JMS Journées de méthodologie statistique de l'Insee 2018

L'ÉCHANTILLON DÉMOGRAPHIQUE PERMANENT A 50 ANS : RETOURS SUR UN DISPOSITIF STATISTIQUE ORIGINAL.

Sébastien DURIER (*)

(*) Insee, Direction des statistiques démographiques et sociales

sebastien.durier@insee.fr

Mots-clés : sondage, appariement, croisement de sources

VERSION PROVISOIRE

Résumé

L'échantillon démographique permanent est un panel d'individus produit par l'Insee depuis 50 ans. Il consiste à compiler au fil du temps, pour chacun des individus retenus dans l'échantillon, les informations issues de plusieurs sources statistiques (recensement généraux puis enquêtes annuelles de recensement, bulletins d'état-civil, fichier électoral, panel « tous salariés », données fiscales) . On profite de cette date anniversaire pour faire un point sur ce dispositif statistique original. En particulier, si l'EDP a régulièrement été présenté (voir bibliographie) avec comme objectif de préciser son histoire, son contenu et les usages que l'on peut faire, les aspects proprement méthodologiques sont plus rarement abordés et beaucoup moins développés.

On essaiera notamment de donner des éléments de réponse aux questions suivantes :

- quelles sont les spécificités et les conséquences de la méthode d'échantillonnage qui est basée sur la sélection de jours de naissance dans l'année ?
- comment les données des différentes sources sur lesquelles s'appuie l'EDP sont-elles appariées ? Quelle est la qualité qui en résulte ?
- doit-on utiliser des pondérations pour réaliser des études à partir de l'EDP ?

L'échantillon de l'EDP est formellement un sondage en grappe (chaque jour de naissance représentant une grappe) mais qui est généralement assimilé à un sondage aléatoire simple. Peut-on néanmoins déplorer l'existence d'effet de grappe ? Les naissances sont-elles aléatoirement réparties dans l'année, indépendamment de l'année et du lieu de naissance ainsi que des caractéristiques des parents ? Inversement, être né à certaines périodes de l'année a-t-il des conséquences sur la vie future des individus ?

Pour la collecte, l'EDP repose entièrement sur la possibilité d'identifier les individus appartenant à l'échantillon dans les sources qu'il compile. La méthode générale consiste à retrouver à partir de ses nom, prénoms, date et lieu de naissance, un individu EDP dans le RNIPP. Quels sont les taux de succès de cette méthode ? La non identification d'individu EDP, assimilable à une non-réponse, introduit-elle des biais ?

Traditionnellement aucune pondération (ou a minima un facteur correctif de 365/nombre de jours EDP) n'est utilisée par les chercheurs dans leurs études utilisant l'EDP. Ce choix est basé sur l'hypothèse d'un sondage aléatoire simple pour le choix des jours et d'une exhaustivité des sources mobilisées. Avec l'introduction des enquêtes annuelles de recensement, l'usage des pondérations devient primordiale, en particulier dans le cas de l'utilisation en panel des EAR. On montrera comment dans le dernier cas les pondérations adéquates peuvent être obtenues.

Introduction

L'échantillon démographique permanent est un panel d'individus produit par l'Insee depuis 1968. Il consiste à compiler chaque année, pour chacun des individus retenus dans l'échantillon, les informations issues de plusieurs sources statistiques (recensement généraux puis enquêtes annuelles de recensement, bulletins d'état-civil, fichier électoral, panel « tous salariés » et enfin données fiscales et sociales). L'histoire et la construction progressive de ce dispositif statistique a été relatée dans diverses publications : notamment pour les débuts de l'EDP on peut se référer à l'article sur la modernisation de l'EDP au début des années 1990 (Rouault, 1995) ainsi qu'à un document de travail plus récent (Jugnot, 2014) ; pour l'ajout récent des sources fiscales et sociales (panel « tous salariés », dispositifs Filosofi et Fidéli), une description en est fournie dans (Durier, 2016) ou (Costemalle, 2017).
On propose dans ce texte de présenter les éléments méthodologiques propres à l'EDP. L'objectif n'est pas d'être exhaustif mais plutôt de fournir aux utilisateurs de l'EDP un cadre « théorique » simple pour réaliser des études à partir de ce dispositif statistique original. Le cadre proposé s'articule autour de trois éléments clés de la qualité dans l'EDP : la méthode de tirage des individus appartenant à l'échantillon qui est basé sur le jour de naissance (partie 1); la méthode d'appariement des sources entre elles qui est basée sur l'identification au RNIPP des individus EDP présents dans les sources (partie 2) et enfin la prise en compte des conséquences du croisement de plusieurs sources entre elles, en particulier lorsqu'une des sources est une enquête annuelle de recensement (EAR) qui ajoute une phase supplémentaire de sondage (partie 3).

1. La méthode de tirage des individus de l'échantillon EDP

1.1. Description de l'échantillon de l'EDP

Comment définir une méthode d'échantillonnage qui permette de suivre des individus au travers des recensements, mais qui fournisse également un échantillon représentatif à une date donnée de la population résidente ? Pour les initiateurs de l'EDP en 1968, le choix s'est porté sur le jour de naissance dans l'année, et cela quelque soit le lieu de naissance des individus. Prenant comme point de départ le recensement de 1968, la stratégie a consisté à inclure dans l'échantillon tous les individus nés les jours qui ont été choisis (dits jours EDP). Pour ces individus, on dispose grâce au bulletin individuel de recensement des nom, prénoms, date et lieu de naissance. Ensuite, au recensement suivant, on renouvelle cette opération de sélection des individus nés un jour EDP. La plupart d'entre eux sont retrouvés et identifiés dans la « base » issue du précédent recensement[1]. Une partie des individus initiaux ne sont plus là (décédés ou partis hors du champ de couverture du recensement) ou n'ont pas pu être identifiés, mais à l'inverse d'autres individus nés un jour EDP sont recensés pour la première fois (naissance depuis le dernier recensement ou migration vers le champ du recensement). En prenant le jour de naissance comme critère d'échantillonnage, on bénéficie donc d'une méthode de sondage automatique (on n'a pas besoin d'avoir une base de sondage prédéfinie) et auto renouvelé (d'où le terme de permanent qui a été retenue pour le nom de l'EDP). Si l'EDP est au départ intrinsèquement lié aux recensements généraux de la population (en ce sens l'EDP est un panel dans le recensement), l'autre avantage de la méthode d'échantillonnage est qu'elle est applicable à d'autres sources. En fait, toute source disposant des date et lieu de naissance ainsi que des nom et prénoms des individus pourrait être intégrée à l'EDP. En particulier, les bulletins d'état-civil que l'Insee gère pour la gestion du répertoire national d'identification des personnes physiques (RNIPP), peuvent être exploités pour l'EDP. Par exemple, les bulletins de décès[2] (ou à défaut l'information sur le décès d'un individu dans le RNIPP) sont depuis ses débuts intégrés à l'EDP

[1] La méthode pour retrouver, c'est à dire identifier, les individus pour lesquels on dispose déjà d'un dossier sera détaillée dans la partie 2.

[2] Ainsi que les bulletins de naissance ou de mariage où un individu est impliqué.

et permettent notamment de contrôler l'attrition du panel. En effet, un recensement étant par définition exhaustif, un individu EDP qui n'est pas retrouvé à un recensement ultérieur est soit décédé, soit parti hors du champ du recensement. Comme son décès éventuel en France est connu, on peut en déduire la population des émigrants entre deux dates de recensement.

1.1.1. Le choix des quatre jours d'octobre

Une fois retenu le principe du jour de naissance comme critère d'échantillonnage, reste à déterminer le nombre de jours (taux de sondage) et les jours eux-mêmes qui seront valable pour toute la durée de vie de l'EDP[3]). Comment choisir ces jours EDP ? Idéalement, un tirage aléatoire des jours serait la meilleure solution[4]. Pour des raisons pratiques, un choix raisonné a été opéré et initialement, quatre jours de naissance dans l'année ont été choisis, à savoir les 1er, 2, 3 et 4 octobre. Le choix de jours consécutifs peut s'expliquer assez bien par le fait de pouvoir rapidement, en visualisant un bulletin individuel de recensement, repérer un bulletin EDP ; un seul mois de naissance et des numéros de jours simples et consécutifs facilite en effet grandement ce travail. Le choix du mois d'octobre s'explique quant à lui par le fait que d'autres projets à l'époque prévoyait aussi de sélectionner des individus du mois d'octobre (notamment pour le futur panel DADS)[5].
Dans l'hypothèse d'une répartition uniforme des naissances dans l'année, le taux de sondage est alors d'environ 4/365,25 (soit 1,1 %) de la population. A l'époque une légère saisonnalité des naissances est observée, mais elle n'était pas de nature à remettre en cause la méthode d'échantillonnage de l'EDP. Cette saisonnalité a d'ailleurs par la suite perdu de son intensité : « Alors qu'en 1975 il naissait davantage d'enfants au printemps, c'est aujourd'hui en septembre que l'on enregistre le plus de naissances, mais avec des fluctuations nettement moins marquées qu'auparavant. La forme du mouvement saisonnier de l'ensemble des naissances ne révèle plus aujourd'hui de particularités. » (Regnier-Loillier, 2010). De fait, avec un taux moyen de 1,11 %, le taux de sondage effectif de l'EDP est très proche du taux théorique, et les fluctuations que l'on peut observer selon les générations d'avant les années 1970 semblent bien aléatoires (figure 1).

Figure 1 : taux de sondage EDP selon la génération, le lieu de naissance et le nombre de jours EDP

Générations	1900-1919	1920-1939	1940-1959	1960-1979	1980-1999	2000-2016	Ensemble
Taux EDP 16 jours	4,45	4,52	4,52	4,51	4,46	4,43	4,49
taux théorique (*)	*4,38*	*4,38*	*4,38*	*4,38*	*4,38*	*4,38*	*4,38*
taux né en France	*4,41*	*4,46*	*4,43*	*4,43*	*4,39*	*4,42*	*4,42*
taux né à l'étranger	*4,83*	*4,87*	*4,90*	*4,85*	*4,81*	*4,63*	*4,84*
Taux EDP 4j d'octobre	1,11	1,08	1,09	1,11	1,12	1,13	1,11
taux théorique (*)	*1,10*	*1,10*	*1,10*	*1,10*	*1,10*	*1,10*	*1,10*
taux né en France	*1,10*	*1,07*	*1,08*	*1,09*	*1,11*	*1,13*	*1,10*
taux né à l'étranger	*1,24*	*1,12*	*1,13*	*1,17*	*1,16*	*1,20*	*1,16*

(*) taux sous l'hypothèse d'une répartition uniforme des naissance dans l'année

Source : Répertoire national d'identification des personnes physiques (RNIPP), Insee

1.1.2. L'extension à seize jours

Avec l'abandon des recensements exhaustifs décennaux (le dernier ayant eu lieu en 1999) et le passage à un mode de recensement par sondage annuel d'environ 14 % de la population, la taille de l'échantillon de l'EDP aurait été divisée mécaniquement par sept. Il a alors été décidé de multiplier la taille de l'échantillon de l'EDP par quatre. À nouveau un choix raisonné a été opéré pour sélectionner

[3] A. Sauvy parle d'ailleurs d'échantillon fixe pour qualifier l'EDP naissant (Sauvy, 1968)

[4] Éventuellement avec des contraintes pour le rendre optimal, comme une stratification par trimestre.

[5] On avait pour projet à un moment de coupler l'EDP avec ces sources naissantes. Le projet pour le panel DADS ne verra cependant le jour que bien plus tard.

les douze nouveaux jours[6] : les 2, 3, 4 et 5 janvier ainsi que les 1er, 2, 3 et 4 des mois d'avril et de juillet ont été ajoutés à la liste des jours EDP. Le taux de sondage théorique est donc depuis les années 2000 avec un échantillon de seize jours EDP, de 16/365,25 (soit 4,4%) de la population. Avec un taux de sondage effectif de 4.49 % de la population, l'écart avec le taux théorique devient significatif (voir sous-partie suivante) mais l'hypothèse d'uniformité des naissances dans l'année peut encore raisonnablement être acceptée pour la plupart des utilisations de l'EDP. C'est pourquoi, l'EDP est généralement présenté comme s'apparentant à un sondage aléatoire simple. Cependant, formellement il s'agit d'un sondage en grappe, chaque jour de naissance représentant une grappe. Deux conditions sont nécessaires pour qu'un tel sondage soit « correct » : la taille des grappes ne doit pas varier de manière trop importante (dans le temps notamment pour l'EDP) et la composition des grappes doit être uniforme pour éviter les effets de grappe. Ces conditions ne sont pas parfaitement remplies par l'EDP, comme nous allons le montrer maintenant.

1.2. Les défauts occasionnés par le choix des jours

On peut schématiser les spécificités de l'échantillon EDP, en séparant les effets « en amont », c'est à dire le fait qu'être né certains jours de l'année informe sur les origines de l'individu (points 1.2.1 et 1.2.2), des effets « en aval », le jour de naissance pouvant influer en partie le devenir des individus (points 1.2.3 et 1.2.4). Enfin la structure des jours choisi peut aussi influer sur la taille des cohortes EDP (1.2.5).

1.2.1. La planification des naissances

On a déjà évoqué ce point précédemment pour rappeler que si une saisonnalité existe, elle reste globalement faible. La question reste cependant de savoir si cette planification est discriminante socialement. Si la volonté des parents de planifier est bien un fait avéré, les différences sociales sont infimes, notamment car la réussite de cette planification est souvent incertaine. Seul le cas des institutrices semble être significatif, celles-ci souhaitant accoucher plus souvent au printemps (Regnier-Loillier, 2010). L'échantillon de l'EDP à quatre jours pourrait ainsi sous-représenter (mais de manière à peine perceptible) les enfants de certaines catégories sociales. Quant à lui, l'échantillon à 16 jours ne souffre pas de ce défaut.

1.2.2. Sur-échantillonnage des individus nés à l'étranger

La conséquence la plus importante du choix des jours concerne la population des individus nés à l'étranger. Ceux-ci ont en effet un taux de sondage EDP légèrement supérieur à celui des individus nés en France pour l'échantillon à 4 jours, et nettement supérieur dans le cas de l'échantillon à 16 jours (figure 1). L'échantillon EDP sur-représente donc cette population. Ceci n'est pas en soi un problème, car on pourrait de fait vouloir augmenter la précision des estimations pour cette population. Mais, il faut dans ce cas en tenir compte et pondérer adéquatement les individus.
La raison de ce phénomène s'explique par la sélection dans l'échantillon de premiers jours du mois. Les individus nés à l'étranger ont en effet plus souvent une date de naissance en début de mois[7]. Ceci résulte du fait que la parfaite connaissance de sa date de naissance par un individu dépend étroitement de la société dans laquelle il vit et en particulier de l'existence ou non d'un état-civil permettant de valider cette date de naissance. Or, la pratique de l'administration française lorsqu'elle est confrontée à un individu ne connaissant pas ou ne pouvant pas prouver sa date de naissance est de lui attribuer une date arbitraire.

[6] Idéalement, un tirage d'un jour pour chacun des douze mois de l'année aurait été envisageable mais aurait pu rendre le repérage des individus EDP plus difficiles.

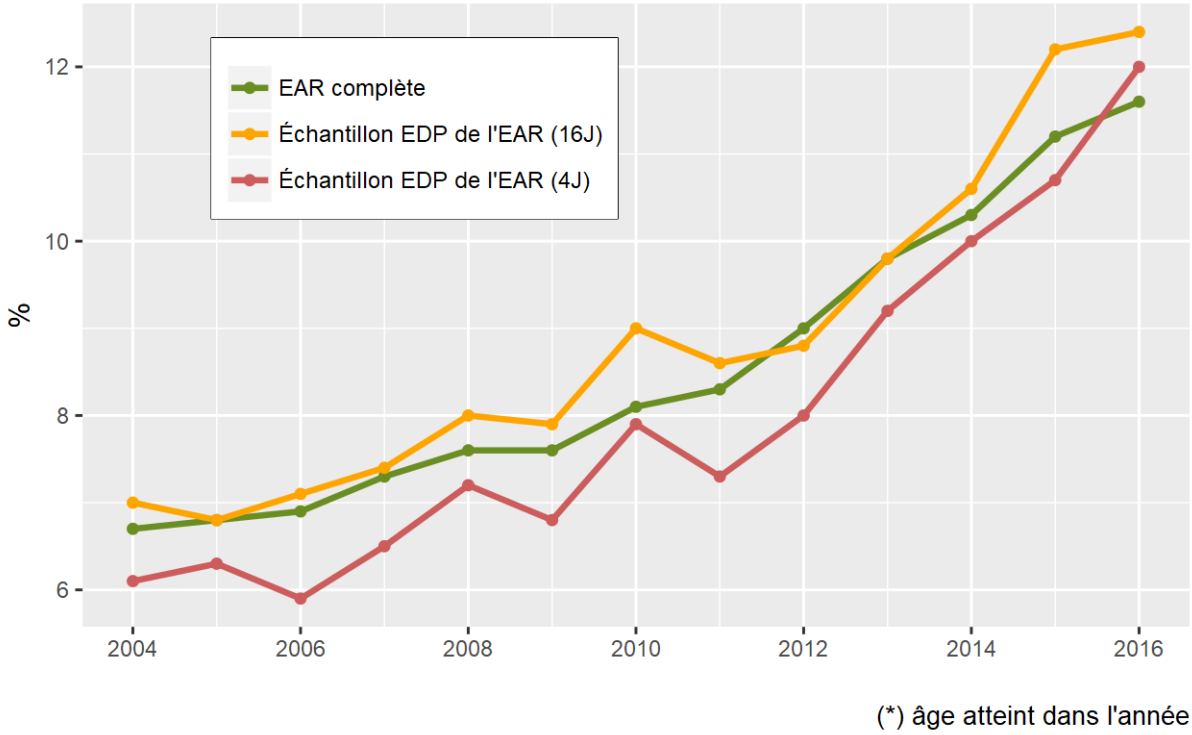
[7] L'effet est le plus fort pour le 1er janvier qui n'a d'ailleurs pas été retenu dans l'échantillon.

On verra dans la partie 2 que cette sur-représentation est « sur-compensée » (mais pas forcément de la bonne manière) par le fait que les individus nés à l'étranger sont plus difficilement identifiables.

1.2.3. Un âge moyen différent

Le choix des jours a aussi comme conséquence que les individus EDP ont un âge moyen différent de celui de leur génération. Plus précisément, avec l'échantillon à quatre jours, les individus EDP étant nés vers la fin de l'année, en octobre, ils sont plus jeunes d'environ trois mois en moyenne. L'ajout des douze jours sur les trois autres trimestres a corrigé en partie ce biais. Cependant, le choix de maintenir le début du trimestre pour sélectionner les jours a, à l'inverse, comme conséquence que les individus EDP sont alors en moyenne un peu plus vieux de un mois et demi environ.

Figure 2 : proportion de bacheliers parmi les 15-19 ans



Source : enquêtes annuelles de recensement (2004-2016), Insee

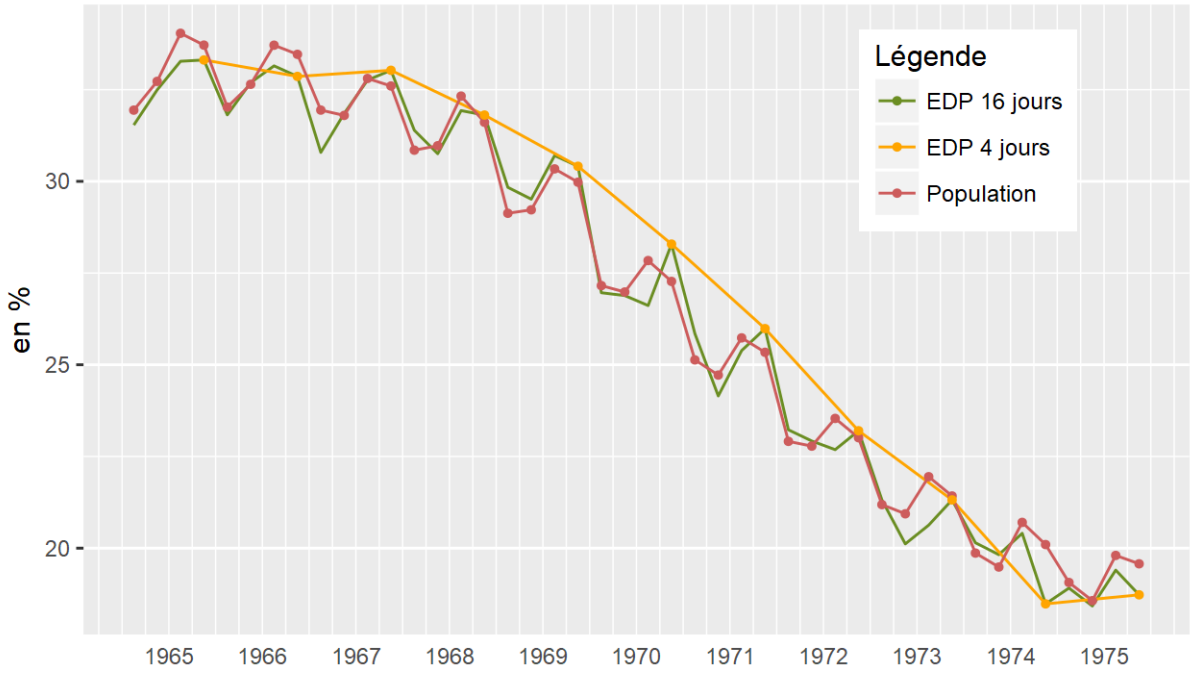
Ces effets, même s'ils restent faibles, doivent être pris en compte. C'est le cas par exemple dans l'étude de la mortalité, plus précisément de l'espérance de vie, où il peut être souhaitable de recaler les résultats obtenus avec l'EDP sur des résultats connus pour l'ensemble de la population (Blanpain, 2015). On peut également en voir une illustration sur la figure 2, où on a représenté l'évolution de 2004 à 2016 de la proportion de diplômés de niveau baccalauréat parmi les jeunes de 15 à 19 ans[8]. Les individus EDP de l'échantillon à quatre jours étant plus jeunes ont une proportion de bacheliers un peu plus faible que l'ensemble des individus des générations concernées, et à l'inverse les individus de l'échantillon à seize jours, légèrement plus âgés, ont une proportion légèrement plus élevée.

[8] Au moment du recensement, c'est à dire en janvier d'une année N, l'âge modal auquel les individus sont bacheliers est de 19 ans en âge atteint dans l'année. Ils ont passé leur baccalauréat en juin de l'année N-1 à 18 ans.

1.2.4. Conséquences sur la vie future des individus de leur jour de naissance

Une des hypothèse faite par un sondage basé sur le jour de naissance, est que ce dernier n'influe pas sur la vie future des individus. Cette hypothèse n'est pas complètement vérifiée. Par exemple, un lien est avéré entre la période de l'année dans laquelle un individu est né (été versus hiver notamment) et le risque de mortalité (Doblhammer, 2002). Toutefois, l'effet est de faible ampleur et demande souvent des tailles d'échantillon très importante pour être mis en évidence.

Figure 3 : proportion de diplômés de niveau V selon la génération et le trimestre de naissance



Source : Enquêtes annuelles de recensement (2010-2014)
Champ : individus âgés de 25 ans ou plus

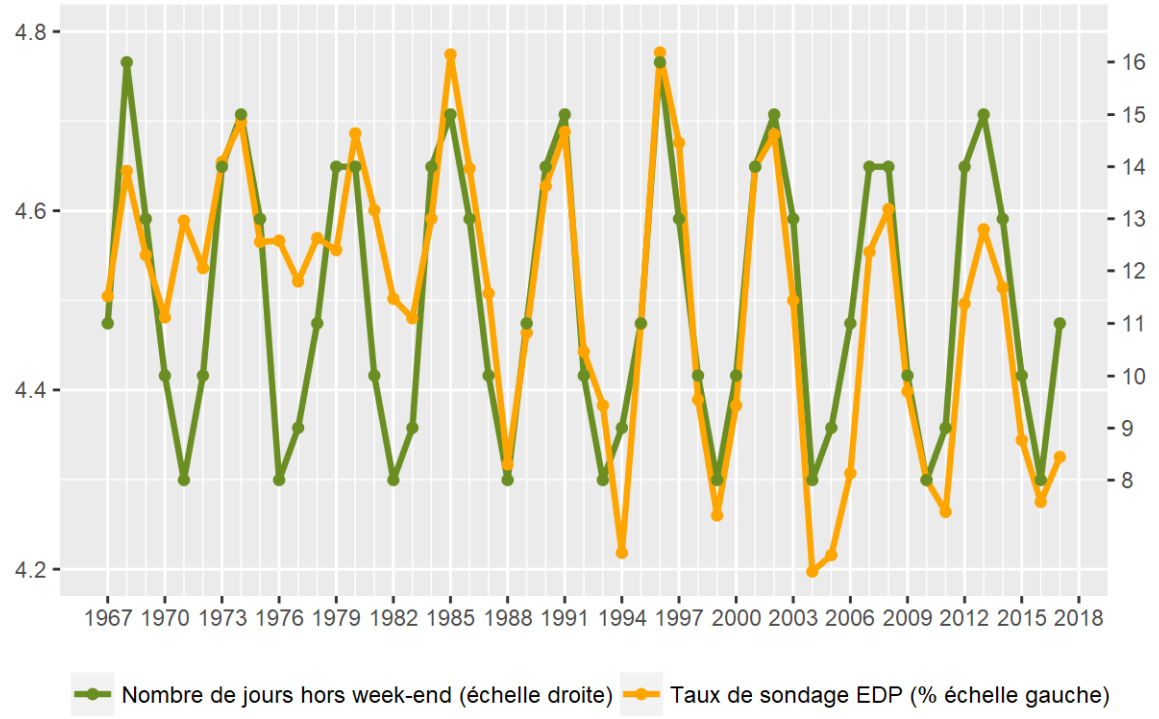
L'autre exemple concerne l' « effet d'âge relatif », c'est à dire le fait qu'être plus âgé que les autres individus d'une cohorte apporte des avantages. Celui-ci a été mis en évidence dans le domaine du sport (Drut et alii, 2014), mais également pour le niveau d'éducation atteint par un individu (Grenet, 2010). Pour ce dernier domaine, on peut donner l'illustration de la proportion de titulaire d'une diplôme de niveau V (CAP et BEP) selon la génération et le trimestre de naissance. Les individus nés au quatrième trimestre de l'année sortent plus souvent avec un diplôme de l'enseignement professionnel que les autres individus de leur génération. La conséquence pour l'échantillon EDP à quatre jours d'octobre est simple : le phénomène est alors totalement invisible, et les individus titulaires d'un diplôme de niveau 5 très faiblement sur-représentés (figure 3). Pour l'EDP à 16 jours, le phénomène est bien pris en compte et l'échantillon EDP tout à fait convenable.

1.2.5. Une fluctuation cyclique de la taille des cohortes EDP

Avoir choisi des groupes de quatre jours consécutifs a aussi une conséquence sur la taille des cohortes des individus EDP. En théorie, cette taille devrait varier uniquement en fonction des tendances générales de la fécondité, auquel s'ajouterait éventuellement une composante de fluctuations aléatoires autour du taux de sondage EDP théorique. En pratique, les cohortes de l'EDP sont bien influencées principalement par la fécondité générale, mais les fluctuations sont quant à

elles « écrasées » par un phénomène cyclique en première approche surprenant (figure 4).

Figure 4 : corrélation entre le taux de sondage EDP et le nombre de jours EDP en semaine



Source : Répertoire national d'identification des personnes physiques (RNIPP), Insee

Ce cycle n'existait pas avant les années 1960 (non représenté sur la figure), et commence à apparaître dans les années 1970, puis s'installe fortement à partir des années 1980. Il s'explique par la présence dans les seize jours EDP d'une année donnée (d'une génération), d'un nombre variable de samedi/dimanche (de zéro jours en week-end comme en 1996 à huit jours en week-end comme en 1999). En effet, la généralisation des accouchements dans les maternités s'est accompagné depuis les années 1970 d'une planification fine via les déclenchements d'accouchement, qui consiste à limiter les naissances en week-end où le personnel médical est moins nombreux.
La taille des cohortes d'individus EDP est ainsi fortement corrélée au nombre de jours en semaine (hors week-end) et varie ainsi de plus ou moins 6 % autour de la moyenne. La plupart du temps l'effet est sans importance pour les études basées sur l'EDP. Néanmoins, lorsqu'on étudie des événements fortement liés à un âge spécifique, il peut s'avérer nécessaire de pondérer adéquatement les différentes générations d'individus EDP. Par exemple, dans la figure 2, où était présenté l'indicateur de la proportion de bacheliers parmi les 15-19 ans, les évolution annuelles sont imparfaites mesurées avec l'EDP, car la taille de la cohorte des individus ayant 19 ans, c'est à dire celle qui contribue le plus à l'indicateur, varie fortement (par exemple en 2010 et en 2016).

Si une sélection raisonnée de jours de naissance n'est pas exempt de défauts, ceux-ci restent cependant minimes et dans la grande majorité des cas peuvent être corrigés en pondérant de manière adéquate l'échantillon d'individu EDP obtenu.

2. Un élément clé de la qualité : l'identification des individus EDP

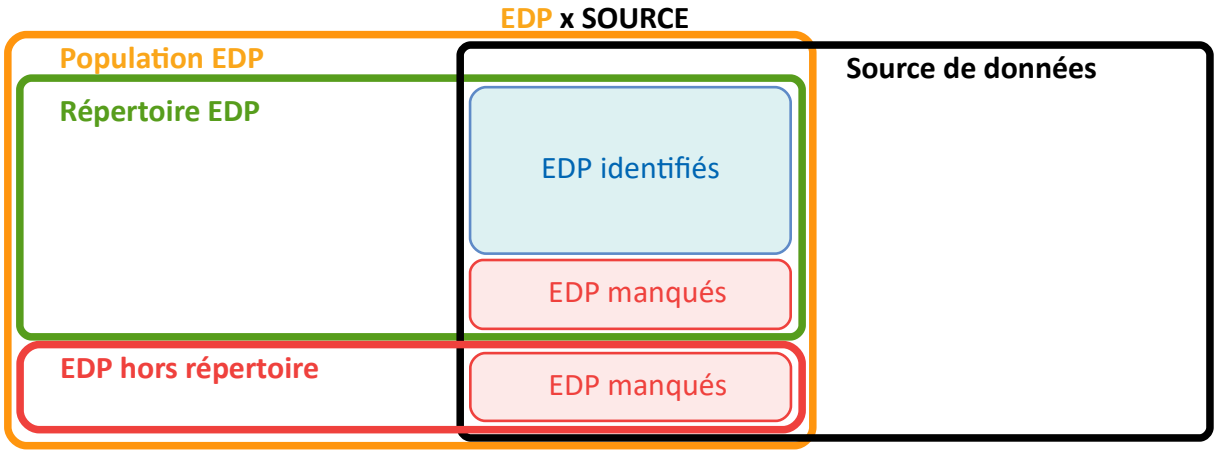
L'objectif du dispositif EDP est de compiler pour les individus sélectionnés les informations provenant de plusieurs sources de données. Cette phase de « collecte » de l'EDP nécessite donc de retrouver dans une source données les individus nés un jour EDP. En pratique, cela nécessite d'abord d'avoir les

informations nécessaires dans la source (nom, prénoms, date et lieu de naissance[9]) pour repérer les individus potentiels, de vérifier ensuite qu'il s'agit bien d'un individu EDP car ceux-ci peuvent déclarer une mauvaise date de naissance et enfin et surtout de confronter l'individu avec les individus déjà « connus » pour lesquels on dispose de données collectées dans d'autres sources. On peut en particulier identifier deux risques majeurs dans la collecte des données dans l'EDP : rater des individus EDP, avec comme conséquence des biais dans la sélection des individus effectivement repérés ; ou attribuer de nouvelles informations au mauvais individu, avec comme conséquence d'avoir de fausses trajectoires.

2.1. La constitution d'un répertoire EDP

Pour garantir la qualité de l'EDP, il faudrait être en mesure de pouvoir déterminer avec certitude si un individu fait bien partie de l'échantillon. La « vraie » population EDP n'est cependant pas connue[10]. Il faut donc soit disposer déjà d'un répertoire des individus avec la qualité la meilleure possible, soit en construire un progressivement avec l'ajout progressif des sources. Pour l'EDP, il s'agit d'un mixte des deux solutions. Il a en effet commencé par une construction progressive d'un répertoire propre (voir Rouault, 1996), avec cependant un lien vers le RNIPP via le numéro d'immatriculation au répertoire (NIR), notamment pour les individus né en France. Depuis les années 2010, la gestion de l'EDP est à l'inverse complètement liée à celle du RNIPP.

Figure 5 : croisement de l'EDP et d'une source de donnée



La situation actuelle est sans doute la meilleure que l'on puisse espérer, néanmoins à tout moment il existe une partie de la population EDP (par exemple résidente) qui échappe au répertoire : ces « hors répertoire », sont dus par exemple au délai de mise à jour du répertoire, ou au fait que certains individus n'ont pas forcément besoin d'un NIR. Il est cependant très peu probable qu'un individu vive durablement sur le territoire français sans finir par être immatriculé au répertoire. On peut donc faire l'hypothèse qu'ils sont très peu nombreux et ne sont pas de nature à nuire à la qualité de l'EDP.

2.2. La couverture de la population EDP

Ces « hors répertoire » (voir figure 5) constituent la première catégorie du défaut de couverture de l'EDP par rapport au source sous-jacente. Cependant, parmi les individus EDP présents dans une

[9] Actuellement, à l'exception du panel « tous salariés », où appariement avec l'EDP se fait sur le numéro d'immatriculation au répertoire (NIR), toutes les sources sont basées sur une identification nominative des individus EDP.

[10] Le RNIPP est lui-même une source, avec laquelle on essaye de capter la population EDP. Cependant, en pratique, on le considère comme référence à laquelle on confronte les autres sources.

source et connus du répertoire, il peut arriver qu'on ne parvienne pas à confirmer leur présence, c'est à dire à les identifier de manière certaine. On parle alors d'échec de l'identification. C'est le deuxième type de défaut de couverture. À l'inverse, les individus EDP bien retrouvés feront partis de l'échantillon utilisable pour les études, à savoir la population EDP couverte par l'échantillon.
On peut dès lors définir un premier indicateur de la qualité dans l'EDP : le taux de couverture de la population EDP qui est égal au rapport entre le nombre d'EDP identifiés sur le nombre d'EDP dans le champ[11] de la source. Celui-ci donne en quelque sorte le nombre attendu d'EDP dans une source ou encore la population EDP qu'on souhaiterait captée dans le champ de la source. Par exemple, au recensement, c'est la population EDP résidente au sens du recensement. Pour les données fiscales, c'est la population EDP résidente au sens de la taxe d'habitation principale ou la population devant faire une déclaration de revenu.

figure 6 : taux de couverture de la population EDP aux RP et EAR

en %	RP 1968	RP 1975	RP 1982	RP 1990	RP 1999	EAR 2004-2008 (*)	EAR 2009-2016
EDP 4J Octobre	89,3	91,7	93,2	94,7	96,6	96,5	95,8
EDP 12J (hors octobre)						80,9	95,4

Source : EDP, RP (1968-1999) et enquêtes annuelles de recensement (2004-2016)
(*) l'extension de 12 jours commence en 2008 et pour cette première année, il n'y a pas eu de reprise gestionnaire d'où le taux de couverture très bas.

Ce taux de couverture peut être estimé pour le dénominateur, soit en appliquant à la population totale un taux de sondage théorique EDP (voir Couet, 2003), soit en prenant la population des individus qui ont déclaré être nés un jour EDP dans la source. Pour les recensements de la population (RP) puis les enquêtes annuelles de recensement (EAR), le taux de couverture est présenté dans la figure 6 pour les deux sous-échantillons de l'EDP : on constate une nette amélioration jusqu'en 1999 du taux de couverture. Depuis, cette date le taux stagne entre 95 et 96 % de la population EDP.

2.3. Mise en œuvre de l'identification dans une source

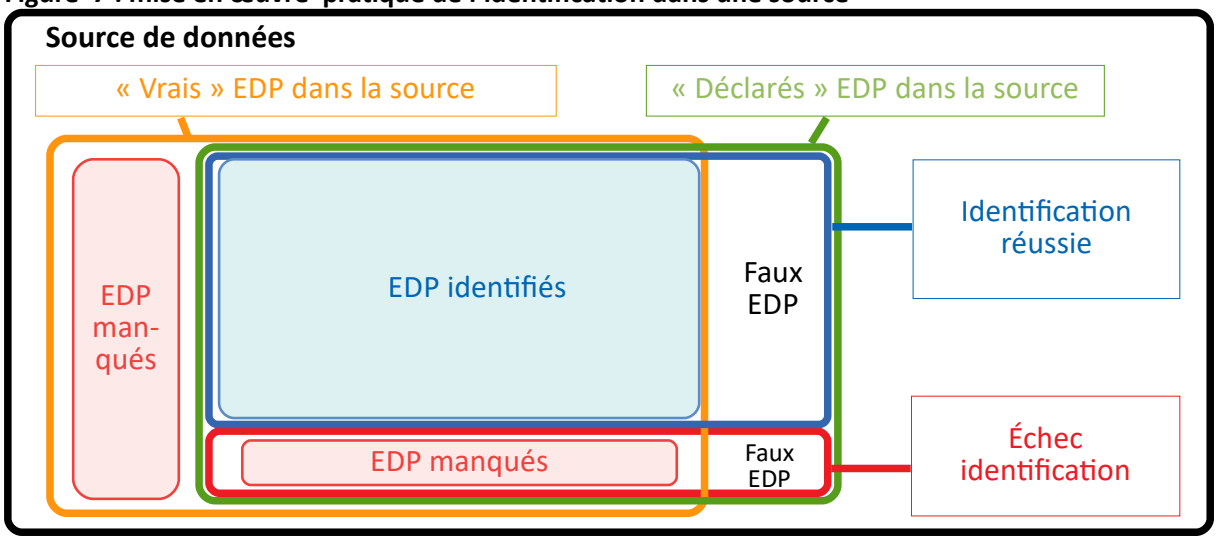
La figure 7 donne le schéma général de la manière dont on procède dans l'EDP pour identifier dans une source les individus EDP. Notons tout d'abord, qu'est inclut dans la « source » la non-réponse ou le défaut de couverture de la source elle-même par rapport à son champ théorique. Par exemple, la non-réponse aux EAR (feuille de logements non enquêtés) fait nécessairement manquer à l'EDP des individus. Comme celui-ci ne fait que prendre ce que les sources lui donnent, il n'y a pas de solution en amont pour corriger ce problème[12].
Toutefois, même lorsque les individus sont dans les « répondants » de la source, il est toujours possible qu'ils aient commis une erreur de déclaration (inversion du jour et du mois de naissance par exemple) ou qu'ils n'aient répondu que partiellement[13]. Ces individus ne sont pas alors dans la population des individus « déclarés EDP » et aucun tentative d'identification au répertoire n'est réalisée pour eux. Cette population pourrait être en partie récupérée si on procédait à l'identification de tous les individus de la source, mais le rapport gain (en nombre d'EDP gagnés)/coût (en temps notamment) ne justifie pas une telle approche.

[11] Cela inclut donc la non-réponse dans la source, voire partie sur l'identification

[12] Il faut le faire via les pondérations en se recalant sur les marges de la source.

[13] Par exemple dans la source fiscale, les personnes à charge mineures n'ont qu'une année de naissance et donc ne peuvent pas être identifiées directement. Une procédure spécifique pour les identifier a été mise en place et des données sont disponibles dans la base Études. Elle utilise les informations connues dans l'EDP sur les parents de l'enfant EDP (date, lieu de naissance, et éventuellement noms et prénoms). Outre que la méthode est imparfaite et variable selon la génération de l'enfant, elle introduit une dépendance entre les sources de l'EDP, qui rend difficile les études.

Figure 7 : mise en œuvre pratique de l'identification dans une source



Si maintenant, on se restreint à la population des individus qui, dans la source, sont « déclarés EDP », on obtient le cœur de la mécanique de l'EDP : l'identification des individus dans le répertoire. Cette procédure était, au début de l'EDP, entièrement manuelle. Depuis les années 1990, l'identification passe par une application informatique ; elle est donc d'abord réalisée automatiquement[14], puis si la procédure n'aboutit pas, un gestionnaire prend le relais pour essayer de retrouver manuellement l'individu. Dans le cas de la source fiscale, ajoutée récemment à l'EDP, la procédure est quant à elle entièrement automatique et sans reprise gestionnaire, et se fait par relâchement progressif des contraintes sur la qualité de l'appariement.
Pour mesurer la qualité de cette procédure, on peut calculer le taux d'identification qui rapporte le nombre d'EDP identifiés au nombre d'EDP déclarés. Cet indicateur, globalement très bon, permet cependant de mettre en évidence des différences significatives selon les caractéristiques des individus. Tout d'abord, les individus nés à l'étranger sont moins souvent identifiés que ceux nés en France (figure 8 et figure 9).

Figure 8 : taux de réussite de l'identification des individus EDP dans les données fiscales (Fidéli)

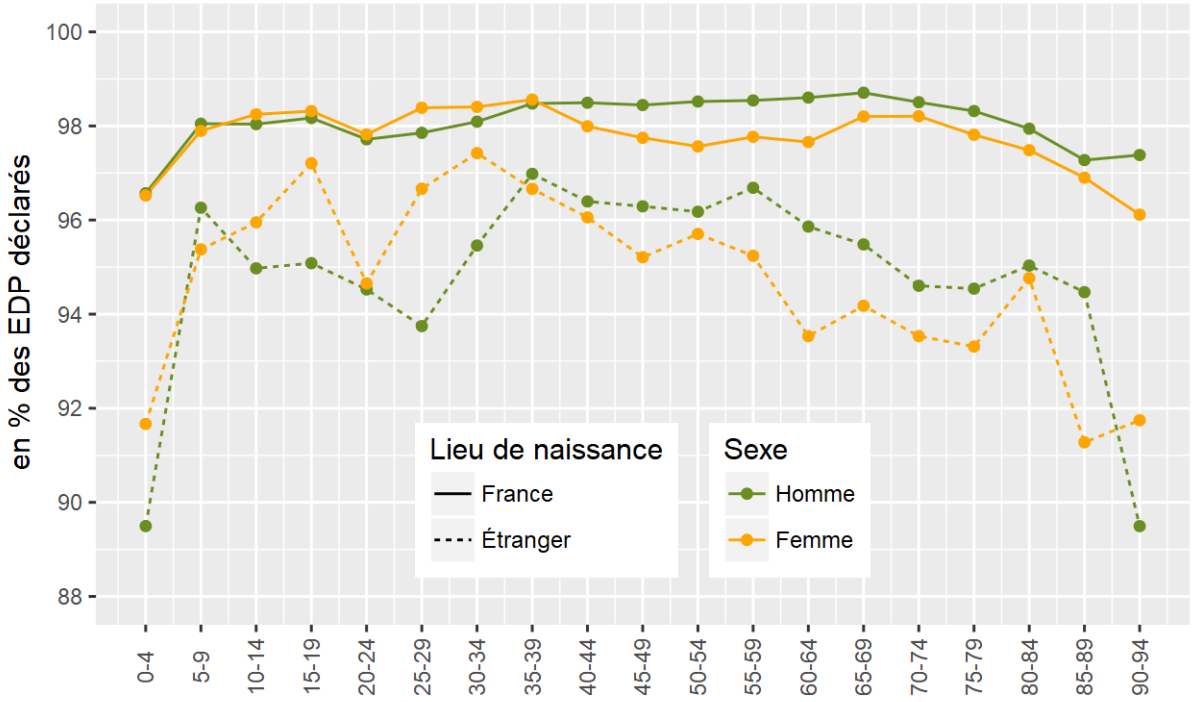
Année de la déclaration fiscale	Nombre d'EDP « déclarés »	Taux d'identi-fication	*dont « faux » EDP*	Taux Homme	Taux femme	Taux né en métro-pole	Taux né dans les Doms	Taux né à l'étranger
2012	2 303 187	96,6	*1,1*	97,4	95,8	98,1	96,6	89,1
2013	2 318 498	96,6	*1,0*	97,4	95,9	98,2	96,8	89,2
2014	2 342 229	96,6	*1,0*	97,4	95,9	98,2	96,4	89,1
2015	2 357 524	96,9	*0,9*	97,5	96,4	98,5	96,5	89,8
2016	2 399 710	97,0	*0,9*	97,6	96,5	98,5	96,8	90,3

Source : base de production de l'EDP 2016.

Cela s'explique par le fait que pour cette population le lieu de naissance est moins discriminant car il s'agit uniquement du pays de naissance et non pas de la commune. Ensuite, on constate des taux d'identification plus faibles pour les très jeunes enfants et pour les personnes les plus âgées. Enfin, bien qu'il soit minime, un écart apparaît entre les hommes et les femmes après quarante ans. Il s'explique par l'ambiguïté liée au nom marital, le RNIPP étant basé sur le nom de naissance.

[14] Pour les bulletins individuels de recensement, environ 83 % de l'identification est réalisée automatiquement.

Figure 9 : taux de réussite de l'identification des individus EDP aux EAR



Source : EDP et enquêtes annuelles de recensement (2011-2016)

La figure 8 montre aussi que dans la population « déclarée EDP » de la source fiscale, environ 1 % s'avère être de « faux EDP », c'est à dire des individus qui ont commis une erreur dans la déclaration de leur date de naissance.

En conclusion de cette partie, il est important de rappeler que le dispositif EDP repose entièrement sur la capacité à identifier correctement les individus EDP dans les sources qui sont mobilisées. De surcroît, cette identification doit être stable dans le temps pour un même individu afin de pouvoir disposer de trajectoires correcte. Le lien du répertoire EDP avec le RNIPP est en ce sens un gage de qualité. Cependant, la procédure d'identification reste imparfaite. Pour les utilisateurs, la stratégie consiste à considérer la phase d'identification, comme un degré supplémentaire de sondage[15], qui peut être corrigé au moyen d'une pondération obtenue par calage sur les marges des sources utilisées.

3. L'intérêt de l'EDP : croiser des sources entre elles

Une fois les individus EDP identifiés dans les différentes sources mobilisées, on dispose donc d'un échantillon représentatif de chacune des sources. L'EDP ne collectant pas d'informations directement auprès des individus, qu'apport l'EDP par rapport aux sources sous-jacentes ? L'intérêt de l'EDP est en fait d'apparier les sources entre elles, et donc de les enrichir mutuellement. Cet enrichissement peut se faire soit en transversal, soit en longitudinal. Par exemple, en croisant les individus recensés à l'EAR 2016 avec les mêmes individus EDP inscrits dans le fichier électoral au début de l'année 2017, on obtient un échantillon représentatif de résidents en France, de nationalité française et inscrits au fichier électoral, ce qui rend possible les études sur les comportements d'inscriptions sur les listes électoral (Durier, 2017).

[15] Plus précisément, la probabilité d'inclusion dans l'échantillon est la probabilité d'être identifié sachant que l'individu EDP est connu du répertoire.

Le croisement ou l'appariement de sources peut également se faire entre différents millésimes de la même source (le recensement de 1990 avec celui de 1999 par exemple). L'intérêt est dans ce cas de « paneliser » la source. Le cadre théorique n'est cependant pas différent de celui où deux sources différentes sont mobilisées.

3.1. Formulation générale du croisement de source dans l'EDP

Le premier point à prendre en compte lors du croisement de source est qu'on obtient un échantillon qui est à l'intersection entre les champs des deux sources. De fait, si les cinq sources de l'EDP ont des champs proches les uns des autres, ils ne sont pas pour autant identiques : pour les RP et les EAR, le champ est celui de la population résidente de France[16] ; pour les bulletins d'état-civil le champ est celui des événements (naissance, mariage, décès) ayant lieu en France[17] ; pour le fichier électoral, il s'agit des individus inscrits au fichier électoral, y compris les individus de nationalité française résidant à l'étranger ; pour le panel « tous salariés », il s'agit des salariés dont les employeurs font une déclaration fiscale en France, y compris donc pour des salariés résidant à l'étranger[18] ; enfin pour les données fiscales le champ retenu est celui des individus présents dans les déclarations fiscales soit au titre de la taxe d'habitation, soit au titre de l'impôt sur le revenu.

Le second point porte sur la question de la pondération. Doit-on utiliser une pondération pour chaque individu de l'échantillon obtenu ? Traditionnellement (jusqu'au RP 1999), les utilisateurs de l'EDP prenaient l'hypothèse générale d'un sondage uniforme (éventuellement en utilisant un facteur d'expansion de 365,25/4) et d'une procédure d'identification correcte (a minima uniforme sur les sous-groupes). Comme les sources (RP et état-civil) étaient exhaustives sur leur champ, la plupart des études n'utilisaient pas de pondération. Comme on l'a vu dans les deux premières parties, si cette pratique était acceptable, des biais éventuels pouvaient néanmoins subsister. La pratique prudente serait donc de munir les individus de l'échantillon obtenu par croisement de sources d'une pondération. Avec l'apparition des EAR dans l'EDP depuis 2004, l'usage d'une pondération devient quasiment obligatoire en raison du taux de sondage différent entre les individus des petites et des grandes communes.

On pourrait formuler la probabilité d'inclusion d'un individu EDP dans le croisement de deux sources de la manière suivante :

$$\pi_i = \pi_{s_1 \cap s_2} * \pi_{EDP} * \pi_{I_1 \cap I_2}$$

avec $\pi_{s_1 \cap s_2}$ probabilité d'inclusion dans l'échantillon formé par le croisement de deux sources (S1 et S2) et cela indépendamment de l'EDP,

π_{EDP} probabilité de sondage EDP, c'est à dire d'être individu EDP, indépendante des sources,

et $\pi_{I_1 \cap I_2}$ probabilité d'un individu EDP d'être identifié sachant qu'il appartient au croisement entre les deux sources.

Pour la presque totalité des croisements de sources dans l'EDP on a $\pi_{s_1 \cap s_2} = \pi_{s_1} * \pi_{s_2}$, car les sources (et éventuellement les tirages d'échantillon à l'intérieur des sources) sont indépendantes entre elles. De même pour la plupart des sources dans l'EDP, on a $\pi_{s_i} = 1$ car les sources sont exhaustives sur leurs champs respectifs (la probabilité d'inclusion vaut 0 sinon pour les individus hors du champ de la source). Actuellement, parmi les sources mobilisées dans l'EDP, seules les EAR ne sont pas exhaustives sur leur champ, et dans ce cas, on récupère dans l'EDP la probabilité d'inclusion des

[16] Plus exactement, il s'agit de la France métropolitaine jusqu'au RP 1999, puis de la France y compris les Doms avec les EAR.

[17] Y compris, les naissances ou les décès d'individus non résidents (des touristes par exemple). Mais à l'inverse, les événements ayant lieu à l'étranger ne sont pas collectés.

[18] Par exemple, les travailleurs frontaliers.

individus dans une EAR, en pratique $\pi_{s_i} = 1/poids\_ea$, à savoir l'inverse du poids de sondage fourni dans les fichiers de diffusions des EAR.
Enfin pour l'ensemble des sources (sauf un cas particulier[19]), on a $\pi_{I_1 \cap I_2} = \pi_{I_1} * \pi_{I_2}$ car les identifications des individus EDP sont réalisées indépendamment les unes des autres. Dans le cas particulier où le croisement a lieu entre deux dates de la même source (par exemple entre le RP1990 et le RP1999) cette condition tient toujours[20].
Si on reprend l'exemple initial du croisement entre une EAR et le fichier électoral, la probabilité d'inclusion d'un individu EDP dans l'échantillon est :

$$\begin{aligned}\pi_i &= \pi_{EAR \cap FE} * \pi_{EDP} * \pi_{I\_ear \cap I\_fe} \\ &= \pi_{EAR} * \pi_{FE} * \pi_{EDP} * \pi_{I\_ear} * \pi_{I\_fe} \\ &= 1/poids\_ea * 1 * \pi_{EDP} * \pi_{I\_ear} * 1\end{aligned}$$

$\pi_{EDP} * \pi_{I\_ear}$ peut être estimée en recalant au préalable l'échantillon EDP sur les marges de l'EAR, ce qui permet de corriger à la fois les défauts du sondage EDP et les défauts d'identification.

3.2. Le cas du panel dans les EAR

L'hypothèse d'indépendance entre sources appariées n'est pas vérifiée dans le cas où on croise deux EAR, ce qu'on pourrait appeler le panel EDP-EAR, c'est à dire les échantillons formés par les individus EDP recensés à deux EAR[21]. Dans la figure 10 sont présentés l'ensemble des panels EAR qui sont disponibles actuellement dans l'EDP.

Figure 10 : effectifs d'EDP recensés en N et proportion (%) recensés à nouveau les années suivantes

EAR	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9	N+10	N+11	N+12
2004	92 094	1,7	2,6	3,1	3,7	58,6	4,6	4,8	5,1	5,3	46,5	5,6	5,7
2005	94 173	1,7	2,5	3,2	3,6	58,1	4,3	4,7	5,0	5,1	46,3	5,3	
2006	95 199	1,7	2,5	3,3	3,7	57,0	4,3	4,6	5,0	5,2	45,4		
2007	95 302	1,7	2,5	3,1	3,5	57,1	4,4	4,6	4,8	4,9			
2008	339 725	1,8	2,5	3,2	3,7	59,0	4,5	4,6	4,8				
2009	389 750	1,7	2,5	3,1	3,6	57,2	4,3	4,5					
2010	391 232	1,7	2,6	3,1	3,6	57,4	4,2						
2011	391 392	1,6	2,5	3,1	3,5	56,8							
2012	389 955	1,7	2,4	3,0	3,3								
2013	393 893	1,6	2,4	3,0									
2014	393 828	1,6	2,4										
2015	394 921	1,6											
2016	391 746												

Source : Échantillon démographique permanent BE2016, Insee
Lecture : le panel effectif entre l'EAR 2004 et l'EAR 2009 est composé de 53 967 individus EDP (58,6 % des 92 094 individu recensés en 2004).

Les échantillons avec un pas de cinq années (N+5 et N+10) sont les plus nombreux, et pour les pas

[19] Il s'agit de l'identification des enfants EDP dans la source fiscale, car on utilise des informations provenant des autres sources, notamment l'état-civil pour retrouver les parents de l'enfant EDP.

[20] On considère que la non identification est un processus aléatoire.

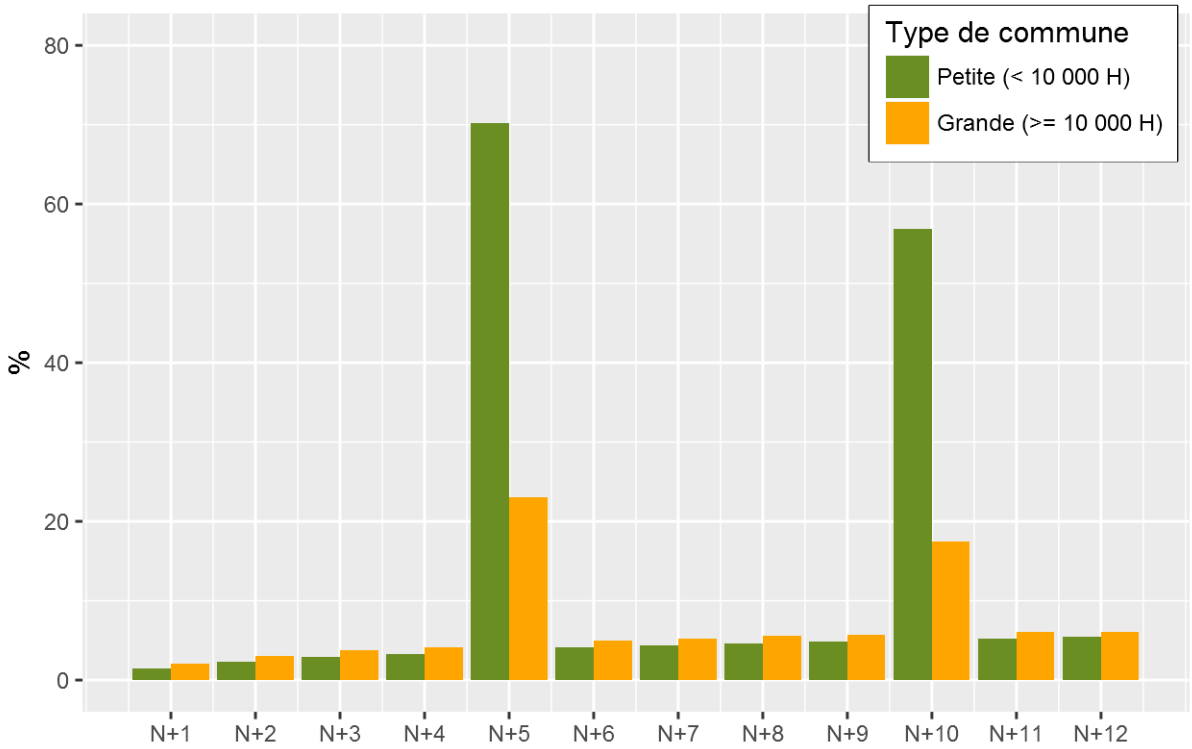
[21] Indépendamment de ce qu'ils ont fait entre les deux dates.

intermédiaires, on constate une augmentation régulière de la taille des échantillons. Comment expliquer ces faits ? Et comment dès lors calculer la valeur de $\pi_{EAR_t \cap EAR_{t'}}$, où t et t' sont deux millésimes des EAR ? Pour cela il faut commencer par rappeler le plan de sondage des EAR.

3.2.1. Rappel du plan de sondage des EAR

Pour les EAR, le plan de sondage (voir Jacob, 2010) est stratifié par région, puis à l'intérieur des régions, les logements sont répartis en trois strates : les petites communes (PC < 10 000 habitants) dont les unités primaires sont les communes elles-mêmes ; les grandes communes (GC ≥ 10 000 habitants) dont les unités primaires sont les adresses (issues du répertoire des immeubles localisés, le RIL) ou les districts pour les grandes communes des Doms ; et enfin la strate des communautés dont les unités primaires sont les communautés elles-mêmes. Dans chaque strate, toutes les unités primaires sont réparties aléatoirement dans cinq groupes de rotation[22]. Chaque année dans un groupe de rotation, il y a une seconde phase de tirage avec une méthode différente pour chacune des strates : pour les PC l'ensemble des logements de la commune sont enquêtés ; pour les grandes communes environ 40 % des adresses de la commune sont échantillonnées[23] ; enfin pour les communautés l'ensemble des communautés du groupe de rotation sont enquêtées. Après cinq ans, on réalise à nouveau la seconde phase de tirage dans le 1er groupe et ainsi de suite.

Figure 11 : proportion d'individus EDP retrouvés au fil du temps selon le type de commune en N



Source : Échantillon démographique permanent BE2016, Insee
Champ : ensemble des EAR de 2004 à 2016.

Pour le panel EDP dans les EAR, les conséquences du plan de sondage des EAR sont les suivantes : un individu qui ne change pas de logement ne peut être à nouveau recensé que tous les cinq ans. En effet, l'attribution d'un logement à un groupe de rotation est fixe. De même, un individu qui change

[22] Pour les communautés, la répartition n'est pas aléatoire mais obéit également à des objectifs pratiques de collecte.

[23]

de logement à l'intérieur d'une même petite commune, à l'intérieur d'une même adresse de grande commune ou à l'intérieur du même district, ne pourra être interrogé que tous les cinq ans. Pour les autres individus changeant de logement, leur nouveau logement peut appartenir aléatoirement à l'un des cinq groupes de rotation, ils pourront donc être recensés en N+1, …, N+5.
En N+1, …, N+4, le panel ne comporte que des individus ayant connu une mobilité résidentielle et les échantillons ne sont donc pas représentatifs de l'ensemble des individus présents aux deux dates (il manque les « immobiles »), ce qui explique notamment leur faible taille. Par contre, pour tous les panels avec un pas de cinq ans (N+5, N+10, …), tous les cas de figures sont bien présents et l'échantillon est bien représentatif de la population présente aux deux dates. Mais, comme on peut le voir sur la figure, les individus des petites communes en N, déjà sur-représentés en transversal, le sont également en longitudinal. Il nous faut donc trouver une pondération adéquate.

3.2.2. La pondération du panel EDP-EAR

Pour les échantillons N/N+1 (N+2, …, N+4), comme l'individu EDP a changé de groupe de rotation, on peut considérer qu'il y a indépendance entre le tirage des deux logements de l'individu, ainsi la probabilité d'inclusion dans l'échantillon est : $\pi_{EAR_t \cap EAR_{t'}} = \pi_{EAR_t} * \pi_{EAR_{t'}}$ (cas A).
Pour le cas N/N+5 ou N/N+10, le logement étant toujours dans le même groupe de rotation, l'indépendance du tirage des deux logements n'est plus garantie. Il faut alors détailler l'ensemble des cas de figure (Ardilly, 2018), ce qui peut être résumé de la manière suivante :
- si les deux logements en N et N+5 de l'individu EDP sont dans la même petite commune, dans la même adresse de GC , dans la même communauté, ou enfin dans le même district pour les GC des Doms, la première phase de sondage (l'attribution à un groupe de rotation) est la même, et donc pour la sélection du logement en t', seule la deuxième phase de sondage doit être prise en compte. En pratique, cela consiste à prendre comme probabilité d'inclusion :

$\pi_{EAR_t \cap EAR_{t'}} = \pi_{EAR_t} * \pi_{EAR_{t'}} * 5$ (cas B)

- pour les autres autres cas des panels N/N+5 ou N/N+10, les deux logements peuvent être considérés comme issus de tirages indépendants. C'est avéré lorsque les deux logements appartiennent à des strates différentes (deux régions par exemple), mais n'est qu'une approximation à l'intérieur d'une même strate car le sondage est un sondage équilibré (par exemple, pour le tirage des petites communes d'une même région). Si on accepte cette approximation, on peut utiliser la formule présentée plus haut :

$\pi_{EAR_t \cap EAR_{t'}} = \pi_{EAR_t} * \pi_{EAR_{t'}}$ (cas A)

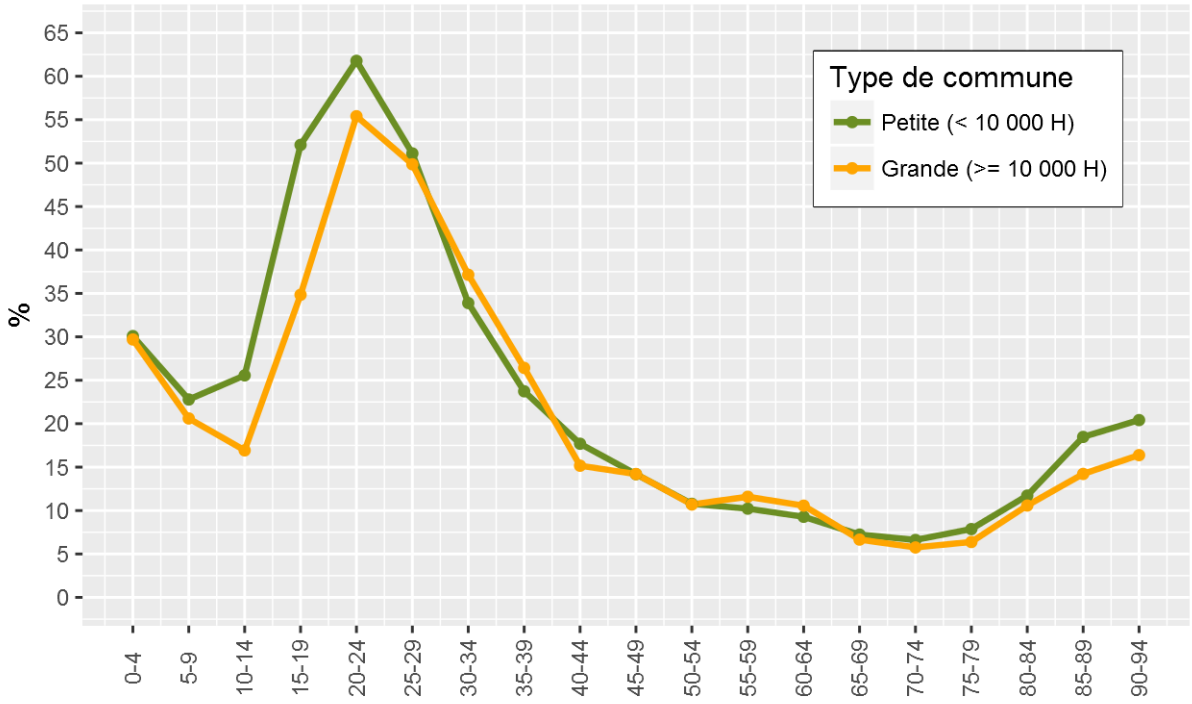
figure 12 : mobilité entre N et N+5 en % selon le type de commune en N

Panel	Même commune en N+5				Commune différente en N+5			
	PC en N	GC en N	*dont même adresse(*)*	*dont changement d'adresse*	PC vers PC	PC vers GC	GC vers PC	GC vers GC
2004-2009	76,2	74,8	*52,9*	*21,8*	14,7	9,0	12,9	12,4
2005-2010	76,8	75,4	*52,6*	*22,7*	15,5	7,7	11,2	13,4
2006-2011	77,0	74,1	*55,4*	*18,7*	15,0	8,1	12,4	13,5
2007-2012	76,1	73,4	*53,6*	*19,8*	14,9	9,2	11,2	15,6
2008-2013	76,8	74,0	*54,6*	*19,4*	14,5	8,9	11,2	15,1
2009-2014	76,7	74,2	*55,2*	*19,0*	14,4	9,2	11,1	14,6
2010-2015	76,6	74,8	*55,0*	*19,9*	15,1	8,6	11,5	14,3
2011-2016	77,0	75,0	*55,5*	*19,5*	14,8	8,4	10,4	14,9

Source : Échantillon démographique permanent BE2016, Insee
(*) ou district pour les Doms. PC = Petite Commune. GC = Grande Commune

Dans la figure 12, sont présentés en utilisant les pondérations les mobilités des individus en rapport au plan de sondage pour les panels N/N+5. En moyenne les trois quarts des individus présents à la fois en N et en N+5 résident toujours dans la même commune, et cela aussi bien pour les habitants des petites ou grandes communes. Pour les individus des PC on est dans le cas B. Pour les individus des grandes communes, un peu plus de 50 % des habitants n'ont pas changé d'adresse dans la même commune (cas B). Les autres qui sont restés dans la même grande commune sont donc dans le cas A. Tous les individus qui ont changé de commune sont dans le cas A. parmi ceux-ci, deux tiers des individus ayant quitté une petite commune ont changé pour une autre petite commune. La situation est inverse, bien que moins marqué, pour les individus ayant quitté une grande commune.

figure 13 : proportion de changement de commune selon l'âge et le type de commune en N



Source : Échantillon démographique permanent BE2016, Insee

La mobilité résidentielle est fortement lié à l'âge des individus, et le panel EDP-EAR offre une outil intéressant pour ce type d'analyse. La figure 13 montre ainsi que les changements de commune de résidence sont plus important pour les jeunes de 15 à 30 ans, et diminuent ensuite fortement avec l'âge. On constate cependant que les très jeunes enfants et les personnes les plus âgés sont relativement plus mobiles.

3.2.3. La dynamique du plan de sondage pour le panel

La solution pour pondérer les individus du panel dans les EAR est basée sur la stabilité du plan de sondage des EAR, en particulier, la stabilité des groupes de rotation au cours du temps. En pratique, des changements apparaissent au fil du temps.

Il y a tout d'abord la question des logements nouveaux : ceux-ci sont automatiquement pris en compte dans les petites communes (cas B). Pour les grandes communes, les adresses de logements neufs sont affectés aléatoirement sur les cinq groupes de la commune, donc il y a bien à nouveau tirage de 1ère phase (cas A).

Ensuite, la stratification PC/GC est basée sur le nombre d'habitants de la commune qui peut évoluer

au fil du temps et certaines communes peuvent franchir le seuil de 10 000 habitants. Pour les individus qui sont dans la commune à la fois en t et t', les pondérations doivent être adaptées de la manière suivante :
- pour les franchissements à la hausse ; dans ce cas les logements de l'ancienne petite commune sont répartis an cinq groupe de rotation (cas A).
- pour les franchissements à la baisse : la nouvelle petite commune sera affecté à un des cinq groupes (cas A).

Conclusion

Si l'EDP est incontestablement un dispositif statistique original, on a montré que sa méthode d'échantillonnage des individus, si elle n'est pas exempte de défauts, fonctionne correctement. De même, si la collecte des données dans l'EDP, à savoir la procédure d'identification des individus EDP dans les sources, pourrait encore être améliorée, elle affiche des taux de réussite très élevé qui sont un garant de la qualité générale de l'EDP. Enfin, si le passage aux enquêtes annuelles de recensement depuis 2004 a rendu certains croisements de source plus compliqués, des solutions via les pondérations existent. De fait, l'usage d'une pondération dans les études mobilisant de l'EDP devrait se généraliser, notamment parce qu'il permet de parer aux inconvénients de la méthode d'échantillonnage EDP et aux insuffisances de la procédure d'identification.

Bibliographie

Ardilly P., « Sur les pondérations de l'EDP en panel dans les EAR », *note interne,* n° 2018/4373/DG75-L110 (31/05/2018), Insee, 2018.

Blanpain N., « L'espérance de vie par catégorie sociale et par diplôme - Méthode et principaux résultats », *Documents de travail,* n° F1602, Insee, février 2016.

Couet C. , « La couverture de l'échantillon démographique permanent par rapport au recensement de la population de 1999», *note interne,* n°2003/094/F170, Insee, 2003.

Couet C. , « L'échantillon démographique permanent de l'Insee », *Courrier des statistiques,* n°117-119, Insee, 2006.

Costemalle V., « Les données fiscales de l'EDP : une nouvelle source d'informations sur les couples et les familles ? », *Document de travail,* n° F1708, Insee, 2017.

Doblhammer G., « Differences in Lifespan by Month of Birth for the United States: The impact of early life events and conditions on late life mortality », *MPIDR working paper* , n°019, 2002.

Drut B., Duhautois R., « L'effet d'âge relatif. Une expérience naturelle sur des footballeurs », *Revue économique*, 2014/3 Vol. 65, 2014.

Durier S., « Une nouvelle source de données sur la famille : l'EDP enrichi de données socio-fiscales », *Actes du colloque de l'AIDELF*, Strasbourg, juin 2016.

Durier S., Touré G., « Élections de 2017 : 6,5 % des citoyens ont fait une démarche volontaire pour s'inscrire », *Insee Focus,* n° 80, Insee, mars 2017.

Grenet J., « La date de naissance influence-t-elle les trajectoires scolaires et professionnelles ? Une évaluation sur données françaises», *Revue économique*, 2010/3 Vol. 61, p. 589-598.

Héran F., « Présentation générale », *Économie et statistique*, n° 316 et 317, Insee, juin-juillet 1998

Jacob J., « Le calcul des Pondérations de l'enquête annuelle de recensement », *note interne*, n° 2010/621/F520 (19/03/2010), Insee, 2010.

Jugnot S., « La constitution de l'échantillon démographique permanent de 1968 à 2012», *Document de travail, n° F1406*, Insee, septembre 2014.

Mikol F., "Refonte de l'EDP : note sur les conséquences du choix de 12 jours de sélection EDP variables", *note interne*, n°2005/031/F170 (01/08/2005), Insee, 2005.

Régnier-Loilier A., « Évolution de la saisonnalité des naissances en France de 1975 à nos jours », *Population,* 2010/1 (Vol. 65), p. 147-189

Régnier-Loilier A., « La planification des naissances dans l'année : une réalité peu visible en France », *Population*, 2010/1 (Vol. 65), p. 191-206

Rouault D., « L'échantillon démographique a pris un coup de jeune », *Courrier des statistiques*, n°73, Insee, mars 1995.

Sautory O., « L 'échantillon démographique permanent », *Courrier des statistiques,* n° 41, Insee, janvier 1987.

Sauvy A., « Progrès et innovations de l'I.N.S.E.E. en matière de statistique démographique. », *Population*, n°3, 1968.