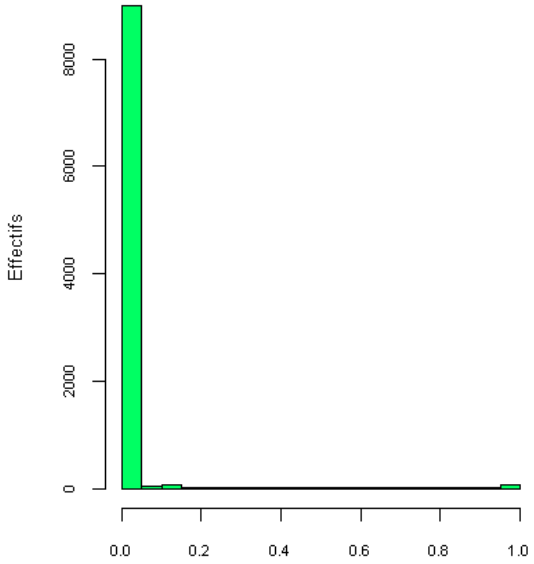
Erreur sur le sexe
(discordance entre l'EAR et le RNIPP)
n = 2 336 personnes



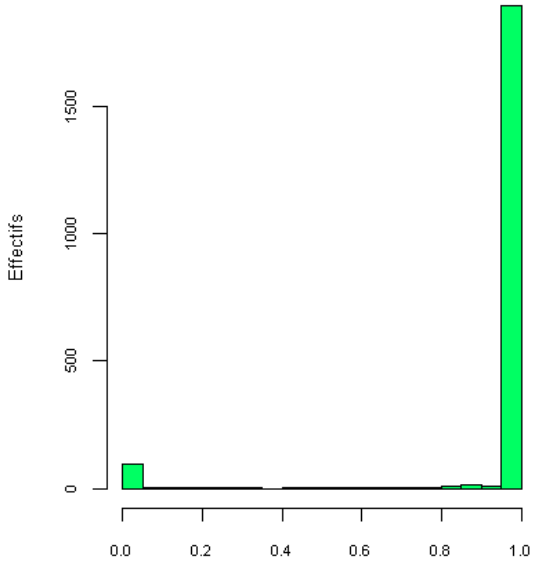
Proportion de personnes portant le même prénom et ayant un sexe différent

b) Personnes EDP apparemment en CMS (11 349 personnes)

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



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



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

Source : Base étude 2016 de l'EDP, Insee, données non pondérées.

Lecture : Parmi les 1 341 572 individus de plus de 15 ans recensés entre 2010 et 2016, qui ont déclaré vivre en couple lors de ce recensement, et pour lesquels le sexe déclaré au recensement a été confirmé dans le RNIPP, la quasi totalité porte un prénom cohérent avec le sexe déclaré au recensement : les valeurs de l'indicateur d'erreur construit avec les prénoms sont concentrées au voisinage de 0. Au contraire, pour la grande majorité des 2336 individus EDP pour lesquels le sexe déclaré au recensement a été invalidé par le RNIPP, l'indicateur construit sur la base du prénom signale la forte probabilité d'erreur sur le sexe déclaré au recensement, avec des valeurs concentrées au voisinage de 0. Les valeurs autour de 0 correspondent pour l'essentiel à des individus en échec d'appariement avec le dictionnaire des prénoms (le prénom porté est trop rare par exemple).

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

En utilisant le fait que ces erreurs de codage du sexe sont ainsi avérées dans l'EDP, il est possible de tester la capacité de la procédure utilisant les prénoms à repérer les individus concernés. La meilleure procédure sera celle qui permettra de repérer la proportion la plus importante possible des individus concernés par une erreur de codage du sexe et qui donc permettra de repérer les faux CMS de façon efficace. Cette procédure doit toutefois éviter de repérer des erreurs de codage du sexe là où il n'y en a pas.
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 EDP en couple, 2 059, soit 88 %, concernent une des 11 349 personnes en CMS apparent. Les 277 autres erreurs doivent être recherchées parmi 1,3 million de personnes, et pèsent donc peu (0,02 % des individus qui ne sont pas en CMS apparents). Proposer une correction du sexe déclaré grâce au prénom déclaré semble donc très prometteur dans le cas des CMS apparents. Dans cette sous-population, 18 % des individus ont une erreur sur le sexe. Une procédure qui permettra de réduire ce taux d'erreur améliorera la situation d'origine, ce qui paraît un objectif que l'on peut atteindre. Cela semble en revanche plus risqué si on généralise la correction à l'ensemble de la population. En effet, il faut trouver ici une procédure qui permettrait de faire mieux que l'absence de correction : cela paraît illusoire de mettre en place correction affichant au final moins de 0,02 % d'erreur de codage sur le sexe.

Tableau 3 : 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
CSO apparents	277	1 332 282	1 332 559
Total	2 336	1 341 572	1 343 908

Source : Base étude 2016 de l'EDP, Insee.
Champ : Personnes de plus de 15 ans vivant en couple.

- **La détermination de la procédure optimale**

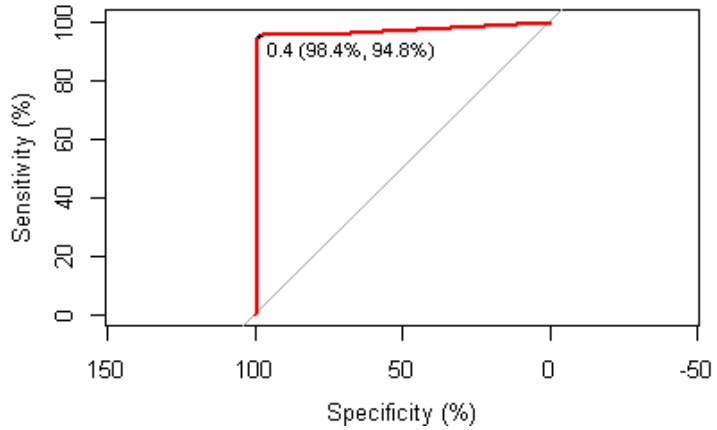
La détermination de la procédure optimale s'apparente au développement d'un test diagnostique en épidémiologie (encadré 2). C'est pourquoi les outils mobilisés sont les mêmes : mesures de spécificité et de sensibilité, détermination du seuil optimal et comparaison de tests grâce à la courbe ROC (Robin, 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 dans la base de production. Les indicateurs en sortie de cette application des dictionnaires ont ensuite été appariés à la base études 2016 de l'EDP afin de tester leur capacité à repérer les erreurs de codage.

- **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 une personne qui se déclare de sexe masculin, 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 en réalité très peu de cas où la valeur du dictionnaire est intermédiaire : la presque totalité des valeurs sont très proches de 0 ou au contraire de 100 % (graphique 4), illustrant la relative rareté des

prénoms ambigus comme Dominique. La fixation du seuil a un impact modéré dans un contexte aussi favorable. Avec le dictionnaire retenu et le seuil de 41 %, cela permet d'atteindre une sensibilité de 98 % et une spécificité de 95 % (encadré 1). 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 étude 2016 de l'EDP, EAR 2017, fichiers des prénoms de l'Etat-civil, Insee.

Tableau 4 : 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	16 950	147
Corrections valides	2088	1 951

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

3.2 La transposition à l'EAR 2017 et la nécessité d'adapter la procédure selon le mode de collecte

- **Les résultats inattendus de la transposition de la procédure de l'EDP vers l'EAR**

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. Malheureusement, la transposition n'a pu être faite de façon directe, comme cela est apparu lorsque nous avons 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'indicateur mesurant le risque d'erreur selon les différentes situations, et notamment selon que la personne est ou non en CMS apparent. En revanche, il n'est pas possible de connaître comme sur l'EDP les

erreurs avérées de codage puisque nous ne disposons pas d'une validation par une source extérieure comme le RNIPP.
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) (encadré 3). En effet, pour des raisons de coût, 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, contrairement aux bulletins destinés à être intégrés dans l'EDP.
. Sans prise en compte du mode de collecte nous étions amenés à considérer que 12 % 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. Ceci, alors même que nous cherchons une procédure pour repérer les erreurs de codage du sexe, dont la fréquence est estimée à 0,2 %, soit 60 fois moindre par rapport à 12 %. Or en fait, il s'est avéré que les erreurs portaient bien plus probablement sur l'acquisition du prénom que sur le sexe.

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 intègre 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. 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.

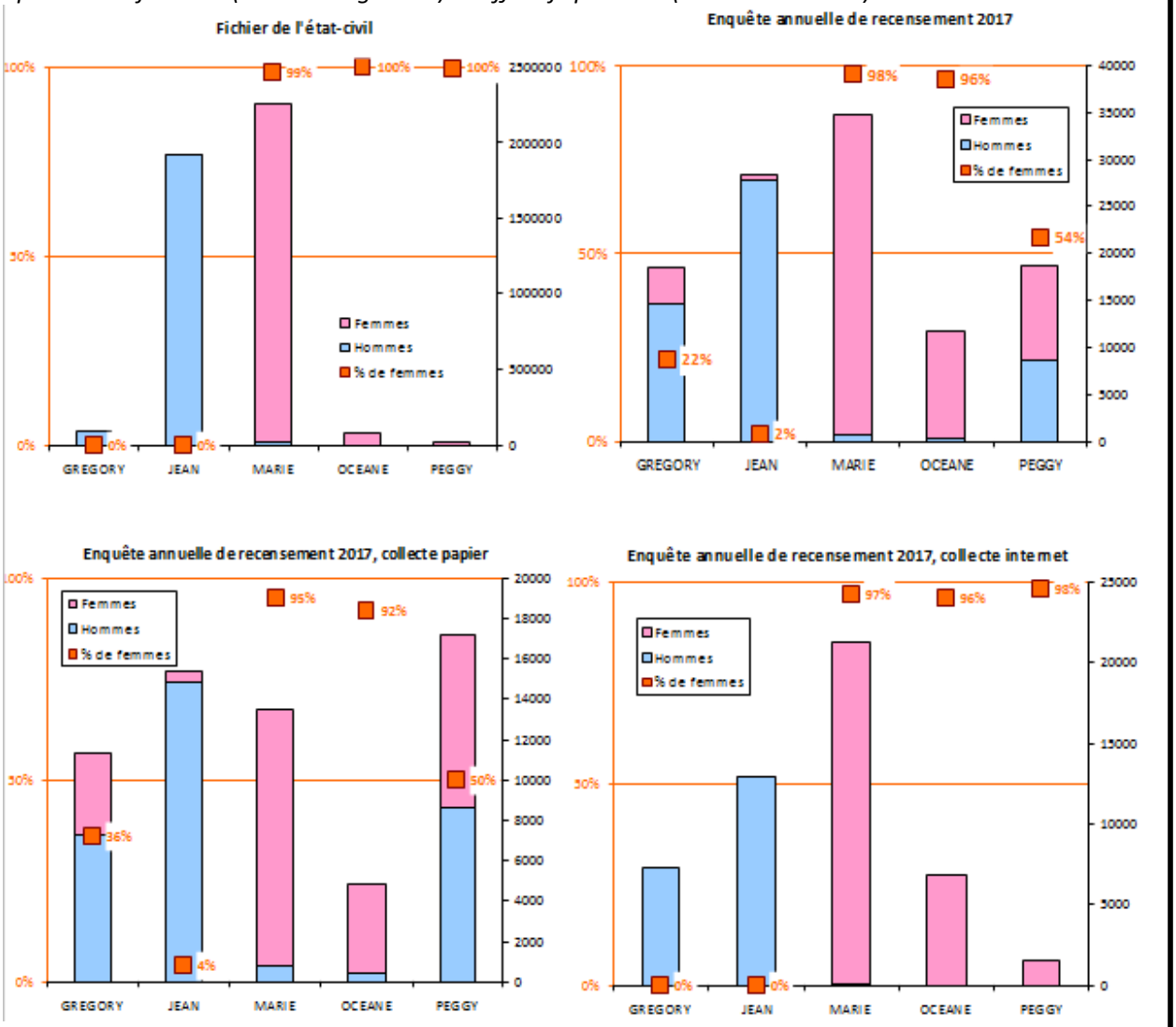
Du côté de la collecte internet, le traitement des prénoms déclarés est similaire entre individus EDP et autres individus recensés. Il peut y avoir de la non-réponse partielle sur le prénom ou des réponses inadaptées (un surnom par exemple). Les prénoms collectés sont plus variés et peuvent être affectés 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 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)



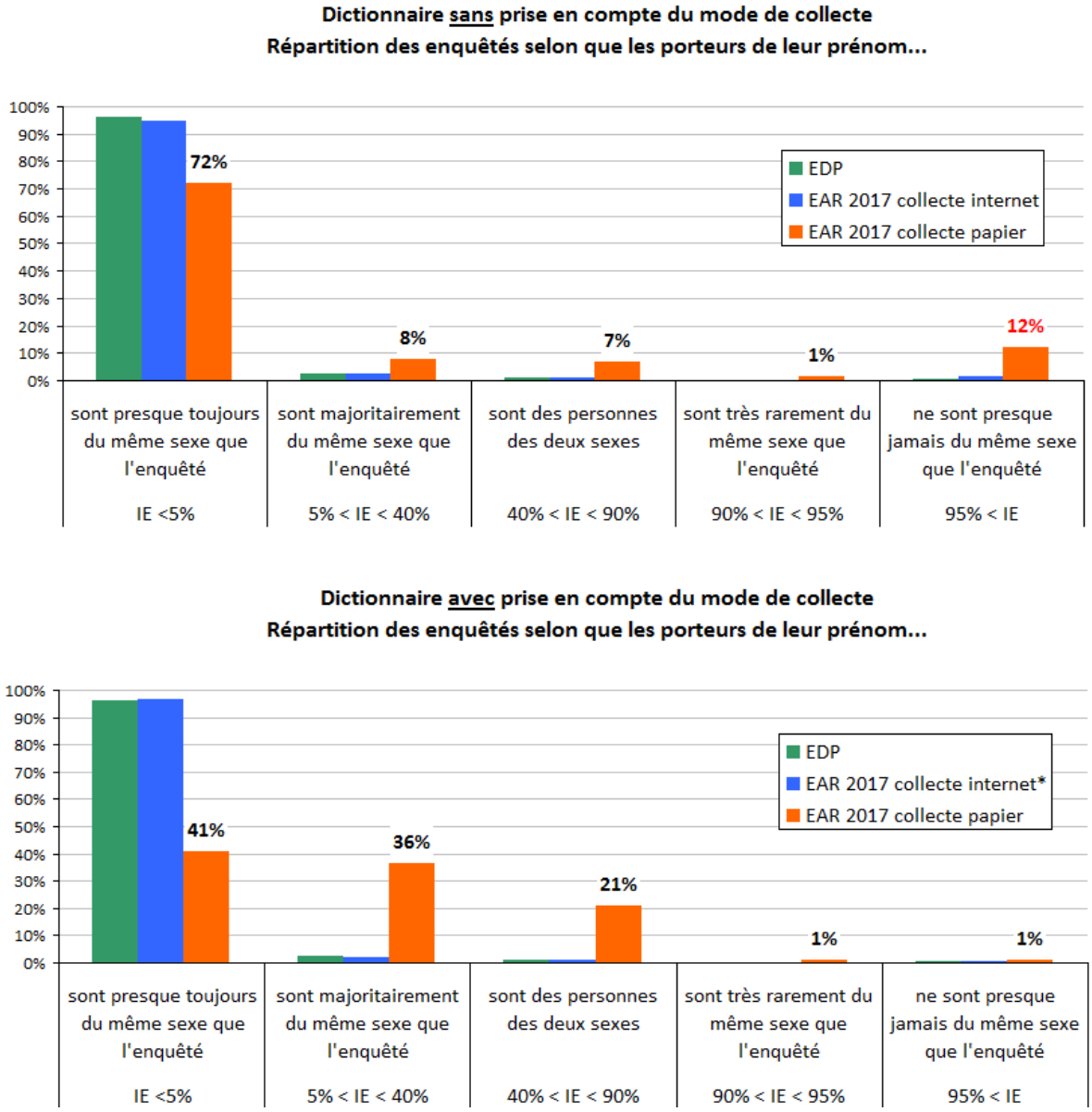
Note : On a choisi ici des exemples contrastés: deux prénoms très courants (MARIE et JEAN) et deux prénoms (PEGGY et GREGORY) pour lesquels les incohérences sont particulièrement fréquentes entre le sexe du recensement et le sexe le plus souvent associé au prénom dans l'état-civil.

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

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

Pour contourner la difficulté, la solution retenue a été d'intégrer le mode de collecte dans la construction des dictionnaires (encadré 4). 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 déclaré 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.
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.

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.
IE : indicateur d'erreur
Lecture : 72 % des personnes enquêtées sur papier portent (d'après la saisie de leur bulletin individuel de recensement) un prénom qui est presque toujours porté par des personnes du même sexe qu'elles : le dictionnaire indique une proportion inférieure à 5 % de porteurs du sexe opposé.
Champ : Personnes en couple co-résident.
Sources : Base étude 2016 de l'EDP, EAR 2017, Insee.

Encadré 4: La prise en compte du mode de collecte dans la construction des dictionnaires

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 MARIE HELENE du dictionnaire B internet est construite en ajoutant aux 37 623 personnes prénommées MARIE HELENE d'après l'état civil (toutes des femmes), les 195 MARIE HELENE nées à l'étranger et recensées par internet, également toutes des femmes. La proportion par sexe n'est en pratique pas affectée par cet ajout.
Pour le dictionnaire E internet, la combinaison se fait dans l'ordre suivant : A internet > C internet > B internet > D internet.

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 MARIE HELENE du dictionnaire B (sans critère d'année de naissance), il y a 2209 femmes et 68 hommes ainsi prénommés recensés sur papier en 2017, soit 97 % de femmes. C'est cette proportion qui est utilisée.
Pour le dictionnaire E, la combinaison se fait en cherchant un appariement par ordre de priorité dans les dictionnaires suivants : Pour le dictionnaire E papier, la combinaison se fait dans l'ordre suivant : A papier > C papier > B papier > D papier > A internet > C internet > B internet > D internet.

Tableau : Proportion de valeurs manquantes, dictionnaires construits avec mode de collecte

		Valeurs manquantes EAR 2017 INTERNET		Valeurs manquantes EAR 2017 PAPIER	
	en %	Ensemble	Personnes nées à l'étranger	Ensemble	Personnes nées à l'étranger
A	Ensemble du prénom / année de naissance	7,1	36,2	22,5	33,3
B	Ensemble	3,1	15,5	13,8	18,9
C	Première partie du prénom / année de naissance	5,6	31,1	7,2	12,5
D	Première partie du prénom	2,0	10,2	1,2	1,9
E	Combinaison	2,0	10,2	1,2	1,9

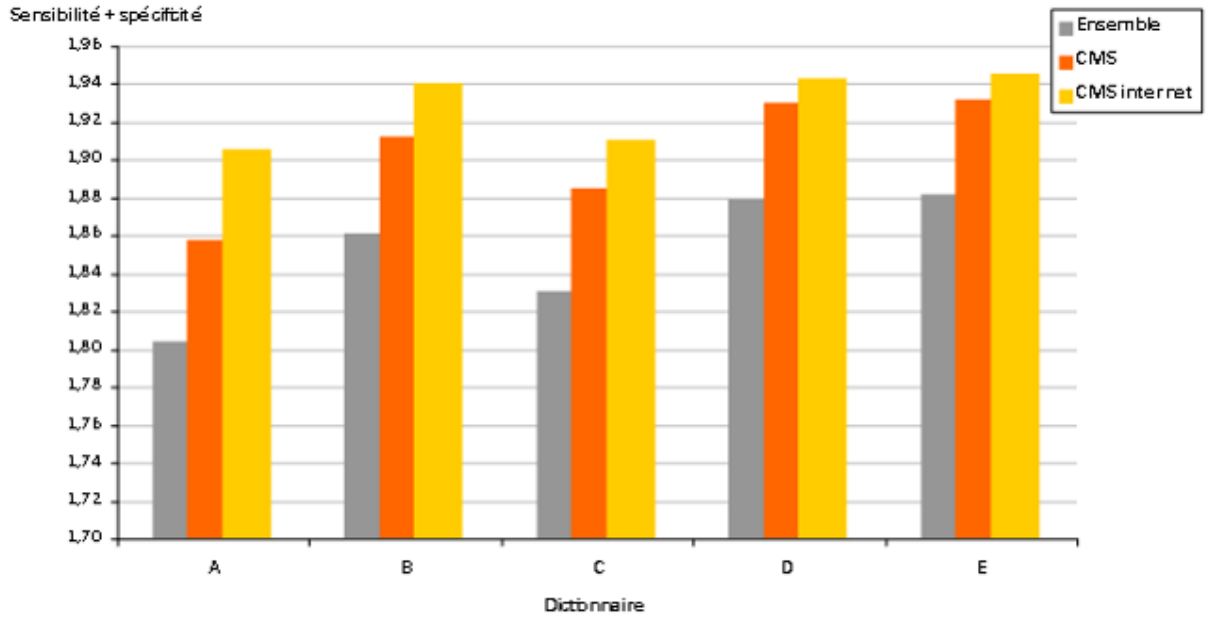
Source : EAR 2017, Insee.
Lecture : Lorsque l'on applique à la partie collectée sur internet de l'EAR 2017 le dictionnaire A construit spécifiquement pour ce mode de collecte, 7,1 % des personnes ont une valeur manquante à l'indicatrice d'erreur : aucune entrée du dictionnaire ne correspond à leur prénom (ensemble du prénom) et leur année de naissance.

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 la saisie ou l'acquisition optique (comme PEGGY). Une fois ces problèmes spécifiques réglés, l'usage des dictionnaires appuyés sur l'état civil et la collecte internet ne pose plus de difficultés.

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. Cette façon de faire est risquée compte tenu des spécificités des prénoms collectés sur papier.

Pour la collecte internet, le critère principal pour choisir le dictionnaire appliqué a été sa capacité à repérer les erreurs de codage du sexe avérées dans l'EDP. Comme dans l'encadré 2 (avant prise en compte du mode de collecte), on choisit de retenir le dictionnaire E car il semble très légèrement supérieur aux autres.

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



Source : Base étude 2016 de l'EDP, Insee

Pour la collecte papier, dans la mesure où EDP et EAR se ressemblent moins, la sensibilité et la spécificité appliquées à l'EDP sont moins importantes. On privilégie le dictionnaire qui minimise la proportion de valeurs manquantes. Là encore, le dictionnaire E semble le plus adapté.

- **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 51 % similaire à celui obtenu sur la collecte internet de l'EDP. Ce seuil est donc quelque peu différent de celui utilisé pour l'EDP dans son ensemble (41%). L'indicateur prend rarement des valeurs intermédiaires, sur l'EDP comme sur la collecte internet de l'EAR et les résultats sont en conséquence peu sensibles au seuil. 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 supérieure ou égale à 51 %. D'après l'EDP, 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 5 : Comportement du dictionnaire appliqué à l'EDP selon le mode de la collecte pour les personnes en CMS apparent

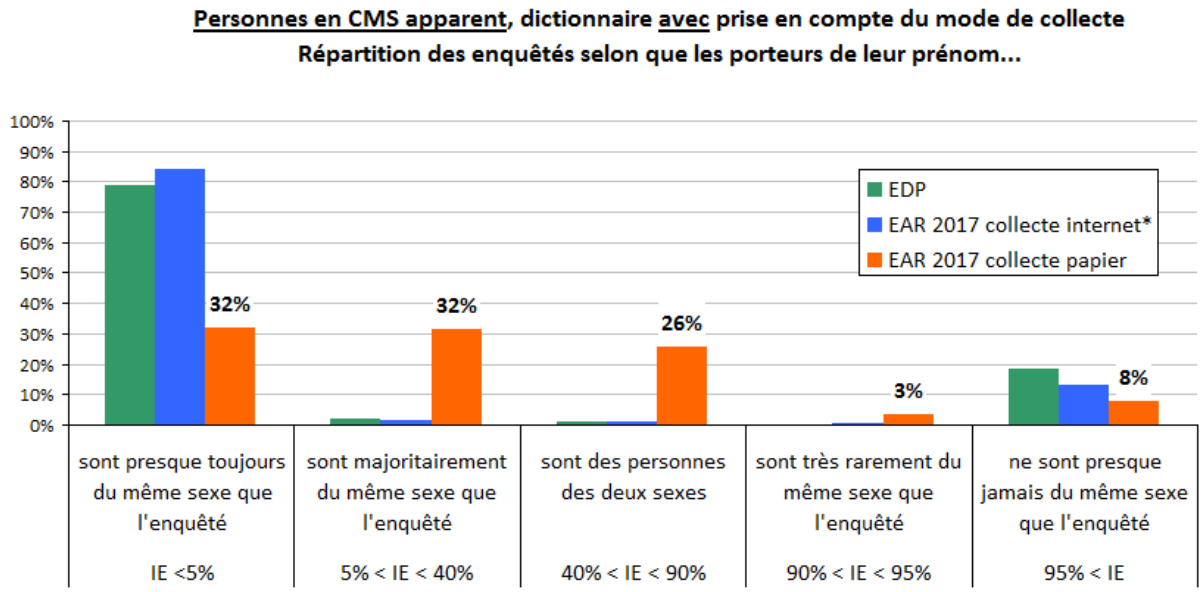
	Ensemble	***Papier***	***Internet***
Effectif	11 349	9 605	1 744
Erreurs avérées	2 059	1 852	207
Seuil (en %)	**41**	**41**	**51**
Spécificité (en%)	98	98	99
Sensibilité (en%)	95	95	95
Correction abusive	147	134	9
Correction valide	1951	1754	197

Source : Base étude 2016 de l'EDP, Insee, données non pondérées.
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 n'est pas possible 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. Alors que 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.
IE : indicateur d'erreur
Lecture : 32 % des personnes en CMS apparent enquêtées sur papier portent (d'après la saisie de leur bulletin individuel de recensement) un prénom qui est presque toujours porté par des personnes du même sexe qu'elles : le dictionnaire indique une proportion inférieure à 5 % de porteurs du sexe opposé.Champ : Personnes en CMS apparent
Sources : Base étude 2016 de l'EDP, EAR 2017, Insee.

L'incertitude est donc assez forte sur la proportion estimée de CMS, car celle-ci est très dépendante du seuil fixé. Si on corrige dans l'EAR 2017 les sexes des personnes en CMS apparent en appliquant un seuil de 50 %, on obtient 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 6: Estimation des « vrais CMS » selon le seuil

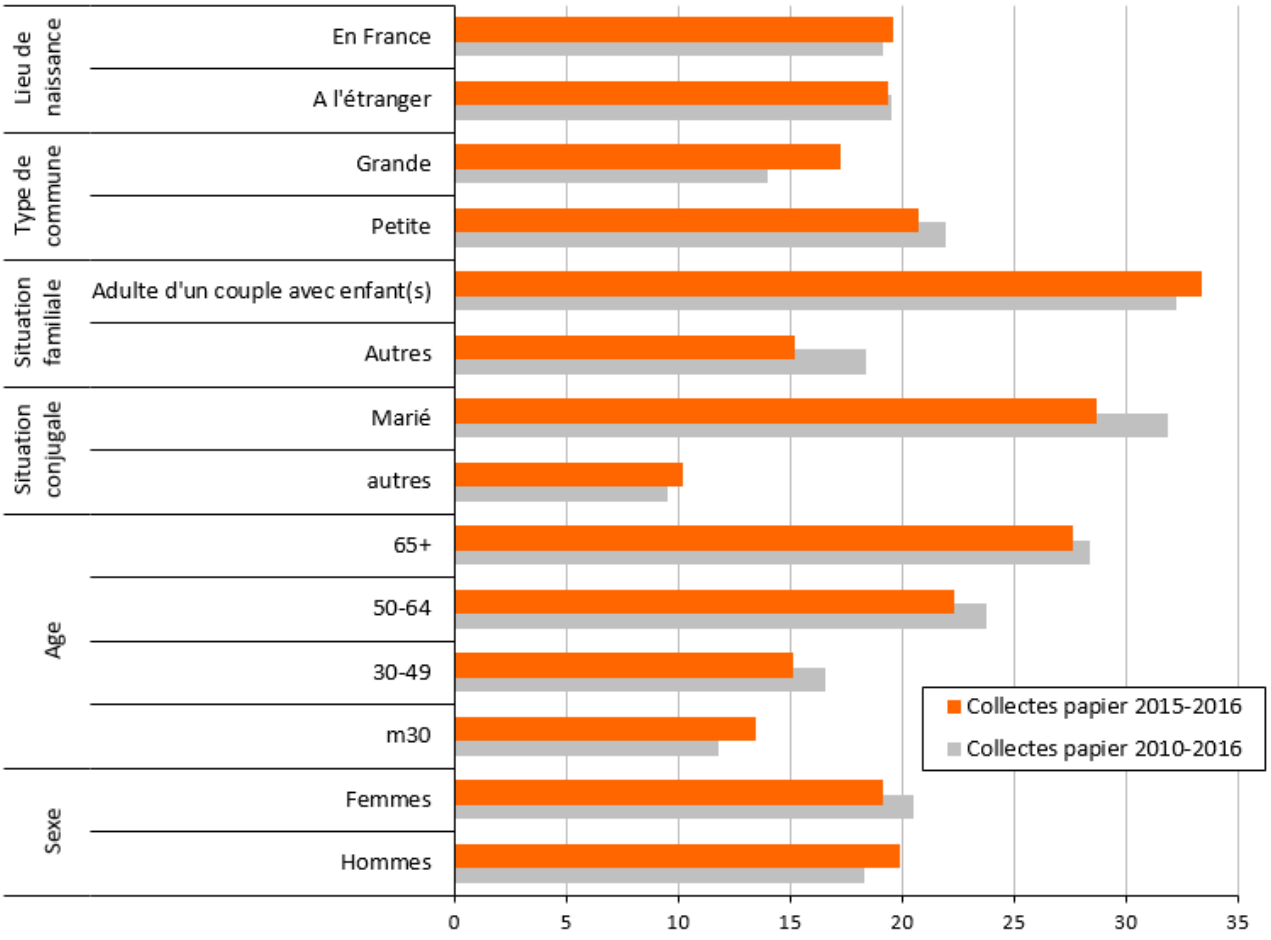
En %	*Papier*	*Internet*	*Ensemble*
Proportion d'individus en CMS apparent avant correction	1,09	1,23	1,17
Proportion estimée d'individus en « vrai » CMS, *seuil de 50%*	0,63	0,92	0,79
Proportion estimée d'individus en « vrai » CMS, *seuil de 95%*	0,94	0,94	0,94

Champ : Personnes en couple, population des ménages, hors FLNE et hors ménages où le sexe d'un des deux conjoints a été imputé
Source : EAR 2017, effectifs pondérés, 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, car il permet de mesurer la proportion d'erreurs effectives sur le sexe pour les personnes en CMS apparent selon différentes caractéristiques.
Plusieurs variables ont été testées. Les proportions d'erreurs de codage du sexe sont calculées 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 qu'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).

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



Source : Base étude EDP 2016, Insee.
Champ : Personnes en couple et apparemment en CMS.

Il s'agit ensuite de trouver les combinaisons de caractéristiques qui présentent des taux d'erreurs de codage très contrastés.

Pour l'EAR 2017, 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 %).
À partir de l'EAR 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. Cela permettra de distinguer **5 groupes**, avec davantage de contraste dans les proportions d'erreurs, puisque celles-ci s'échelonneraient de 6 à 37 %.

Tableau 7 : 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 étude 2016 de l'EDP, Insee.
Champ : Personnes en couple, en CMS apparent, recensées sur papier en 2015 ou 2016.

L'identification précise du conjoint dans l'analyse ménage-familles à partir de 2018 permettra aussi de prendre en compte l'indicateur de la proportion d'erreurs pour ce conjoint. On pourra alors tenir compte du fait que la probabilité d'une double erreur est très faible : si un des conjoints a été corrigé (parce que son prénom était associé sans ambiguïté à un sexe opposé à celui codé), il ne faut pas corriger le second. A l'inverse, on s'attend à trouver des taux doublés au sein des couples. Par exemple, il faudrait corriger le sexe d'un des conjoints au sein de 74 % des couples mariés avec enfants (37 % x 2). Au sein des couples non mariés sans enfants, ce sont seulement 12 % des couples (6 % x 2) qui seraient concernés par une correction (pour un des conjoints).

3.3 Synthèse de la méthode de correction appliquée 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 nous ont donc conduits 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 recensement 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 exactement 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 l'individu, 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 %[10] 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 (pour 2018 : 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 de l'indicateur, entre 20 % et 90 %. 1 % ont une valeur manquante. Au total, ce sont 40 % des personnes en CMS apparent de la collecte papier et 17 % du total (collecte internet comprise), qui se voient qualifiés de vrais ou faux CMS de façon aléatoire en fonction de leur groupe.

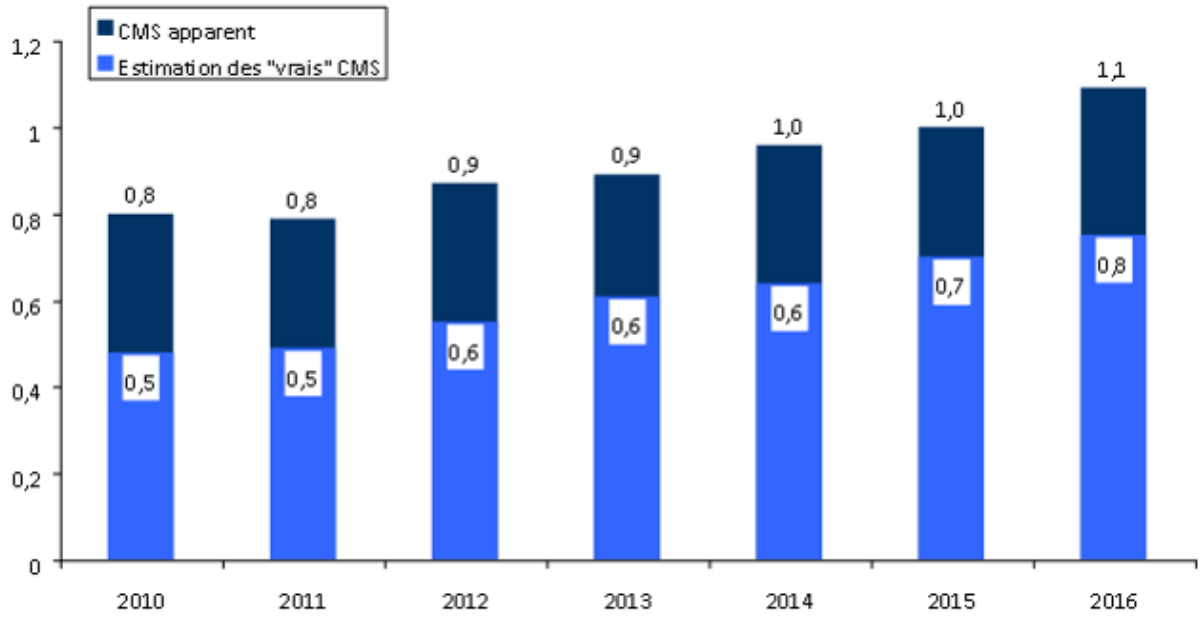
10 L'intervalle est asymétrique parce qu'en cas d'incohérence entre sexe et prénom, nous considérons que l'association d'un prénom à un sexe est davantage bruitée par une mauvaise qualité de la saisie du prénom par rapport à une mauvaise qualité du codage du sexe. On accorde donc toujours du crédit au 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 41 %, aux individus EDP, et non à leur conjoint. Toutefois, en supposant 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 Famille 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 : Base étude 2016 de l'EDP, données pondérées, Insee.

Tableau 8 : 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

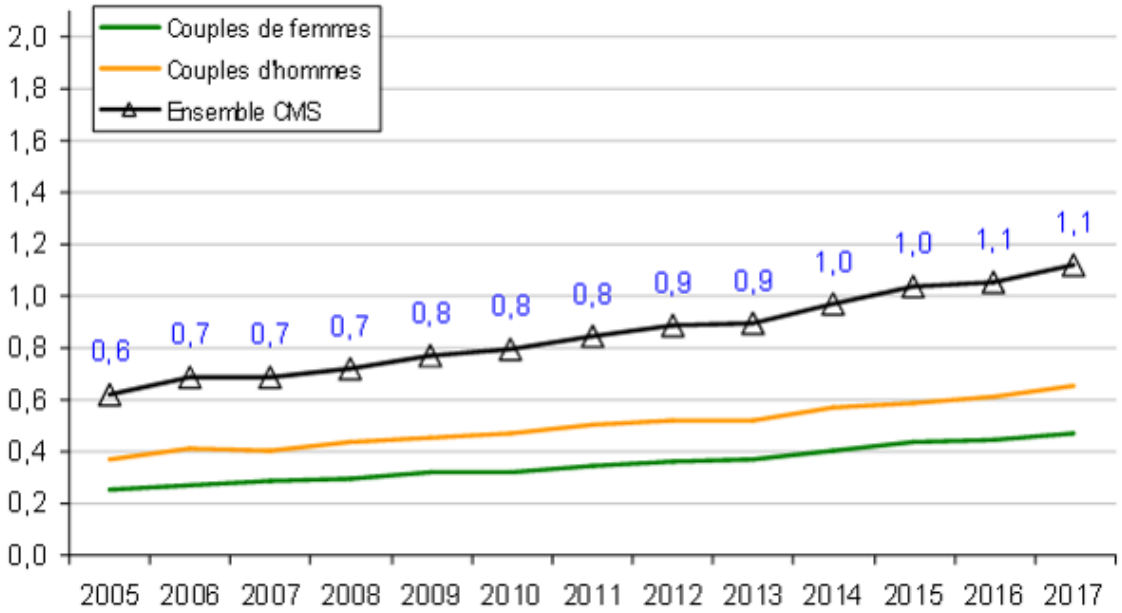
Champ : Personnes de plus de 15 ans ayant déclaré vivre en couple dans leur bulletin individuel.
Source : Base étude 2016 de l'EDP , données pondérées, Insee.

4.2 Dans les enquêtes annuelles de recensements

Dans les enquêtes annuelles de recensement, la hausse des CMS *apparents* s'observe dans des proportions sensiblement identiques à l'EDP, sur une période plus longue, entre 2005 et 2017. Toujours sur les couples apparemment de même sexe, 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 en couple, hors FLNE et hors ménages où le sexe d'un des deux conjoints a été imputé.

Tableau 9 : 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.
Champ : Personnes en couple, hors FLNE et hors ménages où le sexe d'un des deux conjoints a été imputé.

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.

Annexe 1 : La construction des dictionnaires

Cette annexe décrit avec plus de détails et des exemples les différents dictionnaires testés. Elle complète les encadrés 1 et 4. La construction des dictionnaires s'est faite en deux étapes, qu'il est plus facile de distinguer dans la description pour plus de clarté. À l'issue de la première étape, il s'est en effet avéré nécessaire de tenir compte du mode de collecte (encadré 3), aussi bien dans la construction des dictionnaires que dans leur utilisation afin d'affecter un indicateur à chaque personne recensée.

a) : Une première étape sans mode de collecte

Dans les différentes sources, les prénoms ont été traités afin d'enlever les accents, les tirets, les points, les apostrophes. Les prénoms de longueur inférieure à 3 caractères ont été mis à blanc car ils sont souvent tronqués et ambigus (abréviation par exemple, surnom). Quand ce sont de vrais prénoms, comme « JO », ils peuvent souvent être portés par des femmes et des hommes.

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 des dictionnaires alternatifs ou complémentaires ont été construits à partir de l'enquête annuelle de recensement 2017, qui couvre l'ensemble des personnes résidant en France en 2017.

Lors de cette première étape 21 dictionnaires ont été constitués. La fréquence minimale retenue pour qu'un prénom figure dans le dictionnaire est de 10 occurrences. Cette limite s'ajoute aux conditions[11] qui sont appliquées pour constituer le fichier des prénoms de l'état-civil et s'applique aussi aux dictionnaires construits avec les prénoms de l'enquête annuelle de recensement. Elle assure que l'information disponible sur une entrée du dictionnaire soit suffisamment robuste. La façon de les construire et leurs différences sont explicitées ici, à partir de trois exemples :

– Marie-Hélène, née en Belgique en 1980, recensée en 2017 ;

– « Epouse », née en Islande en 1980, recensée en 2017 ;

– Peggy, née en France en 1980, recensée en 2017 sur papier.

Les dictionnaires 1 à 12 ont été construits avec les prénoms figurant sur les bulletins individuels de l'EAR 2017. L'idée initiale était en effet que ces prénoms permettent d'améliorer le traitement des personnes nées à l'étranger puisque les premiers résultats montraient que le fichier des prénoms de l'État-civil est insuffisant. Cela s'explique puisque le fichier de l'État-civil concerne les naissances enregistrées en France tandis que les prénoms trouvés dans l'enquête annuelle de recensement sont ceux portés par les personnes résidant en France, quel que soit leur lieu de naissance.

11 3 conditions sont mentionnées (https://www.insee.fr/fr/statistiques/2540004#documentation) :
1. Sur la période allant de 1900 à 1945, le prénom a été attribué au moins 20 fois à des personnes de sexe féminin et/ou au moins 20 fois à des personnes de sexe masculin
2. Sur la période allant de 1946 à 2016, le prénom a été attribué au moins 20 fois à des personnes de sexe féminin et/ou au moins 20 fois à des personnes de sexe masculin
3. Pour une année de naissance donnée, le prénom a été attribué au moins 3 fois à des personnes de sexe féminin ou de sexe masculin
Les effectifs des prénoms ne remplissant pas les conditions 1 et 2 sont regroupés (pour chaque sexe et chaque année de naissance) dans un enregistrement dont le champ prénom (PREUSUEL) prend la valeur «_PRENOMS_RARES». Les effectifs des prénoms remplissant la condition 2 mais pas la condition 3 sont regroupés (pour chaque sexe et chaque prénom) dans un enregistrement dont le champ année de naissance (ANNAIS) prend la valeur «XXXX».

En première intention, il paraissait souhaitable d'intégrer le maximum d'information : un même prénom peut être attribué plutôt à des femmes ou à des hommes selon la période (comme le prénom Camille), le lieu de naissance (Andréa), ou encore selon qu'il est porté seul ou dans un prénom composé (Jean-Marie). Les dictionnaires testés explorent donc les différents critères qu'il est possible de prendre en compte : l'année de naissance, le prénom en entier ou sa première partie, le lieu de naissance.

Dictionnaires 1 à 12, construits à partir des prénoms de l'EAR 2017

Les dictionnaires 1 à 4 ont en commun d'avoir des entrées distinctes selon le pays de naissance, ceux de 5 à 8 de distinguer les entrées entre naissance en France et naissance à l'étranger, tandis que les dictionnaires 9 à 12 ne distinguent pas les entrées selon le pays de naissance. Les dictionnaires 13 à 16 sont construits à partir de l'état-civil et les dictionnaires A à E sont des combinaisons des dictionnaires 1 à 16.

Le dictionnaire 1 prend en compte, en plus du pays de naissance, le prénom et l'année de naissance. Dans l'exemple de Marie-Hélène, l'indicateur issu du dictionnaire 1 est construit en calculant la proportion de femmes parmi les « MARIE HELENE », nées en Belgique en 1980 et recensées en 2017. Mais elles sont en nombre insuffisant (inférieur à 10) pour construire l'indicateur. Pour le dictionnaire 2, la contrainte sur l'année de naissance est levée : la proportion est calculée sur les « MARIE HELENE », nées en Belgique et recensées en 2017. Elles sont 12, toutes des femmes et l'indicateur vaut 100 %.

Le dictionnaire 3 prend en compte année de naissance et lieu de naissance, mais seulement la première partie du prénom (avant un tiret ou un espace). Dans l'exemple de Marie-Hélène, l'indicateur est la proportion de femmes parmi toutes les personnes dont la première partie du prénom est « MARIE », nées en Belgique en 1980. Le dictionnaire 4 est identique au 3, si ce n'est que l'année de naissance n'est pas prise en compte.

Le dictionnaire 5 (resp. 6, 7 et 8) est identique au dictionnaire 1 (resp. 2, 3 et 4) à la différence près qu'il considère, au lieu du pays de naissance détaillé, le lieu de naissance France / Étranger. Ainsi pour le dictionnaire 5, la proportion de femmes est calculée sur les « MARIE HELENE » nées à l'étranger en 1980 et recensées en 2017. Elles sont moins de 10 donc l'indicateur est manquant.

Marie-Hélène, née en Belgique en 1980, recensée en 2017		Femmes	Hommes	% de femmes
1 MARIE HELENE, nées en Belgique en 1980	recensées en 2017	-	-	-
2 MARIE HELENE, nées en Belgique		12	0	100,0
3 MARIE*, nées en Belgique en 1980		-	-	-
4 MARIE*, nées en Belgique		546	16	97,2
5 MARIE HELENE, nées à l'étranger en 1980		-	-	-
6 MARIE HELENE, nées à l'étranger		351	17	95,4
7 MARIE*, nées à l'étranger en 1980		99	14	87,6
8 MARIE*, nées à l'étranger		12 629	830	93,8
9 MARIE HELENE, nées en 1980		36	0	100,0
10 MARIE HELENE		5879	74	98,8
11 MARIE*, nées en 1980		5915	74	98,8
12 MARIE*		190 069	4506	97,7
13 MARIE HELENE, nées en 1980	inscrites à l'état-civil	228	0	100,0
14 MARIE HELENE		37 623	0	100,0
15 MARIE*, nées en 1980		11 997	4	100,0
16 MARIE*		3 116 784	28 056	99,1
A MARIE HELENE, nées en 1980	combinaison[1]	228	0	100,0
B MARIE HELENE		37 974	17	100,0
C MARIE*, nées en 1980		12 096	18	99,9
D MARIE*		3 129 413	28 886	99,1
E Le premier renseigné parmi A > C > B > D, ici A		228	0	100,0

1 : inscrites à l'état-civil (et nécessairement nées en France) ou recensées en 2017 et nées à l'étranger.

« Epouse », née en Islande en 1980, recensée en 2017		Femmes	Hommes	% de femmes
1 EPOUSE, nées en Islande en 1980	recensées en 2017	-	-	-
2 EPOUSE, nées en Islande		-	-	-
3 EPOUSE*, nées en Islande en 1980		-	-	-
4 EPOUSE*, nées à l'étranger		-	-	-
5 EPOUSE, nées à l'étranger en 1980		-	-	-
6 EPOUSE, nées à l'étranger		-	-	-
7 EPOUSE*, nées à l'étranger en 1980		-	-	-
8 EPOUSE*, nées à l'étranger		29	1	96,7
9 EPOUSE, nées en 1980		-	-	-
10 EPOUSE		25	0	100,0
11 EPOUSE*, nées en 1980		-	-	-
12 EPOUSE*		188	1	99,5
13 EPOUSE, nées en 1980	inscrites à l'état-civil	-	-	-
14 EPOUSE		-	-	-
15 EPOUSE*, nées en 1980		-	-	-
16 EPOUSE*		-	-	-
A EPOUSE, nées en 1980	combinaison[1]	-	-	-
B EPOUSE		25	0	100,0
C EPOUSE*, nées en 1980		-	-	-
D EPOUSE*		188	1	99,5
E Le premier renseigné parmi A > C > B > D, ici A		25	0	100,0

1 : inscrites à l'état-civil (et nécessairement nées en France) ou recensées en 2017 et nées à l'étranger.

Peggy, née en France en 1980, recensée en 2017		Femmes	Hommes	% de femmes
1 PEGGY, nées en France en 1980	recensées en 2017	175	74	70,3
2 PEGGY, nées en France		9 168	7 562	54,8
3 PEGGY*, nées en France en 1980		194	90	68,3
4 PEGGY*, nées en France		11 092	9 257	54,5
5 PEGGY, nées en France en 1980		175	74	70,3
6 PEGGY, nées en France		9168	7 562	54,8
7 PEGGY*, nées en France en 1980		194	90	68,3
8 PEGGY*, nées en France		11 092	9257	54,5
9 PEGGY, nées en 1980		201	91	68,8
10 PEGGY		10 218	8582	54,4
11 PEGGY*, nées en 1980		225	112	66,8
12 PEGGY*		12 468	10 557	54,1
13 PEGGY, nées en 1980	inscrites à l'état-civil	879	0	100,0
14 PEGGY		17 532	74	99,6
15 PEGGY*, nées en 1980		879	0	100,0
16 PEGGY*		17 532	74	99,6
A PEGGY, nées en 1980	combinaison[1]	901	16	98,3
B PEGGY		18 452	968	95,0
C PEGGY*, nées en 1980		906	21	97,7
D PEGGY*		18 746	1 228	93,9
E Le premier renseigné parmi A > C > B > D, ici A		901	16	98,3

1 : inscrites à l'état-civil (et nécessairement nées en France) ou recensées en 2017 et nées à l'étranger.

Le dictionnaire 9 (resp. 10, 11 et 12) est identique au dictionnaire 1 (resp. 2, 3 et 4) à la différence près qu'il ne prend pas en compte le lieu de naissance. Ainsi pour le dictionnaire 9, la proportion de femmes est calculée sur les Marie-Hélène nées en 1980 et recensées en 2017. Elles sont 36, là encore toutes des femmes.

Dictionnaires 13 à 16, construits à partir des prénoms de l'état-civil

Le dictionnaire 13 (resp. 14, 15 et 16) prend en compte les mêmes critères que le dictionnaire 1 (resp. 2, 3 et 4) mais il s'appuie sur l'état-civil, et non sur l'EAR 2017, donc uniquement sur des personnes nées en France. Ainsi pour le dictionnaire 13, la proportion de femmes est calculée sur les Marie-Hélène nées en 1980 en France et enregistrées à l'état-civil. Elles sont 228.

Comme les données sont exhaustives et reposent sur les naissances depuis 1900, les effectifs sur lesquels s'appuie le calcul de l'indicateur sont plus conséquents.

En revanche, dans certains cas, les données de l'enquête annuelle de recensement apportent davantage d'information. Ainsi, il peut arriver qu'à la place du prénom figure « Epouse » sur le bulletin individuel. Cette information est évidemment une bonne indication du sexe, mais ce n'est pas un prénom enregistré à l'état-civil.

Dictionnaires combinés A à E

Les dictionnaires combinés permettent de prendre en compte les informations de l'état-civil mais aussi de l'EAR. Ils ont été construits de façon à privilégier l'état-civil, qui est exhaustif pour les naissances enregistrées en France. Si la vague 2017 de l'EAR comportait des entrées pour des personnes nées à l'étranger, elles ont été ajoutées aux entrées de l'état-civil.

Par exemple, alors que dans le dictionnaire 15 figuraient 11 997 femmes nées en 1980 avec un prénom débutant par Marie, dans le dictionnaire C y sont ajoutées les 99 femmes nées à l'étranger (voir dictionnaire 5), ce qui porte l'effectif de femmes à 12 096. Dans ce cas, la proportion n'est pas affectée par cet ajout. L'autre modification est que les entrées absentes de l'état-civil sont reprises de l'EAR : ainsi les entrées « Epouse » sont ajoutées. Cela ajoute des prénoms étrangers (Maria de Fatima, Malgorzata, El Hassan, Iwona), quelques fautes d'orthographe (Dider, Dominque), des abréviations pour les prénoms composés (Jluc) ou l'absence d'un tiret entre les deux prénoms (Jeanclaude). Du côté du prénom simplifié, on ajoute aussi J (Un prénom de moins de 3 caractères n'est pas pris en compte, mais « J LOUIS » par exemple contient plus de 3 caractères et la première partie est « J »), souvent en première partie comme abréviation de Jean.

Enfin, le dictionnaire E est construit a posteriori, par étapes : s'il n'y a pas d'appariement dans le dictionnaire A, on prend le C (première partie du prénom) puis le B (prénom mais sans année de naissance) et enfin le D (première partie du prénom, sans année de naissance).

Le comportement des dictionnaires testés à l'issue de cette première étape

À l'issue de cette étape, 21 dictionnaires sont disponibles. Ils comprennent entre 14 000 et 237 000 entrées.

Une entrée n'est maintenue dans un dictionnaire qui si elle repose sur au moins 10 personnes.
Par exemple, moins de 10 personnes prénommées « MARIE HELENE », nées en 1980 à l'étranger sont trouvées dans l'état-civil. Cette entrée n'apparaît pas dans le dictionnaire 5 et une Marie-Hélène née en 1980 sera en échec d'appariement pour le dictionnaire 5. En revanche, dans le dictionnaire 6 qui fait abstraction de l'année de naissance, cette même personne pourra être appariée. Plus les critères sont précis, plus le taux de valeurs manquantes est donc élevé. Cela nous a conduit à renoncer à certains critères comme le pays de naissance détaillé et même la distinction entre personnes nées en France et à l'étranger.

N°	Lieu de naissance	Prénom	Année	Source	Nombre d'entrées	% valeurs manquantes	
						Ensemble	Pers. nées à l'étranger
1	Pays	Ensemble	Oui	recensées en 2017	83 398	23,7	90,2
2	Pays	Ensemble	-		26 821	11,7	47,5
3	Pays	Première partie	Oui		92 412	16,1	86,4
4	Pays	Première partie	-		24 717	4,9	35,8
5	France / Pays étranger	Ensemble	Oui		92 590	21,5	67,3
6	France / Pays étranger	Ensemble	-		20 210	9,5	23,2
7	France / Pays étranger	Première partie	Oui		104 068	13,3	56,1
8	France / Pays étranger	Première partie	-		15 949	2,5	9,8
9	-	Ensemble	Oui		92 574	18,0	40,5
10	-	Ensemble	-		18 988	8,7	18,5
11	-	Première partie	Oui		99 785	9,4	25,1
12	-	Première partie	-		14 351	1,9	6,1
13	-	Ensemble	Oui	inscrites à l'état-civil	236 632	17,3	40,2
14	-	Ensemble	-		28 996	9,4	19,3
15	-	Première partie	Oui		223 154	11,4	30,0
16	-	Première partie	-		26 488	2,0	6,5
A	-	Ensemble	Oui	combinaison[1]	247 773	14,6	34,3
B	-	Ensemble	-		34 549	8,1	17,2
C	-	Première partie	Oui		239 862	6,7	20,1
D	-	Première partie	-		28 580	1,3	4,9
E	-	Combinaison**			//	1,3	4,9

Lieu de naissance : « Pays » signifie par pays détaillé, «France/Pays étranger » signifie la distinction par une indicatrice des personnes nées en France d'avec celles nées à l'étranger. « - » signifie que le lieu de naissance n'est pas pris en compte.

Lecture : Avec le dictionnaire 1, le plus détaillé (par pays de naissance, année de naissance et prénom), 23,7 % des personnes recensées en 2017 ne sont pas trouvées, et même 90,2 % de celles nées à l'étranger (car il faut trouver au moins 9 autres personnes recensées en 2017, nées la même année dans le même pays.

Champ : ensemble des personnes recensées en 2017.

Source : Enquêtes annuelles de recensement et base études 2016 de l'échantillon démographique permanent.

b) Une seconde étape intégrant le mode de collecte

Cette seconde étape a été dictée par les résultats qui montraient des différences substantielles selon le mode de collecte dans la qualité des prénoms et de ce fait dans la capacité des dictionnaires à repérer les erreurs (encadré 3).

Pour simplifier, les dictionnaires 1 à 8 de la première étape, prenant en compte le pays de naissance, sont éliminés car trop détaillés. Pour chacun des dictionnaires 9 à 12, dont la source est l'EAR 2017, deux dictionnaires distincts sont constitués, selon que la collecte a été réalisée sur internet ou papier. Les dictionnaires 13 à 16, dont la source est l'état-civil, ne sont pas modifiés.

Pour la partie internet, les dictionnaires A à E ajoutent aux naissances de l'état-civil les personnes recensées sur internet et nées à l'étranger. Comme auparavant, lorsqu'une entrée était complètement absente à l'état-civil, elle est ajoutée[12]. Dans l'exemple de Peggy, recensée sur internet, cela ne change presque rien par rapport à l'étape 1 : avec le dictionnaire E, elle se voit affecter une proportion de 100 % car 879 Peggy nées en 1980 figurent à l'état-civil, et s'y ajoute une Peggy née à l'étranger en 1980 et recensée sur internet en 2017. Ce sont toutes des femmes. Lors de la première étape, on ajoutait aussi 36 personnes prénommées Peggy nées à l'étranger en 1980 et recensées sur papier, dont 16 hommes. Cela conduisait à une proportion globale de femmes de 98,3 %.

Pour la partie papier, priorité est donnée aux dictionnaires issus de la collecte papier, puis en cas de défaut d'appariement, on prend les dictionnaires combinés internet. Donc le dictionnaire A papier est obtenu est prenant le dictionnaire 9 et à défaut le dictionnaire A internet.

Pour le dictionnaire E papier, la combinaison se fait dans l'ordre suivant :

9 papier > 11 papier > 10 papier > 12 papier > A internet > C internet > B internet > D internet.

Les résultats sont sensiblement différents, comme on le voit sur l'exemple. Alors que sans prise en compte du mode de collecte, le dictionnaire E associait à la personne prénommée Peggy une proportion de femmes de 98,3 %, elle est désormais de 57,3 %. L'écart est donc considérable et permet de rétablir un indicateur qui correspond mieux à la réalité, à savoir à la forte incertitude qui pèse sur le sexe des porteurs du prénom Peggy en cas de collecte papier. Une personne aux caractéristiques identiques recensée sur internet se verra en revanche affecter un indicateur de 100 %.

12 C'est le cas pour « Epouse », dont le traitement est inchangé par rapport à l'étape précédente : ce « prénom » n'apparaît que sur internet lorsqu'il est directement saisi par les personnes recensées. Sur papier, le prestataire confronte les prénoms saisis à une liste préétablie dans laquelle « Epouse » n'apparaît pas.

« Peggy », née en France en 1980, recensée sur papier en 2017		Femmes	Hommes	% de femmes
9 PEGGY, nées en 1980	recensées **sur papier** en 2017	122	91	57,3
10 PEGGY		8673	8575	50,3
11 PEGGY*, nées en 1980		145	112	56,4
12 PEGGY*		10913	10548	50,9
13 PEGGY, nées en 1980	inscrites à l'état-civil	879	0	100,0
14 PEGGY		17532	74	99,6
15 PEGGY*, nées en 1980		879	0	100,0
16 PEGGY*		17532	74	99,6
A PEGGY, nées en 1980	combinaison avec priorité aux dictionnaires papier[1]	122	91	57,3
B PEGGY		8673	8575	50,3
C PEGGY*, nées en 1980		145	112	56,4
D PEGGY*		10913	10548	50,9
E Le premier renseigné parmi 9 > 11 > 10 > 12 > A > C > B > D, ici 9		122	91	57,3

1 : En priorité, si l'effectif est suffisant, personnes recensées **sur papier** en 2017. Sinon, personnes **i**nscrites à l'état-civil (et nécessairement nées en France) ou recensées **sur internet** en 2017 et nées à l'étranger.

« Peggy », née en France en 1980, recensée sur internet en 2017		Femmes	Hommes	% de femmes
9 PEGGY, nées en 1980	recensées **sur internet** en 2017	79	0	100,0
10 PEGGY		1545	7	99,5
11 PEGGY*, nées en 1980		80	0	100,0
12 PEGGY*		1555	9	99,4
13 PEGGY, nées en 1980	inscrites à l'état-civil	879	0	100,0
14 PEGGY		17532	74	99,6
15 PEGGY*, nées en 1980		879	0	100,0
16 PEGGY*		17532	74	99,6
A PEGGY, nées en 1980	combinaison[1]	880	0	100,0
B PEGGY		17557	74	99,6
C PEGGY*, nées en 1980		880	0	100,0
D PEGGY*		17559	74	99,6
E Le premier renseigné parmi A > C > B > D, ici A		880	0	100,0

1 : **I**nscrites à l'état-civil (et nécessairement nées en France) ou recensées **sur internet** en 2017 et nées à l'étranger.

c) Le comportement des différents dictionnaires

C'est le dictionnaire E qui a été retenu, après comparaison des dictionnaires (voir encadré 1). Les dictionnaires ont été comparés sur leurs performances pour repérer les erreurs de codage sur le sexe dans l'EDP, mais aussi sur la proportion de valeurs manquantes quand on les applique à l'EAR. Ce dernier critère est le plus important pour la collecte papier.

N°	Prénom	Année	Source	PAPIER				INTERNET			
				Nombre d'entrées	Nombre moyen occurrences / entrée	% valeurs manquantes		Nombre d'entrées	Nombre moyen occurrences / entrée	% valeurs manquantes	
						Ens.	Pers. nées à l'étranger			Ens.	Pers. nées à l'étranger
9	Ensemble	Oui	recensées en 2017	75 204	43	27	42	45 017	88	15	55
10	Ensemble	Non		11 362	330	15	21	10 135	439	5	20
11	1ere partie	Oui		86 721	45	11	18	43 285	93	13	51
12	1ere partie	Non		7 116	606	2	4	9 448	477	3	14
13	Ensemble	Oui	inscrites à l'état-civil	241 919	331	26	41	241 919	331	7	38
14	Ensemble	Non		29 581	2777	16	21	29 581	2777	3	16
15	1ere partie	Oui		228 383	351	16	27	228 383	351	6	33
16	1ere partie	Non		26 988	3043	2	3	26 988	3043	2	11
A	Ensemble	Oui	combi-naison*	**	**	22	33	244 655	328	7	36
B	Ensemble	Non		**	**	14	19	29 980	2750	3	16
C	1ere partie	Oui		**	**	7	12	231 287	347	6	31
D	1ere partie	Non		**	**	1	2	27 321	3018	2	10
E	Combinaison			***	***	1	2	****	****	2	10

* : Personnes inscrites à l'état-civil (nées en France) ou recensées **sur internet** en 2017 et nées à l'étranger.

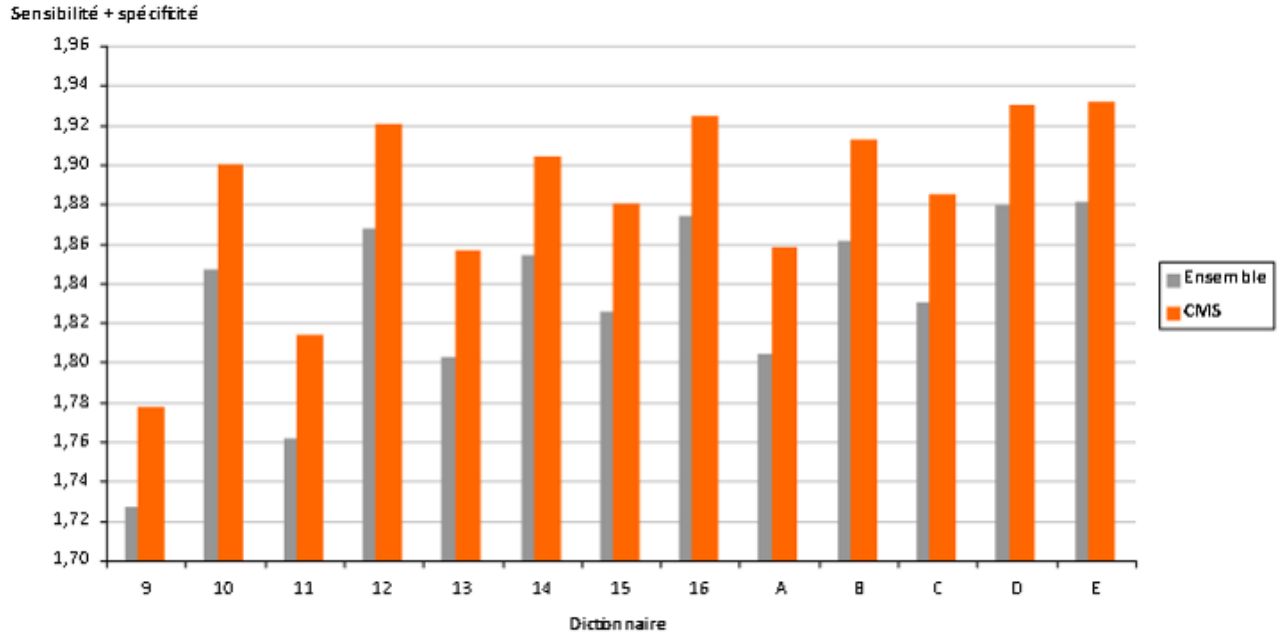
** En priorité, si l'effectif est suffisant, personnes recensées **sur papier** en 2017. Sinon, personnes inscrites à l'état-civil ou recensées **sur internet** en 2017 et nées à l'étranger.

*** Pour les personnes recensées par internet, par ordre de priorité les dictionnaires papiers A, C, B et D, puis les dictionnaires A, C, B et D internet

**** Pour les personnes recensées par internet, par ordre de priorité les dictionnaires A, C, B et D.

Source : Bases études 2016 de l'échantillon démographique permanent et enquête annuelle de recensement 2017.

Graphique : Qualité (Somme de la sensibilité et de la spécificité) des différents dictionnaires selon le type de couple



Source : Bases études 2016 de l'échantillon démographique permanent, Insee.

Série des documents de travail de la DSDS

F1806 : « L'effet d'une variation du montant de certains transferts du système socio-fiscal sur le niveau de vie : résultats sur 2016 à partir du modèle de microsimulation Ines (Cahier de variantes) » - Maëlle Fontaine et Michaël Sicsic

F1805 : « Family, Firms and the Gender Wage Gap in France » - Elise Coudin, Sophie Maillard et Maxime Tô

F1803 : « Trois versions du taux d'effort en matière de logement » - Pascal Godefroy

F1802 : « Heterogeneous exposure to labor earnings risk » - Pierre Pora et Lionel Wilner

F1801 : « L'espérance de vie par niveau de vie Méthode et principaux résultats » - Nathalie Blanpain

F1708 : « Les données fiscales de l'EDP : une nouvelle source d'informations sur les couples et les familles ? » - Vianney Costemalle

F1707 : « La situation matrimoniale dans le recensement : impact de la refonte du questionnaire de 2015. » - Guillemette Buisson

F1706 : « Indices de prix à la consommation » - Patrick Sillard

F1705 : Effet d'un choc d'inflation sur le revenu disponible et ses composantes deux ans après : une approche par microsimulation - Anne-Lise Biotteau et Maëlle Fontaine

F1704 : Scanner data and quality adjustment - Isabelle Léonard, Patrick Sillard, Gaëtan Varlet et Jean-Paul Zoyem

F1703 : Les structures familiales en France : comparaison entre le recensement, l'enquête famille et logements et l'enquête emploi - Guillemette BUISSON et Aude LAPINTE

F1702 : Projections de la population active à l'horizon 2070 - Malik KOUBI et Anis MARRAKCHI

F1701 : Les taux marginaux effectifs de prélèvement pour les personnes en emploi en France en 2014 - Juliette FOURCOT et Michaël SICSIC

F1606 : Projections de population 2013-2070 pour la France : méthode et principaux résultats - Nathalie BLANPAIN et Guillemette BUISSON

F1605 : Les durées passées en famille monoparentale - Méthode d'estimation des durées et résultats - Vianney COSTEMALLE

F1604 : ESeG = European Socio economic Groups - Nomenclature socio-économique européenne - Monique MERON, Michel AMAR, Charline BABET, Milan BOUCHET-VALAT, Fanny BUGEJA-BLOCH, François GLEIZES, Frédéric LEBARON, Cédric HUGRÉE, Étienne PENISSAT et Alexis SPIRE

F1603 : Catégorie sociale d'après les déclarations annuelles de données sociales et catégorie sociale d'après le recensement : quels effets sur les espérances de vie par catégorie sociale ? Comparaison entre les déclarations annuelles de données sociales et les recensements de la population. Comparaison de méthodes d'estimation des espérances de vie - Vianney COSTEMALLE

F1602 : L'espérance de vie par catégorie sociale et par diplôme - Méthode et principaux résultats - Nathalie BLANPAIN

F1601 : Échantillonnage des agglomérations de l'IPC pour la base 2015 - Laurence JALUZOT et Patrick SILLARD

F1508 : Worker-firm matching and the family pay gap: Evidence from linked employer-employee data - Lionel WILNER

F1507 : Effet des nouvelles mesures sociales et fiscales sur le niveau de vie des ménages : méthodologie de chiffrage avec le modèle de microsimulation Ines - Mathias ANDRÉ, Marie-Cécile CAZENAVE, Maëlle FONTAINE, Juliette FOURCOT et Antoine SIREYJOL

F1506 : Nowcasting du taux de pauvreté par la micro-simulation - Maëlle FONTAINE et Juliette FOURCOT

F1505/376-501 : Bilan du projet EDP++ - division Camap et division Enquêtes et études démographiques

F1504 : Contrôles des rémunérations dans les déclarations annuelles de données sociales (DADS) - Une analyse exploratoire pour améliorer la détection des points atypiques - Claire JACOD

F1503 : Précision de l'enquête Patrimoine 2010 - Pierre LAMARCHE et Laurianne SALEMBIER

F1502 : Pourquoi l'indicateur de pauvreté en conditions de vie baisse malgré la crise économique ouverte en 2008 ? Jean-Louis PAN KÉ SHON

F1501 : Évolution de la population de la France entre 1981 et 2011 : contributions de la fécondité, de la mortalité, du solde migratoire et de la structure de la pyramide des âges - Catherine BEAUMEL et Pascale BREUIL-GENIER

F1410 : "Personal network" and retirement: Is retirement bad for friendship and good for family relationships ? Anne LAFERRÈRE

F1409 : Retraités mais pas en retrait : La retraite pousse-t-elle à de nouvelles activités ? Anne LAFERRÈRE

F1407 : Production "aval" de l'enquête emploi en continu EEC2 2013 - 20XX - Fabien GUGGEMOS

F1406 bis : La constitution de l'échantillon démographique permanent de 1968 à 2012 - Stéphane JUGNOT

F1405 (tome 1) : Hommes et femmes vivant en couple en 2009, 1999 et aux recensements précédents - Fabienne DAGUET

F1405 (tome 2) : Hommes et femmes vivant en couple en 2009, 1999 et aux recensements précédents - Fabienne DAGUET

F1404 : L'addition est-elle moins salée ? La réponse des prix à la baisse de TVA dans la restauration en France - Quentin LAFFÉTER et Patrick SILLARD

F1403 : Estimer les flux d'entrées sur le territoire à partir des enquêtes annuelles de recensement - Chantal BRUTEL

F1402 : Une rotation de la main d'œuvre presque quintuplée en 30 ans : plus qu'un essor des formes particulières d'emploi, un profond changement de leur usage - Claude PICART

F1401 : Calculs statistiques de stock et de flux sur la révision électorale 2012-2013 - Christelle RIEG