Statistical Journal of the IAOS 33 (2017) 311–316
DOI 10.3233/SJI-171054
IOS Press

Commentary paper

Undercount of young children and young adults in the new French census

Laurent Toulemon
French Institute for Demographic Studies, Institut National d'études Démographiques (INED), Paris, France
E-mail: toulemon@ined.fr

1. Introduction

In his paper, William P. O'Hare analyses undercount in the US census and shows that it is particularly prevalent at very young ages. In addition to post-enumeration surveys, he presents results from the Demographic Analysis Method (DA), comparing estimates based on births, deaths and net migration by (sex and) age to census counts. The net undercount reaches 4.6% at ages 0–4 among young children. In the 2000s this undercount is specific to young children.

An important result of the paper is that the undercount of young children is not specific to the US, but is on the contrary present in many countries. His Table 3 also shows that, in the US, the net undercount is higher at ages 0–4 than for any other age group, while in all other countries under examination the undercount rate is higher at ages 20–24.

I strongly support the idea the "the widespread nature of this problem [undercount of young children] suggests that countries may benefit from working together to address this issue." I thank the Editor for giving me the opportunity to update some results presented by Guy Desplanques in 2008 [7] on the French Annual census surveys. The French National Institute of Statistics and Economic Studies (INSEE), in charge of the census, did not publish any documents on undercounts or over-counts in the census, but INSEE offers on its website the possibility of downloading census files at the individual level. I will first present some results of undercount rates by age, and then briefly comment on the similarities and differences between France and the US.

2. How to estimate population of census surveys by sex and age?

Since 2004, the General Population Census has been replaced by a set of annual census surveys. Each 5-year set is merged into a "census" file. Each annual survey takes place in one small municipality out of 5, and in 8% of addresses in large towns (more than 10,000 inhabitants). The total enumerations are weighted using external information (housing tax for small municipalities, housing register for large towns). The census results are thus based on a 5-year moving average. More details on the procedure are available in [3], but suffice here to know that no age-specific post-stratification is performed, so that the age differences are likely to be due to age-specific data collection errors. Census results are 5-year moving averages, but the files include the information on year of birth b and age at 1/1/t, allowing to know the year of survey s, with $s = b + t - 1$. Data from a census dated t include individuals sampled in years $t - 2$ to $t + 2$. From census data dated 2006 to 2013, we thus get information on samples included in the surveys in years 2004 to 2015. This allows us to accurately identify each birth cohort, in order to replicate the Demographic Analysis method.

3. An increasing undercount of young children

For a matter of comparability, results refer to Metropolitan France (excluding overseas departments and are based on the files "*Individus à la region*" from

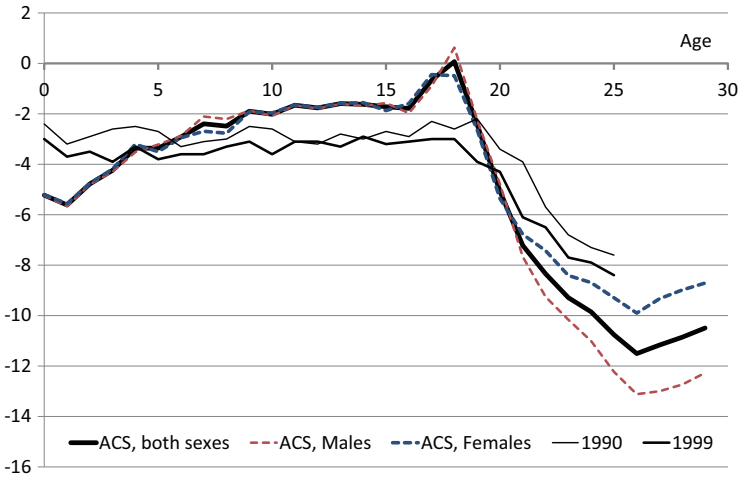

Fig. 1. Net apparent migration from birth to observation in the Population Censuses 1990, 1999, and 2004–2015 Annual Census Surveys (All, Males and Females), for persons born in Metropolitan France, by age at census (percent). Source: Censuses 1990, 1999: Deplanques, 2008, Fig. 2. Annual census surveys 2006–2013: own estimates from downloaded files, average of census surveys 2004–2015.

the Annual Census Surveys [9a–d].[1] As available estimates of net migration by sex and age mainly come from the census, especially at young ages [1,4], undercount can be better estimated with looking at "apparent net native migration". Figure 1 presents the rate of apparent net native migration in the 1900, 1999 and 2006–2013 censuses. This rate is estimated, for each age, as the relative difference between the enumerated native population at census survey and the number of births in France in the same cohort (corrected for the deaths among the cohort). For age 0, the comparison is made between population of age 0 counted in the annual census survey at 1/1/t and births in year t − 1 (minus deaths in t − 1 of children born in t − 1). For age 1, the comparison is made between population of age 1 counted in the annual census survey at 1/1/t and births in year t − 2 (minus deaths in t − 2 and t − 1 of children born in t − 1). Deaths are estimated from cohort life tables. As in the US, the numbers of births and deaths are assumed to be perfectly accurate.

In 1990 and 1999, at all ages before 20, around 3% of children were not present at census, without any age-trend, showing that a small undercount was likely among children. After age 20, the difference widely increased with age. Out-migration may have occurred at these ages, but a large undercount at ages 20–24 was more likely, the difference reaching 8% at age 24. With the new census, the shape of the curve becomes dramatically different. On the one hand, the undercount of young children exceeds 5% at ages 0 and 1; on the other hand, the undercount decreases with age and is almost stable between ages 11 and 16. Re-entries of children born in France and gone just after their birth are possible, but they are likely almost stable with children's age, and may not explain the large decline of undercount between ages 0 and 10. An over-count can be seen at age 18, which implies that double counts are more numerous than omissions, as native net migration may by definition not be positive. At ages 20 and more the undercount rate increases widely, especially for men. A similar undercount of young adults has been documented in many countries, as pointed by O'Hare.

Double counting is more likely to occur with the new census. First, most double counts are invisible for the inhabitants: only a sample of households is included in each Annual Census Survey, so that people who share their time between two homes are rarely included in the same Annual Census Survey in the two dwellings. Second, the INSEE has made large efforts

[1] In the data available on the web, the complete annual survey samples are available, but variables on household family types and Socioeconomic status are constructed for only one household out of four (one out of five from 2014) in small municipalities. Data used in this comment come from those subsamples for which all variables are available. The sampling rate for each annual survey is thus equal to 5% or 4% in small municipalities, and 8% for large towns. As the weights for each annual sample are not provided by INSEE, they were estimated. All census files were merged, and each annual sample was weighted by its total weight in all census files, multiplied by 5 and divided by the number of occurrences of the merged file. The weight of each survey sample was then corrected for its use at the survey year in the present comment, compared to its use at different census years in the original data.

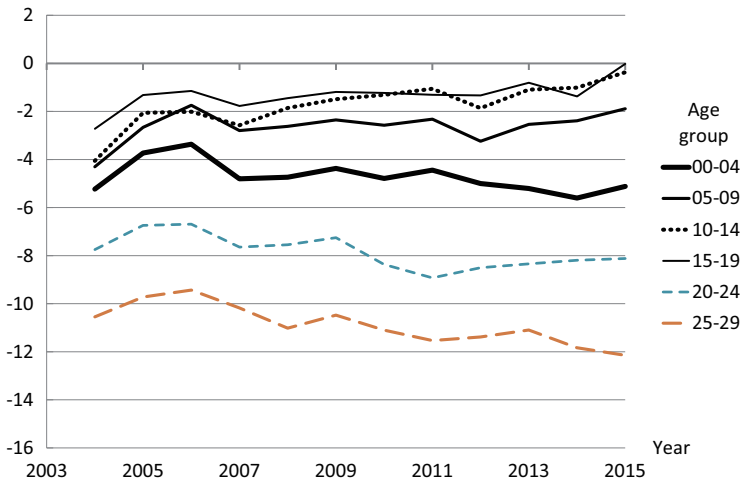

Fig. 2. Net apparent migration from birth to observation in the 2004–2015 annual census surveys, for persons born in Metropolitan France, among different age-groups at census, by census survey year (percent). Source: Annual census surveys 2004–2015: own estimates from downloaded files.

to avoid omissions, but the incentive to avoid double counts is weak: the municipalities, as well as the enumerators, benefit from a larger population being included in the survey. Double counts may concern children with separated parents, sharing their time between both parental homes, with both parents keen to include them as "permanent members" of their household; at ages 17 and 18, young adults may live away from the family home for their studies, and return during the week-end. In that case, they are supposed to be included in the census only at their family home up to age 18, and only at their week residence at higher ages, but the instructions are poorly read in practice, and this rule is incorrectly applied.

After improvements of the coverage in 2005 and 2006 at all ages, the undercount of young children is almost constant between 2007 (4.8%) and 2015 (5.1%), as can be seen in Fig. 2. It is also stable at ages 5–9 and decreases at ages 10–19. The undercount of young adults (aged 20–29) is probably increasing, but including age-specific estimates of net migration would be necessary to accurately evaluate this trend.

4. Three different estimates of population by sex and age

In addition to census data and annual survey data, INSEE is publishing population estimates by sex and age, based on the census but slightly different, especially at very young ages and around age 20. These three estimates of population by sex and age are very close, but not identical. Census data are 5-year moving averages of Annual Census Surveys estimates, with additional post-stratification on the number of dwellings coming from a specific dwelling register and from housing tax data and the mean number of persons per dwelling. This 5-year moving average dated "January 1st, year t" is using age at census for all surveys, thus merging 5 cohorts for each age: people "aged x in t" are actually aged x in $[t - 2; t + 2]$, and thus born in $[t - x - 3; t - x + 1]$. This makes a direct comparison between births and age-specific population size irrelevant, and INSEE has decided to produce demographic population estimates at the national level by sex and age based on "within cohort" moving averages: the population aged x in t is based on census data related to population included in the Annual Census Surveys in years $t - 2$ to $t + 2$ at ages $x - 2$ to $x + 2$ respectively, thus all belonging to cohort born in $t - x - 1$. This has been precisely documented [11], as the original aim of INSEE was to use one only set population figures by sex and age directly based on census data. INSEE publishes two series of population by sex and age, one directly based on census data (by age), one based on census data (rearranged by cohort). The two estimates are different for two reasons. On the one hand, census data were not adjusted by sex and age, because of internal migrations and the important issue of the distribution of the adjustment among the territory; on the other hand, it appeared necessary to use demographic population estimates by sex and age consistent with demographic data on births by age of parents and deaths by sex and age, in order to accurately estimate sex- and age-specific birth and death rates.

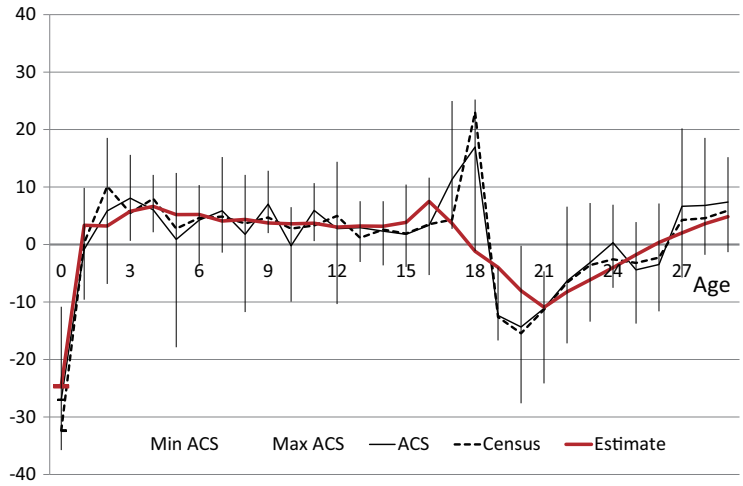

Fig. 3. Net apparent migration during one year, by age at the end of the year (mean 2006–2012), according to 2006–2013 Census data, Annual Census Survey data and demographic Population Estimates, by age (Thousands). Source: Annual census surveys and censuses 2006–2013: own estimates from downloaded files. Population estimates: [2] Table_fm_ T6.

In addition, some smoothing is performed in the population estimate in order to:

1. Smooth the negative net migration between birth and the end of the year, from one year to the next.
2. Smooth net migration estimates among cohorts at each age, from one year to the next.
3. Smooth the population age-structure, especially around ages 17–19.

The result of this smoothing procedure is presented in Fig. 3. First, the mean apparent net migration from one year to the next is highly volatile, according to the Annual Census Surveys: the standard deviation of net migration at each age is around 6,500 for ages 0–29 for a mean population of 767,000 inhabitants at each age, a coefficient of variation of more than 1%. This comes from large variations from one annual survey to the next, which order of magnitude is much larger than random variations due to sampling procedures (a Poisson estimate leads to the order of magnitude of 900 for the standard deviation, 0.1% for the coefficient of variation). Second, apparent net migration by age presents some dips and bumps, without any relation with actual population behavior. Census data published on Insee website also presents smaller irregularities at young ages, but the peak at age 18 is more pronounced (and no peak is visible at age 17), because ACS results were computed by cohort, while Census-based published data presented here are presented by "age at census" (five-cohort moving average). The demographic population estimates at January 1st, based on "within cohort" census data, are additionally smoothed to be consistent with other demographic estimates, such as migration and mortality by sex and age, and their trends from one year to the next. Demographic population estimates are then much steadier at all ages, and they also show specific adjustments at ages 0–3 and 15–20. This is indicated in a technical note, without much detail on the adjustment method for the years 2006–2013 [2]. An examination of annual net migration in years 2006 to 2013 (not shown) indicates that since 2012 the adjustment at ages 15–20 has changed in order to produce a more coherent age-structure without introducing a massive immigration at ages 17–18.

5. Why are young children undercounted at census?

In France as well as in the US, "Birth tourism" is likely to be present: women living abroad may come to France to give birth, because it may be safer and cheaper than in their country of residence. This phenomenon is difficult to identify, as no formal proof of residence is needed to register a birth: some births may be falsely declared as to a mother living in France, because the mother can be housed by relatives at the end of their delivery, or have a secondary residence in France. Out of 781,167 births in France in 2014, only 1,889 are registered from a mother living abroad or in an overseas department. More precise checks are made difficult by the fact that little information is common to civil registration and census. Moreover, migrations

may occur just after a birth. Undercount at census may come from omission of young children by their parents and the enumerator; they may also come from indirect collection bias. Partial non-response may occur if some individual forms are missing and, when this happens, it may be that young children are then the most likely to be omitted. The housing form includes only eight rows for the "permanent inhabitants", and very young inhabitants may be listed out of the proper lines in the dwelling form, and then omitted in the computerization. Total non-response also occurs in some households, and the rules used to produce "Non-surveyed dwelling forms" and household imputations may also lead to an undercount of young children.

Undercount may be present at all ages, and over-counts may also be numerous in the new census. These over-counts are likely to increase with age among children. A global assessment of undercount of young children is thus still needed for France. It would imply a careful examination of the different procedures taking place between enumeration and dissemination of data: forms collection, computerization, imputation of individuals and households, weighting procedures. This assessment would inform on undercount of young children, but also on likely over-count of older children.

6. Conclusion

The issue of the overall consistency of population estimates and trends is commonly raised after a census, and may lead to different adjustments: changes in the current or in the previous census population estimates, changes in the inter-census net migration figures, or the inclusion of a "statistical adjustment" in the overall population balance [8]. After 2006, such an adjustment of 666,000 inhabitants has been added to the population balance, in order to maintain the consistency of demographic estimates. Neither the figures of the 2006, nor those of the 1999 census, were changed, despite the fact that the 1999 census had been shown to undercount the population [6]. The detail of this adjustment by sex and age [10] shows a negative adjustment for years 2000–2006 for young children born after the 1999 census, who had been included at birth in the population, while on the contrary children aged 6–20, whose number came from the 1999 census figures, were more numerous at the 2006 census, and were then subject to a positive adjustment. Note that a negative adjustment was performed for young men in their 20s in 2006, likely to have been counted twice in 1999 (at ages 16–20) and to have been omitted in 2006 (at ages 23–27).

With the use of Annual Census Surveys, this issue of demographic consistency and adjustment has become permanent in France, as new estimates are published each year. The solution to produce census data based on 5-year moving averages has proven efficient and careful. An adjustment may be very difficult to be spread consistently among different sub-groups in the population, and Population estimates by sex and age at the national level differ from census data. Nevertheless, the time may have come to explicitly assess the statistical quality of Annual Census Surveys and census data, and to improve the consistency of French population estimates, or at least to better document the discrepancies between them. This is all the more important that census methodology is changing, and additional data are becoming available. More and more census forms are being collected through Internet, which may change the undercount of young children. Income and housing tax files include information on population by sex and age, and will likely be used, in one way or the other, as a complement to census data, not only to post-stratify the number of households at census, but also for their use as a panel [5]. Furthermore, the French demographic panel, a 4% "permanent demographic sample" merging data from civil registration (births, deaths, marriages), census data, tax data, allowance data, is now available for research on a secure remote access basis. This last dataset has been dramatically improved recently [12,13] and may be very useful to estimate the probability of double counts. Even if administrative records have their own shortcomings, a multisource approach may allow statistically assessing the quality of census data. International comparisons, as well as collaboration between statisticians from INSEE and researchers, may be useful to reach a shared diagnosis on this topic. The paper from William P. O'Hare is a perfect example of a welcome and useful path in that direction.

Acknowledgment

This work has been funded by the French National Research Agency (grant "ANR-16-CE41-0007-01").

References

[1] J. Arbel and V. Costemalle, Estimation des flux d'immigration: réconciliation de deux sources par une approche bayési-

enne [Estimating immigration flows: reconciling two sources using a Bayesian approach], Économie & Statistique, 2016, n° 483-484-485, 121–149. https://www.insee.fr/fr/statistiques/2017646?sommaire=2017660.

[2] C. Beaumel and V. Bellamy, La situation démographique en 2014 [Population trends in 2014]. Insee Résultats, N° 182 Société. https://www.insee.fr/fr/statistiques/2045470. See also C. Beaumel and V. Bellamy (2015a). Statistiques d'état civil sur les naissances en 2014. Insee Résultats, N° 171 Société. https://www.insee.fr/fr/statistiques/1406578. C. Beaumel and V. Bellamy (2015b). Statistiques d'état civil sur les décès en 2014. Insee Résultats, N° 172 Société. https://www.insee.fr/fr/statistiques/1406302, 2016.

[3] G. Brihault and N. Caron, Le passage à une collecte par sondage: quel impact sur la précision du recensement? [Transition to data collection by sampling: what impact on census accuracy?], Économie & Statistique, n° 483-484-485, 2016, 23–40. https://www.insee.fr/fr/statistiques/2017638?sommaire=2017660.

[4] C. Brutel, L'analyse des flux migratoires entre la France et l'étranger entre 2006 et 2013. Un accroissement des mobilités [Analysing migratory flows between France and the rest of the world in the period 2006–2013 – Increased mobility], Insee Analyses, n° 22, 2015, 1–4. https://www.insee.fr/fr/statistiques/1521331.

[5] M. Chaleix and S. Lollivier, Panels for social statistics. Courrier des statistiques, English series no.12, 2006, 49–52. C. Couet, INSEE's Permanent Demographic Sample (EDP). Courrier des statistiques, English series no.13, 2007, 29–38. https://www.epsilon.insee.fr/jspui/handle/1/14398. https://www.epsilon.insee.fr/jspui/handle/1/14399.

[6] G. Desplanques, Analyse des écarts entre les résultats du recensement de 1999 et les estimations fondées sur le recensement de 1990. Population métropolitaine: 480 000 personnes de moins que prévu [An analysis of differences between 1999 census results and estimates based on the 1990 census population in mainland France: 480,000 fewer people than expected]. Documents de travail de l'Insee, n°F0403. https://www.epsilon.insee.fr/jspui/handle/1/5738, 2004.

[7] G. Desplanques, Strengths and uncertainties of the French annual census surveys, *Population-E* **63**(3) (2008), 477–501. https://www.journal-population.com/numero_revue/2008-3-volume-63-numero-3.

[8] F. Héran and L. Toulemon, What happens when the census population figure does not match the estimates? Population and Societies, n° 411, April. https://www.ined.fr/en/publications/population-and-societies/what-happens-when-the-census-population-figure-does-not-match-the-estimates-en, 2005.

[9] Insee, Documentation on the French census: a. downloadable individual files at https://www.insee.fr/fr/statistiques?debut=0&theme=1&categorie=3&collection=4 and https://www.insee.fr/fr/statistiques/2409379?sommaire=2409559. b. Documentation https://www.insee.fr/en/information/2517226. c. Technical information on the [in French]: https://www.insee.fr/fr/metadonnees/source/s1321#documentation. d. Census forms: https://www.insee.fr/fr/metadonnees/source/s1321#documentation.

[10] INSEE, Révision des estimations de population nationale par sexe, âge et situation matrimoniale du 1er janvier 2000 au 1er janvier 2006 pour tenir compte des résultats du recensement 2006 [Revision of population estimates by sex, age and marital situation form 1-1-2000 to 1-1-2006, taking into account the results of the 2006 Population Census]. https://www.insee.fr/fr/metadonnees/source/s1169#documentation, 2010.

[11] S. Jugnot, M. Anguis and C. Beaumel, Construire une pyramide des âges pertinente pour le calcul des indicateurs démographiques à partir des enquêtes annuelles de recensement [Constructing an age pyramid relevant for the calculation of demographic indicators from annual census surveys]. Documents de travail de l'Insee, n°F2010/03. https://www.insee.fr/fr/statistiques/1380914, 2010.

[12] S. Jugnot, La constitution de l'échantillon démographique permanent de 1968 à 2012 [Developing the Permanent Demographic Sample from 1968 to 2012], INSEE, Working paper F1406. https://www.insee.fr/fr/statistiques/1381113. Presented in English in [13], 2014.

[13] M. Solignac, Analysis of "La constitution de l'échantillon démographique permanent de 1968 à 2012 [Developing the permanent demographic sample, 1968–2012]", Population, 70, n° 4, 2015, 817–820. https://www.journal-population.com/numero_revue/2015-4-volume-70-numero-4/.