A renewed source of data on families : the French Longitudinal Survey including tax records.

Sébastien Durier, Insee

Launched in 1967, the permanent demographic sample (« Échantillon démographique permanent » in French, hereinafter referred to as EDP) is an individual panel produced by Insee (the French National Statistics Institute) by combining several data sources. At the beginning, the EDP was based on the population censuses and civil registers[1]. In the 2000s, deep improvements were undertaken, which aimed at adapting it to the new census methodology (comprehensive data collection every 10 years were replaced by annual censuses carried out by sample) and to include other available administrative sources, in particular tax data.
The EDP in its new form offers to researchers a powerful tool for demographic analysis.

The EDP, a longitudinal database build on multiple sources.

In the beginning, EDP combined censuses and civil registers

The original idea of the EDP is to mobilize jointly and across the time, socio-demographic information available in the big sources of data produced by Insee. The first sources included were the population censuses and civil registers (birth, marriage and death). Recently, other sources were added as the electoral register or data stemming from tax and social authorities (see infra). The EDP brings two additional dimensions to the sources: a mutual enrichment (« transversal » approach) and a use of sources as a panel (« longitudinal » approach). For example, in the "transversal" approach, one can combine both census data, where the profession and diploma of the individuals is known, and the register of deaths, to study life expectancies by social groups and by level of education (Blanpain, 2016). The « longitudinal » approach gives the opportunity to analyse mobilities even for small groups of people thanks to the big sample size: for example to follow individuals through the censuses allows to compare their geographical location over long period (Sautory, 1988).
In order to build a longitudinal database as the EDP, it is necessary to define a criterion of sampling applicable each year to all the sources included. The chosen criterion consists in selecting all the individuals who were born on specific days of the calendar year (at present 16 days called "EDP days" are selected)[2], regardless of their year or their place of birth. The sample of the EDP could be well equated to a simple random sample. The advantage of such a process is to guarantee an automatic renewal of the sample via births and immigration and to guarantee at any time a transversal representativeness of the sample. Besides, if the sources from which the information is collected are comprehensive on the scope of the EDP individuals, the attrition of the panel due to the deaths and emigration is totally controlled.
One other essential element of the functioning of the EDP is the identification of the individuals in the included sources. To guarantee this quality element, the EDP bases identification on the French national register for identification of persons (RNIPP). So, when finding in a source an individual born on one « EDP day », we try by means of its last and first names, date and place of birth to find him in the register. If the identification succeeds, either we add the information collected to the "file" of the individual if he or she was already included in the EDP, or we create for him or her a new file. The size of the EDP database is

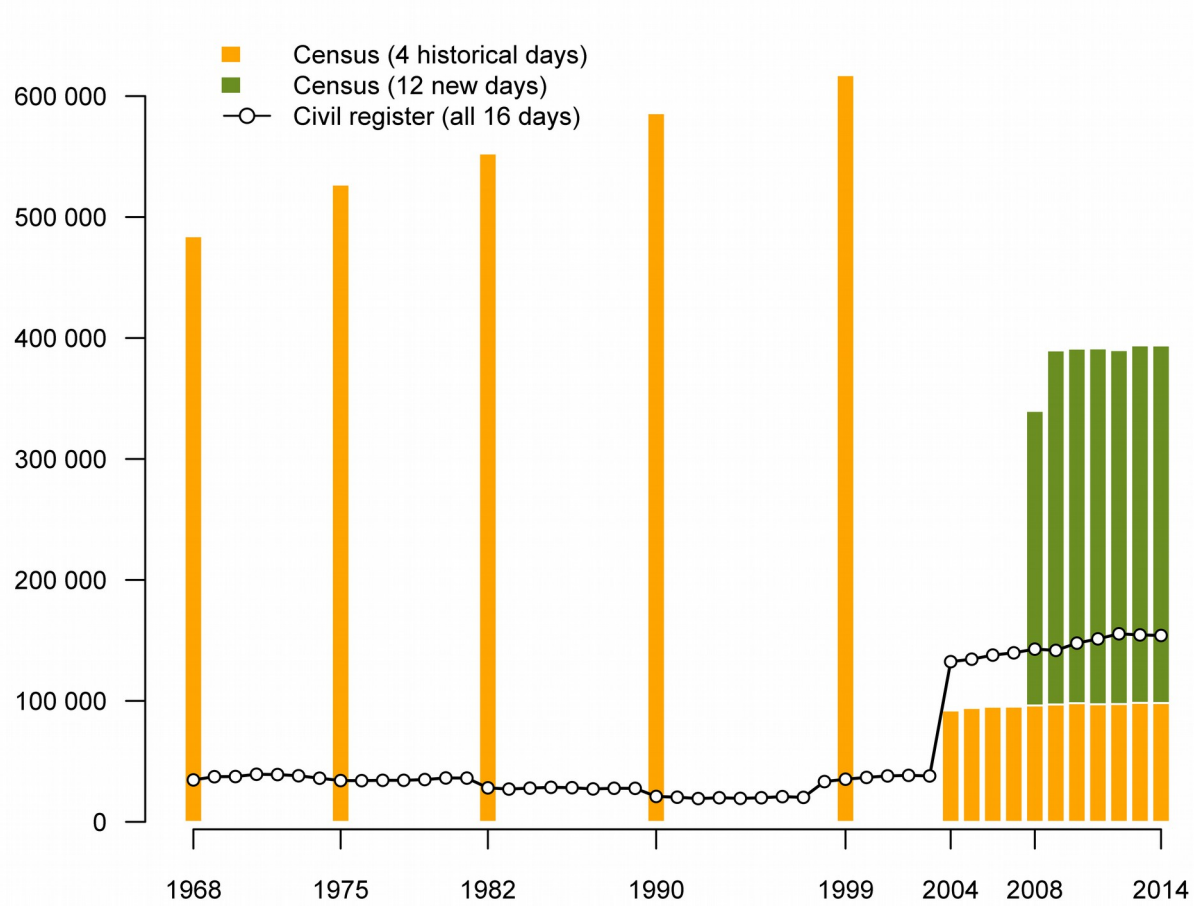
1 This « first » EDP resembles a lot ONS's Longitudinal Study : http://www.ucl.ac.uk/celsius
2 There are the first four days of April, July and October, and the days from 2 to 5 January.

thus continuously increasing year after year with the adding of new individuals (new born or immigrants) and new events.

Longitudinal data over a period of 45 years and more

The follow-up of EDP individuals through the censuses began with the census of 1968. The sample of the EDP was at that time restricted to the individuals born from 1st to 4 October ("4 historical days"). From 1968 till 1999, there are each census year between 500 000 and 600 000 EDP individuals living in metropolitan France for whom information from the census is stored (**figure 1**). The sample of the EDP represents then approximately 1 % (4/365) of the population. Besides the characteristics of the individual such as the diploma or profession, all the information concerning the dwelling and the people living in the dwelling is available.

Figure 1 : Number of census records and civil register records for EDP individuals (1968-2014)



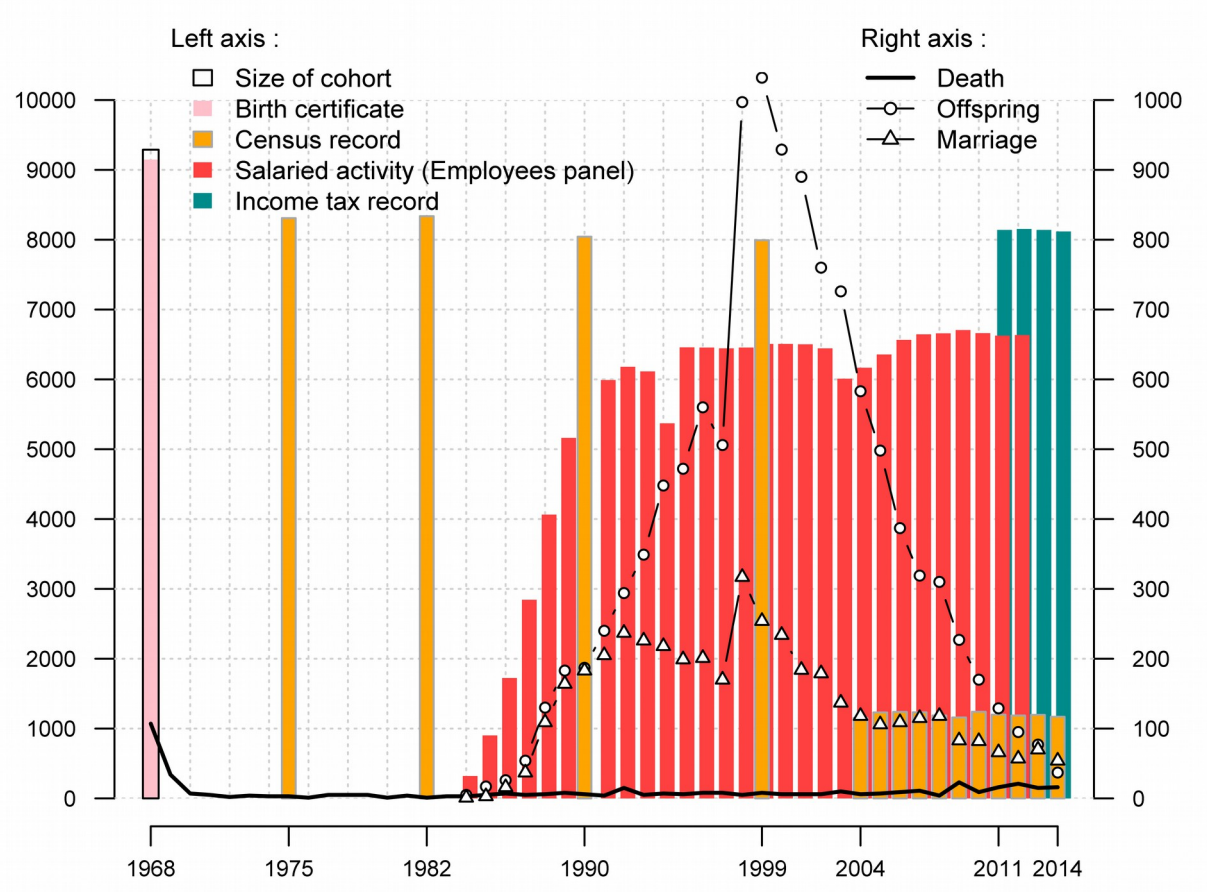
From : échantillon démographique permanent 2014, Insee.

Starting from 2004, the French census is based on a continuous five year rotating sampling survey: every year, approximately 1/7 of the population is surveyed. The advantage for the EDP is that it is now possible to follow cohorts annually, instead of having information only every 10 years. The drawback however is that the sample of the EDP was mechanically divided by seven. To compensate for this phenomenon, the size of the sample was increased fourfold since 2008[3]. Nowadays around 400 000 EDP individuals are identified each year in

3 However, it was not possible to get information from the former censuses for individuals born on the new days, because the nominative data, which are essential to the identification, were not saved at that time.

the census. This change was also the opportunity to include the events of overseas departments.
Civil register is the other key source of EDP from the start. As for the census, the collection of events from civil register began in 1968 for birth certificates, marriage and death[4]. Only the four days of October and the events for metropolitan France were included until 2003. Then, from 2004, the whole scope, including overseas departments, is covered. With the increase in the sample size, we now collect every year approximately 110 000 birth certificates (13 % of all the certificates registered in the year), 20 000 marriages (8.6 % of the marriages) and 24 000 certificates of death (4.4 % of the deaths)[5] (**Figure 1**).

Figure 2 : Available events for the cohort of EDP individuals born on 1st to 4th October 1968 in metropolitan France.



From : échantillon démographique permanent 2014, Insee-Dgfip.
Reading note : amongst the 9289 EDP individuals born on 1st to 4th October 1968 in metropolitan France, 107 died in 1968 (right Y-axis) and 8000 were enumerated in 1999 (left Y-axis).

To give an overview of the available information and the links between the various sources, the **figure 2** shows the set of events collected over the years for the EDP individuals born in 1968 in metropolitan France, for whom the data collection began at their birth. Indeed, for the quasi-majority of the individuals of this cohort, we have information from their birth

4 Judgement of adoption and stillborn child certificate are also collected.
5 Depending on the event, the number of EDP individuals involved varies from one to three. Thus the sampling rate is different. In a death certificate, only the dead person could be EDP; in a marriage, both spouses could belong to the EDP sample; for a birth, the newborn, the father and the mother are possibly born on an EDP day.

certificate. A hundred of them died in 1968, which corresponds to an infantile mortality rate of ten per thousand. Being born in October 1968, the individuals of this cohort couldn't logically be enumerated in the census conducted in March 1968 and their follow-up by a census thus started in 1975. They began to get married, to have children and to work as employees (as monitored by Employees Panel, described farther) in the middle of the 1980s. From 1990 till 1997, events from the civil register was not collected for all the EDP individuals[6]. The number of marriages and the number of children are thus underestimated for this time period[7]. In 2014, the individuals born in 1968 are 46 years old, and most are still alive and still live in France (according to the census or the tax data), but they get married less and less and do not have children any more.

Developing the potential of the EDP

In its historic shape, the EDP clearly showed it was a key datasource for social science, as proved by the long list of articles published on very different topics (Jugnot, 2014). Beyond the adjustment to the statistical environment (new census methodology), the revision of the EDP provided the opportunity to include new administrative sources to enrich the trajectories of EDP individuals and thus to increase the potential of studies. This enhancement of the EDP[8] was made on the one hand by matching the EDP with a panel on employees also produced by Insee and on the other hand by using benefit and tax data from tax income returns and from social services. Without changing the spirit of the EDP, these new sources make it a true multi-purpose panel.

Employees Panel brings comprehensive information for EDP individuals who are wage earners

The Employees Panel began in 1967 upon the annual statements made by employers to tax authorities about their employees. Employees are followed by means of their identification number in the RNIPP. As this number is also available in the EDP, we can retrieve information, until 1967 for 1/8 of the EDP sample[9] and for the rest of the scope the complete follow-up begins in 2002. Since then, and until 2012, more than one million EDP individuals are identified each year in the Employees Panel.
One of the interests of this source is to supply, in a comprehensive way on the scope of EDP individuals who pursue activity as an employed person, information on employment (profession, workplace, business sector of the company) such as they are declared by the employer. The information is sometimes of better quality or is complementary to the one obtained from the employee himself via the census (Costemalle, 2016).
Having all the professional positions of an individual over his or her life on a yearly basis allows to study the links between the careers of individuals and demographic events such marriages and births. For example, it is possible to analyse the impact of the professional activity on fertility (Robert-Bobée, 2006) or on the contrary the impact of family events on wages (Wilner, 2015).

6 Only events for individuals born on the 1st or the 4th October were collected.
7 Information on the death is still available for all individuals thanks to the RNIPP.
8 The use of the electoral register was a part of the renovation of the EDP, but it isn't described in this article.
9 To be precise, individuals born in October of an even year.

Social and tax data: a comprehensive annual source to complete census data

The use of social and fiscal data to enrich households surveys is a more and more common practice at Insee (for example Statistics on Income and Living Conditions). More recently, it was decided to build a localized social and tax database (FiLoSoFi) and an individual and dwelling database (FIDELI) which combine statistical information from different comprehensive administrative data, namely the household income tax return, the local residence tax, and data stemming from CNAF (French social security services). The enrichment of the EDP by social and fiscal data is naturally based on these sources.
In the tax data, the relevant unit is the « family » (specific definition for the tax system) and not the individuals, either for the tax on income, or for the local residence tax, as the amount of tax to pay is calculated on this unit. When in a family unit[10], an individual is identified as belonging to the EDP, all the available information on the dwelling and on all the individuals living in the dwelling (those belonging the family unit or other family units connected with this dwelling) are collected. The "family" links (married, civil solidarity pact[11], divorced, widowed, the dates of the corresponding events, year of births of the dependent children) are essential to set the level of taxation. We thus have access to a large demographic information. Some information (marriage, offspring) are redundant and can be confronted with those already available in the other sources of the EDP, but others (date of divorces, contractings and breaks of civil solidarity pacts) are new in the EDP.
The social and tax data were included for the first time in the 2014 EDP database[12], which was released in March 2016. Tax reference years[13] from 2011 to 2014 are included and every future annual release will include an additional tax reference year. Using this new source in panel is already possible and allows for example to study demographic events cohorts experiment over their life (marriage, divorce, civil solidarity pact, death of one's spouse) by comparing the current marital status of the individuals to the one in previous years[14]. In addition to this promising use of the EDP, the joint use of tax data with other EDP sources allows also new analyses. Let's focus then on three main contributions of the tax data : comprehensiveness on the scope of EDP individuals, additional elements for the analysis of family situations, and the introduction of income dimension, especially the standard of living.

EDP comprehensiveness is back

One of the main interest of the social and tax source for the EDP is to ensure a maximal cover of the scope of EDP individuals, a role played until 1999 by the population censuses. Comprehensiveness allows the EDP to offer samples of important size for studies (for example for fine geographical mobility studies) but also guarantees a control of the panel attrition: an individual not found[15] in the tax data is either dead, or left the national territory. As the deaths are known in the EDP, all the reasons of panel exits are controlled and can be separately studied, in particular the migrations outside France. For example, **figure 2** shows that 8 % of the individuals born in metropolitan France in 1968 and not yet dead were not found in the tax returns of 2014.

10 The family unit could be identical to a household, but is a different concept.
11 In France, a civil solidarity pact is a contractual/legal form of civil union between two adults.
12 The EDP database is released once a year and named after the year of the most recent included events.
13 The year N considered is the year when the tax forms was filled. Geographical location and marital status are valid at the 1st January N. Income and living standard are related to the year N-1.
14 As there is actually only four successive tax reference years, the longitudinal use is limited.
15 An individual can be present in the tax source but not have been identified. This failure of the identification is round 3 % in the tax source for individuals aged 25 years or more.

Table 1 : size of EDP sample in the income tax return and sampling rate from the total French population of each year

Age*	2011		2012		2013		2014	
	n	%	n	%	n	%	n	%
0-19 years	440 292	2.7	453 358	2.8	467 670	2.9	483 270	3.0
20-24 years	145 399	3.6	145 914	3.7	145 218	3.7	143 772	3.7
25 years or +	1 981 290	4.4	2 000 469	4.4	2 017 619	4.4	2 039 471	4.4
All	2 566 981	4.0	2 599 741	4.0	2 630 507	4.0	2 666 513	4.0

From : échantillon démographique permanent 2014, Insee-Dgfip.
* aged reached during the year N of the income tax return.

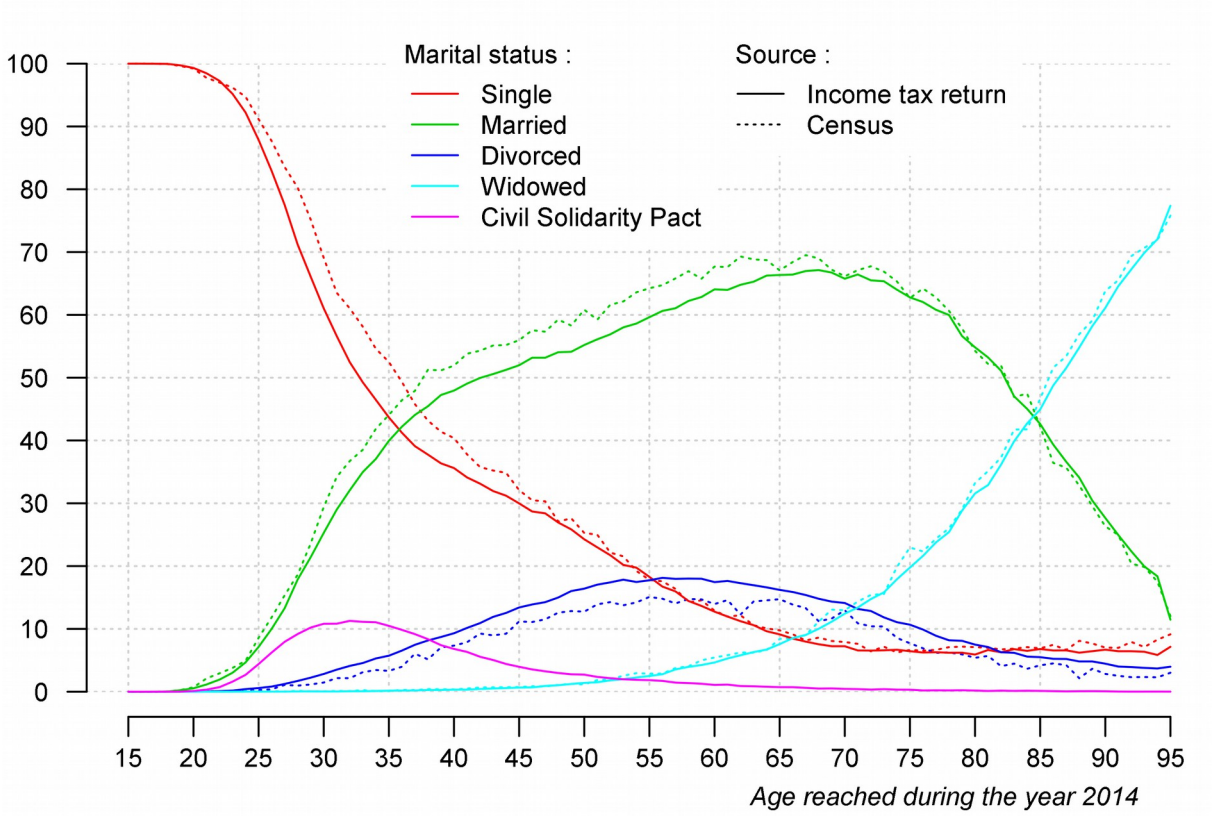
Every year, there are about 2.6 million individuals identified for the EDP use in the tax data, that is 4 % of the total population living in France. The identification is of very good quality for the individuals of 25 years old or more, not so bad for the age bracket of 20-24 years, and of low quality for people under 20 years old. For the latter, only the year of birth is known in the income tax data and the identification is therefore difficult. For this sub-population, there is no comprehensiveness, but the sample remains representative of the corresponding population.

More details on families

In the census, persons are asked for their legal marital status. So, persons separated or engaged in a divorce procedure have to declare themselves as « still » married and the individuals who signed a civil solidarity pact should declare themselves as single. In the tax data, the marital status mixes legal status with de facto situations. First, there is a specific category for civil solidarity pact that does not exist in the census[16]. Married people have to declare in the tax returns that they are married and make a joint tax return with their spouse. Exceptions are however possible: the year of the marriage, spouses can opt for separate returns and individuals engaged in a divorce can, in specific cases, declare themselves as « separated ».
Figure 3 shows age by age, the breakdown of the population by marital status in each source. First, we can see that the shape of the tax data curve is less noisy than the one from the census, as it is based on a sample seven times bigger. The second report is that the two sources offer the same global structure by age. In particular, the proportion of widowers and widows by generation is quite the same. As a third report, we can underline some important differences between the two sources, but which are totally due to differences in the definitions used. In details, married people are more numerous according to the census than according to the tax data, because it includes the separated but not yet divorced people. On the contrary, divorcees are more numerous according to the tax data because separated people are included in this category.

16 It was included in the census in 2015, but is not yet available in the EDP database.

Figure 3 : Global comparison of marital status between the census and the income tax returns, age by age, in 2014



From : échantillon démographique permanent 2014, Insee-Dgfip
Scope : EDP individuals enumerated at the census in 2014 **or** identified in the income tax returns in 2014.
Reading note : at the age of 85 years old, 45 % of individuals are windowed according to both the census and the income tax return.

Finally, the census leads to more single women(men) than the tax source, because this category also includes the majority of the civil solidarity partners, whereas they belong to a distinct category in the tax data. For the latter, it can be useful to analyse how they declare themselves to the census. The EDP allows us to study this point, by cross-tabulating, for the same individuals, information stemming from several sources. In this regard, as **table 2** shows, we can notice that 72 % of the individuals involved in a civil solidarity pact declare that they are single in the census and 7 % are divorced or widowed. But, 11 % do not answer the question (only 4 % for all the individuals), and 10 % consider themselves as being married. This result is in line with the analysis made by the Family and Dwelling survey of 2011 (Breuil-Genier, Buisson, Robert-Bobée and Trabut, 2016).

Table 2 : Comparison between the marital status in the census and the one in the income tax returns in 2014, for each individual present in both sources.

		Marital status according to tax return						
Marital status according to the census	% row	Out of scope*	Single	Married	Divorced (incl. separated)	Widowed	Civil Solidarity Pact	All
	Non-response	25	25	23	8	10	10	**100**
	Single	23	66	1	3	0	8	**100**
	Married (incl. separated)	0	1	96	2	0	1	**100**
	Divorced	0	4	1	91	0	3	**100**
	Widowed	0	1	1	3	95	0	**100**
	All	**9**	**24**	**47**	**9**	**7**	**4**	**100**
	% column	Out of scope*	Single	Married	Divorced (incl. separated)	Widowed	Civil Solidarity Pact	All
	Non-response	11	4	2	3	5	11	**4**
	Single	88	93	1	10	1	72	**34**
	Married (incl. separated)	0	1	97	12	1	10	**47**
	Divorced	0	1	0	73	0	6	**7**
	Widowed	0	0	0	2	93	1	**7**
	All	**100**	**100**	**100**	**100**	**100**	**100**	**100**

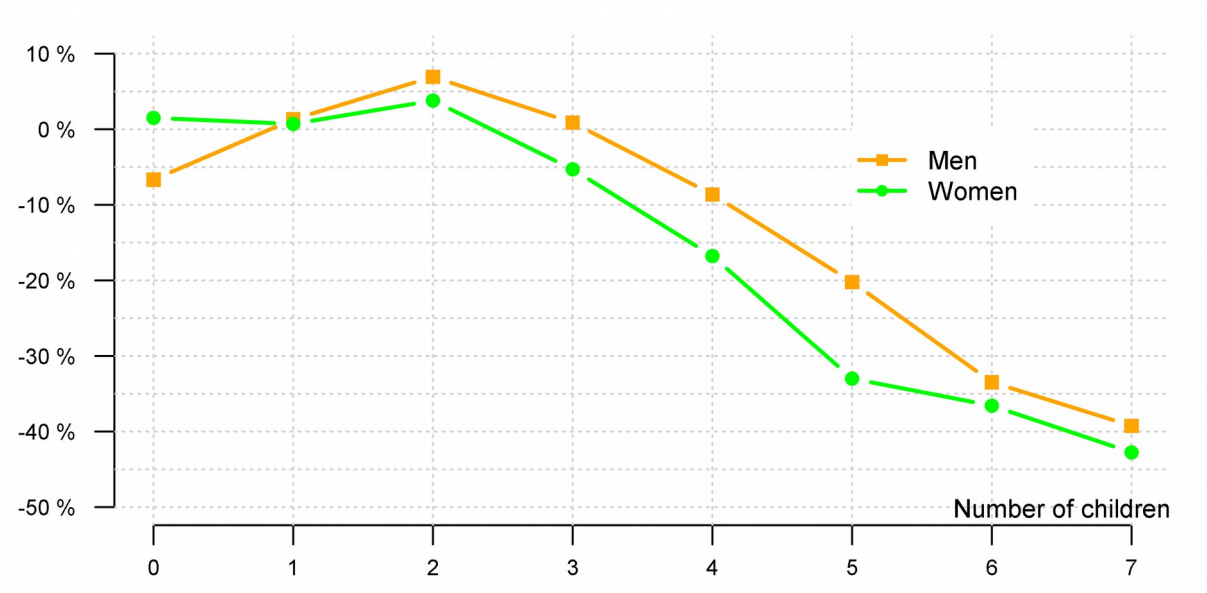
From : échantillon démographique permanent BE2014, Insee-Dgfip.
Scope : 290 000 EDP individuals aged 15 years old or more enumerated at the 2014 census and also identified the same year in the income tax return.
* In the income tax return, marital status of dependent persons in the family unit (aged less than 25 years old) is not known.
Reading note : 96 % of the married persons in the census are also married according to their tax return. Symmetrically, 97 % of married persons according to the income tax return are also married according to the census.

A new variable for analysis : standard of living

Having social and fiscal data (namely almost all income and social allowances perceived by households) allows to enrich the trajectories of the individuals with their standard of living, and we can also even in certain cases focus on some specific components of the income. For example, we can measure the part taken by maintenance or child support in the standard of living after a divorce (Bonnet, Garbinti, Solaz, 2015)[17]. To list the set of the new possibilities offered by EDP is not the purpose of this article and we are going to focus on one specific example .

17 The mentioned study is based on all income tax returns of 2008-2010, with an important preliminary work of « panelization ». The advantage of the EDP in this case is that it supplies data in a suitable format and possibly updated every year.

Figure 4 : Distance to the average standard of living in 2013 depending on the number of children individuals had between ages 15 and 49



From : échantillon démographique permanent 2014, Insee-Dgfip-Cnaf-Cnav-Cccmsa
Scope : individuals born on the 1st or 4th October of the years 1950 to 1969 in metropolitan France and with a tax income return in 2014 (n = 78 000) .
Reading note : individuals who had seven children or more during their life (precisely between 15 and 49 years) have in 2013 a standard of living 40 % lower than the average one of the population.

The X-axis in **figure 4** represents the number of children the men and women born in the year 1950 to 1969 had over their life, in other words, it represents the completed fertility rate of these generations measured through the collection of birth certificates over the period 1968-2014[18]; the Y-axis represents the differences between the standard of living[19] of those individuals calculated with their income-tax return and social-security benefits perceived in 2013 and the overall standard of living in the population .
We can notice that among the individuals aged 45 to 64 years old in 2014, those who had, during their life, a largest number of children have a standard of living much lower than the average one. This negative correlation is clearly marked from and above three children. Also, but to a lesser extent, having had only one or no child is associated with a lower standard of living compared to individuals with two children. Globally, we also notice the lower standard of living of women comparing to the one of men (Insee, 2012), but the gender difference disappears for people with one child and even is inverted for the childless people.

18 For the generation 1950, the follow-up begins only at the age of 18 years old, thus the completed fertility rate is slightly underestimated. Also for the generations 1965-1969, the period 45-49 years was only partially covered, and thus the completed fertility rate is also underestimated. Besides, for the men the fertility could be measured beyond age 49.
19 The standard of living is defined as the disposable income of the household divided by the number of consumption units.

How to access and use the EDP ?

Enriched by social and fiscal data, the EDP becomes a very important tool for researchers in the field of socio-demographic analyses. They can access the data following a specific process : as the EDP is an indirectly identifiable personal database[20], it is available only via a secured procedure, namely after the authorization of the « Comité du secret statistique »[21] and through the Center of Secured Access to Data (CASD)[22].
As it combines several sources, the EDP does not (cannot) propose a "turnkey" product for all the studies, but rather supplies a relatively raw material to the researchers so that they build by themselves their object of studies. As the time needed to investigate the EDP for the first time may be quite long in some cases, Insee tries hard to supply an extremely detailed documentation and organizes regularly meetings of users in order to share experiences.

Bibliography

Blanpain N., « Les hommes cadres vivent toujours 6 ans de plus que les hommes ouvriers », *Insee Première*, n° 1584, Insee, 2016
Blanpain N. «Life expectancy by socio-economic position and diploma in France : the difference between executives and blue-collar workers remains stable », poster session 1, European Population Conference, Mainz 2016.
Bonnet C., Garbinti B. et Solaz A., « Les conditions de vie des enfants après le divorce », *Insee Première*, n°1536, Insee, 2015.
Breuil-Genier P., Buisson G, Robert-Bobée I., Trabut L., « Enquête Famille et logements adossée au recensement de 2011 : comment s'adapter à la nouvelle méthodologie des enquêtes annuelles et quels apports ? », *Économie et statistique*, n° 483-484-485, Insee, 2016
Costemalle V. , « 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 ? », *Document de travail*, n°F1603, Insee, 2016
Insee, « Fiche 4.4 – Niveaux de vie et pauvreté », *Regards sur la parité*, Insee, 2012
Jugnot S., « La constitution de l'échantillon démographique permanent de 1968 à 2012 », *Document de travail*, n°F1406, Insee, 2014
Robert-bobée I., « Ne pas avoir eu d'enfant : plus fréquent pour les femmes les plus diplômées et les hommes les moins diplômés », *France, portrait social*, Insee, 2006
Sautory O., « Près de la moitié de la population a changé au moins une fois de commune en 20 ans », *Économie et statistique,* n° 209, Insee, 1988.
Wilner L., « Worker-Firm Matching and the Family Pay Gap: Evidence from Linked Employer-Employee Data », *Document de travail*, n°F1508, Insee, 2015

20 Neither names or register numbers are available to researchers, but the combination of multiple date of events and locations is often sufficient to identify people.
21 http://www.cnis.fr/cms/Accueil/activites/_trois_comites/Comite_du_secret_statistique
22 https://casd.eu/