United Nations ECE/CES/GE.41/2019/9

Economic and Social Council

Distr.: General
3 July 2019
English
Original: French

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Twenty-first Meeting
Geneva, 18–20 September 2019
Item 2 of the provisional agenda
Results of tests with regard to methodology, technology, participation and other aspects

Using first names to improve the measurement of same-sex couples in the census

Note by the National Institute of Statistics and Economic Studies (INSEE, France)*

Summary

It is difficult to produce reliable statistics on the number of same-sex couples in France on the basis of the French census. A significant number of same-sex couples – more than 40 per cent according to a 2011 study – are counted as such owing to a coding error for the sex of one of the members. That leads to an overestimate. The proposed correction procedure consists in calculating the proportion of people with a particular first name who are male or female and using that information to rectify the sex variable for people who, according to the census data, are living in a same-sex relationship. This method seems effective and gives results that are consistent with other sources.

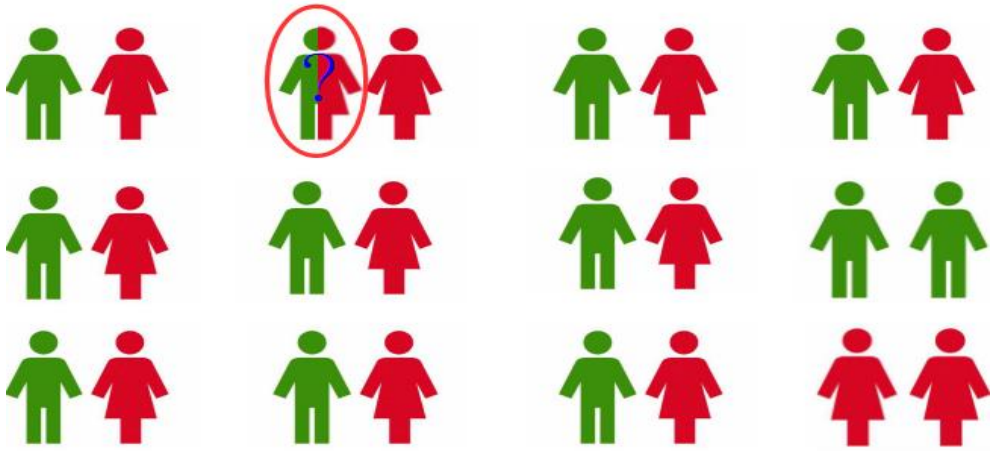
* Prepared by Elisabeth Algava and Sébastien Hallépée.

GE.19-10957 (E) 310719 310719
1910957

I. Introduction

1. It is currently impossible to produce reliable statistics on the number of cohabiting same-sex couples in France based on the census, owing to a methodological difficulty. A reporting or coding error with the sex of only one member of an opposite-sex couple leads to the couple being counted as same-sex. While this affects only a small proportion of opposite-sex couples, it is sufficient to produce a substantial overestimate of the proportion of the population in same-sex relationships. This risk is not specific to same-sex couples. It exists whenever it is necessary to estimate rare populations with only a small number of respondents. When measuring the number of widowed people under 30 years of age or the number of married 18-year-olds, errors in age or marital status must be taken into account: their frequency may exceed the number of early marriages or young widows/widowers. The difference in this case is that an error in the sex of only one member of the couple will lead to both members being considered as having a same-sex partner (Figure 1): each error counts twice.

Figure 1
Impact of a coding error on the sex variable



Note to the reader: This imaginary population contains 12 couples and 24 individuals. Nine are opposite-sex couples, two are same-sex couples and there is a doubt as to the sex coding of one of the individuals in the final couple.

If it is a coding error, an error proportion of 1 in 24 will switch 1 in 12 couples from opposite-sex to same-sex. The error therefore counts double. The number of same-sex couples would be artificially increased by 50 per cent while the number of opposite-sex couples would be artificially reduced by only 10 per cent. Again, low numbers of coding errors have a much more visible impact on small populations.

2. INSEE was able to overcome this difficulty in 2011, when it conducted a survey on families and housing. For a sample of the housing units included in the census, an additional four-page questionnaire was delivered at the same time as the census forms. Approximately 360,000 questionnaires were collected. The survey on families and housing enabled a large amount of research on numerous topics (Bodier *et al.*, 2015; Imbert *et al.*, 2018), including on the subject of same-sex couples. Major verification work was carried out on the responses, by cross-checking the information available in the census and the survey on families and housing (Breuil-Genier *et al.*, 2016). It was then possible to estimate the number of persons in same-sex relationships at 205,000, of whom 173,000 were cohabiting. The percentage of cohabiting couples who were same-sex couples was 0.57 per cent and the percentage of "false" same-sex couples was 0.36 per cent (Buisson and Lapinte, 2013; Banens and Le Penven, 2016). The extent of the correction was therefore quite considerable for same-sex couples: from 295,000 persons before the rectification to 173,000 after it. More than 40 per cent of cases were corrected. This correction step

enabled new analyses of persons in same-sex relationships in 2011 (Rault, 2016, 2017, 2018).

3. The next families survey, which in principle will make it possible to produce a new controlled estimation of the number of same-sex couples, is not planned before 2023. It is important to have data before that date, given the recent changes in legislation (particularly the law of May 2013 that opened marriage to same-sex couples) and the commitments made by France to provide data on same-sex couples for the 2021 European census.[1] Identifying them more reliably in the census will ensure that a high-quality response is provided to the European statistical institute. Taking advantage of the census and the related annual collection of information from millions of people and housing units (Godinot, 2016), this improvement will also make it possible to perform new analyses on this population and its demographic, family and socioeconomic characteristics. Admittedly, the analysis in the census is restricted to cohabiting couples, while conjugal relationships between persons of the same sex are more often "long-distance" than is the case for opposite-sex couples (Toulemon *et al*., 2005; Rault, 2018). However, the information remains very useful, as cohabiting relationships are by far the most widespread form of relationship even among same-sex couples (84 per cent in 2011, Buisson *et al*., 2013).

4. It is therefore justified to implement a solution to make it possible to distinguish between genuine same-sex couples and couples counted as such owing to a sex coding error. To achieve this, it is planned to add a new calculated individual variable to the census processing sequences, indicating what proportion of people with the reported first name are male or female. This variable would then be used to correct the sex variable for persons who the census data show as living in a same-sex relationship. Implementation of the proposed procedure within the processing sequence is planned for the 2020 annual census survey at the earliest. However, an experimental implementation, in addition to the standard processing operations, is planned for the annual census surveys from 2017 to 2019.

5. After introducing the experience of and solutions implemented in other countries and the French census, we will set out the proposed procedure and an initial application.

II. Solutions implemented in other countries

6. Difficulty in measuring same-sex couples is not specific to France and different solutions have been put in place in other countries. Banens and Penven (2016) present estimations of the proportion of all couples in the United States, Canadian and British censuses who are "false" same-sex couples. It ranges from 0.25 to 0.57 per cent. The share of "false" couples in the total number of apparently same-sex couples is between 27 and 55 per cent. These figures are within the same order of magnitude as those measured for France, with the same difficulties: few errors overall but with a significant negative impact on ability to estimate the number of same-sex couples. The countries faced with this difficulty have experimented with different strategies to overcome it.

A. Information redundancy and cross-checking

7. The first group of solutions involve modifying the questionnaire or protocol of the survey or census to obtain additional validation criteria. The general principle is to rely on the fact that errors are rare and the probability of there being two cumulative errors is very low.

8. In the Canadian census since 2001 and similarly in the United States census from 2020, of all the items that can be selected to describe the relationship between two people, four describe a conjugal relationship: Opposite-sex husband or wife, Opposite-sex common-law partner, Same-sex married spouse, Same-sex common-law partner. It is then possible to cross-check this information with the sex of the partners. If two partners of the same sex select the "opposite-sex husband or wife" option, it is probable that the sex of one

[1] European Commission Implementing Regulation No. 1201/2009.

of the two partners has been miscoded; conversely, if they both select the "same-sex married spouse" option, the probability of two errors is very low and this is an "actual" same-sex couple. For online respondents to the 2020 United States census, it is also planned to include a check (window) when the item selected is inconsistent with the reported sexes (for example when a woman reports that she is the opposite-sex spouse of another woman). Advance tests have shown that inconsistencies are significantly reduced with the new question and the automatic checks used in online completion (Kreider, 2017).

9. This approach of cross-checking information collected in different ways is very similar to the one used for the 2011 families and housing survey.

B. Matching with administrative data

10. An experiment to confirm reported and coded sex by matching with administrative data was implemented in the United States of America (Kreider, 2015). By matching census data with social security records, the authors found that inconsistencies between the sex coding in the census and the sex given in the social security records (Numident) for at least one partner were much more frequent when the couples appeared to be same-sex in the census: 72.7 per cent for married couples and 6.4 per cent for unmarried couples.

C. "Statistical validation" using first names

11. When analysing the 2010 census, the United States Census Bureau used a name probability index (O'Connell, 2011). The index was constructed from the responses to the census itself and indicated the proportion of people with the name who were male ("maleness"), between 0 and 1,000. A threshold for correction of 50 out of 1,000, or 5 per cent, was selected by the authors, a threshold they considered "conservative". In other words, if the index indicated that the name of a respondent coded as male was a male name in only 5 per cent of cases or fewer, the sex was corrected. Otherwise, it was maintained. This threshold resulted in the correction of inconsistencies between first name and sex for at least one of the members of 50 per cent of the couples identified as same-sex, more frequently if they were married (69 per cent) than unmarried (21 per cent).

12. The results obtained were subsequently compared with social security records to check whether the corrections made actually corresponded to sex coding errors (Kreider, 2015). In this test, 85 per cent of respondents had a first name considered unambiguous and so the correction could be applied. In 96 per cent of cases, the sex assigned based on first name was the same as the sex in the social security records.

13. The method of statistical validation using first names therefore seems sufficiently reliable, at least as applied in the 2010 United States census. These results have encouraged tests of the possible use of this method in the French census.

III. Recent developments in the French census

14. Since 2004, the French census has been sample-based. Data collection is annual and consists in completion of a paper or online questionnaire. To publish the commune-by-commune results for legal population, the data from five annual collection surveys are used. For example, "results of the 2014 population census" refers to data from the annual collection surveys of 2012 to 2016. By contrast, the "2017 annual census survey" refers to the annual collection of 2017, meaning all the persons and housing units surveyed in that year. It is these individual annual surveys that will be used below.

15. For each housing unit surveyed, the census taker must collect one form for the housing unit, which describes the relationships between the residents of the housing unit, and one individual form for each resident, describing their sociodemographic characteristics. The individual form was redesigned in 2015: the question on legal marital status (married, divorced, widowed, single) was replaced with a question on de facto status (married, in civil union, living with unmarried partner, divorced, widowed, single). This has improved

the quality of responses because people identify more with the options, particularly those in a civil union, who sometimes used to report that they were married as they considered that status closer to their actual way of life (Buisson, 2017).

16. The housing unit form was redesigned for the 2018 collection survey. It will more effectively capture a conjugal relationship between two persons based on what they have reported, rather than deducing it from the fact that each one reported living as a couple (without specifying with whom). This redesign has indirectly created the conditions for improving the measurement of same-sex couples. Its implementation requires systematic matching between the housing unit form – pertaining to the individuals included on the list of residents and the relationships between each pair of residents – and the individual form for each resident. They are matched using the criteria of sex and birth year and, if the first two variables are insufficient, surname and first name. The set of names is therefore used during processing (and then deleted from the files to be disseminated); the new input method was introduced early, beginning with the 2016 collection survey, enabling the procedure proposed below to be used.

17. One of the main recent developments in the census is online collection. Almost non-existent in 2013, in 2017 it accounted for more than half the individuals surveyed. There are thus no longer any response coding errors related to optical character recognition or handwritten corrections to paper forms in respect of those persons. For our purposes, it is for first names that the difference in quality between online and paper collection is most significant. This is particularly the case since, in order to keep costs down to a level commensurate with the use made of first names, the quality criteria used for the input of first names collected from paper forms are rather low. This difference has proved to be critical, requiring an adaptation to be made to the proposed processing method.

18. Aside from the methodological aspect, online collection may appear to offer greater guarantees of confidentiality and improve the veracity of reporting, particularly by same-sex couples. The universal availability of electronic data gathering for the 2020 United States Census is considered by the Census Bureau as one of the ways to improve measurement of same-sex couples (Kreider, 2017).

IV. Finding an optimal solution

19. The example of the United States suggests that it is fairly effective to use first names to identify sex coding errors and thus identify the false same-sex couples among all apparent same-sex couples. Taking advantage of the fact that, since 2016, first names have been entered to conduct quality control on the census data collection, the effectiveness of this type of “statistical validation” using first names was tested.

A. Interim indicator: apparent same-sex couples

20. To estimate the numbers of apparent same-sex couples over time, and to test the proportion to be considered genuine same-sex couples and the proportion to be categorized as sex coding errors, a simplified indicator is used: if exactly two persons report living as a couple in a housing unit and are of the same sex, they are counted as an apparent same-sex couple. That indicator is imperfect since two persons who live together and both report that they “live as a couple” because each one has a partner living in another housing unit will be wrongly counted as forming a couple with each other. However, a comparison of this simplified indicator with the measurement validated using the data of the 2011 family and housing survey shows that the indicator is satisfactory for estimating apparent same-sex couples. As expected, it encompasses too many couples, many of whom are opposite-sex couples counted as same-sex couples owing to a coding error. However, few “actual” same-sex couples are omitted. The redesigned housing unit form will make it possible to obtain a better measurement for 2018.

B. The Permanent Demographic Sample: an ideal tool for testing the capability of the procedure to identify proven sex coding errors

21. The Permanent Demographic Sample is a large-scale sociodemographic panel established in France to study demographic, family, professional and geographic trajectories (Durier, 2018). The general principle is to maintain information on individuals in the sample (approximately 4 per cent of the population) collected from the five statistical sources used in the Permanent Demographic Sample (civil registration forms, censuses, electoral roll, all employee panel for wages and, since 2011, fiscal and social data on income).

22. The primary use of the Permanent Demographic Sample is to compare "actual sex", which is the sex recorded in the national directory for the identification of natural persons, considered accurate because it is used to generate social security numbers, with the sex reported in the annual census surveys. The proportion of sex coding errors in the census can then be estimated. In the 2016 research database, for the 1.3 million Permanent Demographic Sample members (born on a Permanent Demographic Sample day) who participated in at least one census survey between 2010 and 2016 and who live as part of a couple, the sex error rate is 0.17 per cent. It is therefore a very rare phenomenon. However, the error rate is considerably higher (18 per cent) for apparent same-sex couples.

23. In the Permanent Demographic Sample, the national directory for the identification of natural persons is only queried for the Permanent Demographic Sample member to include that person in the sample and supplement the statistical data about him or her. The statistical information about other residents of the same housing unit is also included, but without identification or a query to the national directory for the identification of natural persons. It is therefore not possible to confirm the sex of the other residents in the housing unit, including partners. Since 0.17 per cent of Permanent Demographic Sample members are affected by a coding error, it can be estimated, assuming that errors affecting partners are independent of each other, that 0.31 per cent of couples would be affected by a sex coding error for one partner, leading them to be incorrectly counted as a same-sex couple, and 0.03 per cent by two coding errors (Table 1).

Table 1
Predicted consequence of sex coding errors on couples

Actual situation		*Apparent situation*		
H1: 0.6 per cent of couples are "actual" same-sex couples		*H2: Sex is miscoded for 0.17 per cent of census participants*		
		Couples with one error (0.31 per cent)	*Couples with two errors (0.03 per cent)*	*Couples with no error (99.66 per cent)*
Of 1 million couples	6,000 same-sex couples	➔ 19 apparent opposite-sex couples	2 same-sex couples	5,980 same-sex couples
	994,000 opposite-sex couples	3,078 apparent same-sex couples	287 opposite-sex couples	990,634 opposite-sex couples

Source: 2011 family and housing survey, 2016 Permanent Demographic Sample research database, INSEE.

Scope: Cohabiting couples.

Interpretation: According to the family and housing survey, 0.6 per cent of couples are "actual" same-sex couples and 99.4 per cent are "actual" opposite-sex couples (H1). With a coding error affecting 0.17 per cent of census participants (H2), 0.03 per cent of couples are affected by two coding errors (0.17 x 0.17) while 0.31 per cent of couples are affected by an error for one of the members (0.17 + 0.17 – 0.03). Thus, 0.31 per cent of opposite-sex couples become apparent same-sex couples, i.e. 3,078 in 1 million couples. Conversely, 0.31 per cent of same-sex couples, i.e. 19, become apparent opposite-sex couples. Finally, 0.03 per cent of couples have both members affected by a sex coding error (2 same-sex couples and 287 opposite-sex couples). Even if neither member of

the couple is shown with the correct sex, the apparent situation of the couple reflects the reality: the couple remains an opposite-sex couple. With an error rate of 0.17 per cent, it can be expected that the actual situation, in which same-sex couples account for 0.6 per cent of couples, changes to an apparent situation in which they represent 0.9 per cent of couples in the annual census survey (5,980 + 3,078 + 2 = 9,060 in 1 million).

C. Selected dictionary

24. The second use of the Permanent Demographic Sample is that the first name reported in the census is included in the production database to facilitate matching (however, it is not included in the research database used to produce statistics, to avoid the direct identification of individuals). It is therefore possible to construct a dictionary and error indicator in the same way as in the production sequence for future annual census surveys.

25. For each first name, a name dictionary includes the proportion of persons with that name who are female (or male). It is then matched with the first names of census participants to compare the sex coded for each participant with the most common sex of persons with that name. The purpose is to identify the most probable instances of coding errors. The Permanent Demographic Sample was used to compare the performance of different dictionaries and select the most effective dictionary to identify sex coding errors affecting sample members.

26. A combination of the tested dictionaries was selected, using two sources. The preferred source, exhaustive for persons born in France, is the file of first names recorded in the civil registry since 1900, by sex.[2] A very large number of observations is required to calculate the proportions of men and women among the persons with each first name. The file was completed by adding occurrences of the first names of 2017 census participants born in other countries, since they are not covered by the civil registry.[3] They are more likely to have a first name which is not in the dictionary constructed from the civil registry. The selected dictionary is also a combination in the sense that a match is first searched for in the most detailed dictionary possible (same first name, same birth year). If none is found, a less detailed dictionary is used: first part of the name identical, no condition as to birth year. The proportion finally assigned to a person can thus be either the proportion of women among all persons born in the same year with the same name or the proportion of women among all persons with a first name that begins in the same way. For the sake of simplicity, this indicator will be referred to as the proportion of women among persons with the same name. This proportion is then used to deduce an error indicator: if the sex coded in the census is male, this is the proportion of women with that name according to the dictionary. If the sex coded is female, it is the proportion of men. By convention, if there is no match and so no proportion, the indicator is set at 0: no correction can be made. The higher the value is, the greater the suspicion of a coding error.

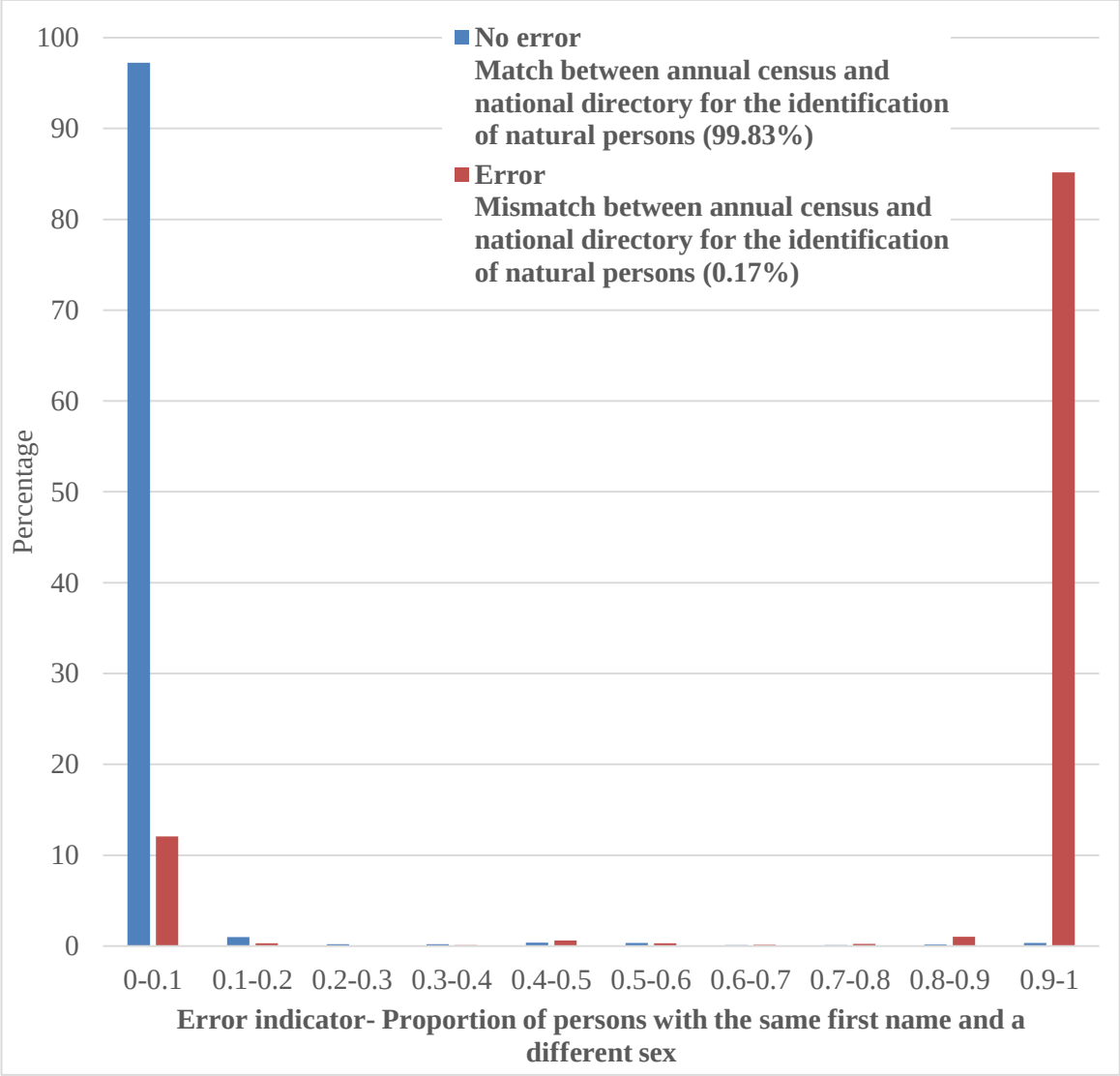
D. For Permanent Demographic Sample members, sex coding error identification using first names is highly effective

27. This indicator is highly effective in identifying coding errors. When there is no proven error, almost all the values are 0 or very close to it (Graph 1). Conversely, in cases of proven error (the sex coded in the annual census survey is different from the sex recorded in the national directory for the identification of natural persons), the values are almost all close to 1. Intermediate values are rare: few people have unisex first names like Dominique.

[2] First names given too rarely are removed.

[3] These first names, available for online respondents, were entered for paper questionnaire respondents for census management purposes. The names were then destroyed as planned once the processing had been completed in 2017.

Graph 1
Error indicator distribution according to the presence of a proven error



Source: 2016 Permanent Demographic Sample research database, INSEE, weighted data, all persons living in a couple

Interpretation: Census participants between 2010 and 2016 of over 15 years of age who reported living in a couple for the census and whose reported sex from the census has been confirmed in the national directory for the identification of natural persons almost all have a first name consistent with the sex reported in the census: the values for the error indicator constructed with first names are concentrated around 0. By contrast, for a large majority of persons whose reported sex from the census was contradicted by the national directory for the identification of natural persons, the indicator constructed based on first name indicates the high probability of an error in the sex reported in the census, with values concentrated around 1. Values around 0 mainly correspond to individuals who could not be matched with the name dictionary (for example, the name was too rare).

28. To determine the optimal threshold above which a correction will be made, the tools are the same as those used to develop a diagnostic test in epidemiology: specificity measures (errors correctly identified) and sensitivity measures (accurate responses not wrongly corrected), optimal threshold determination and test comparison using the receiver operating characteristic (ROC) curve (Robin, 2011). This resulted in the selection of a threshold of 41 per cent: for a person coded as male, if more than 41 per cent of persons with the same name are female, a correction is made. It may seem surprising that a correction is made even when the majority of persons with a name are male, but in reality there are very few cases in which the dictionary value is intermediate: almost all values are

very close to either 0 or 100 per cent. Threshold determination has a limited impact in such a favourable context. The selected dictionary and the threshold of 41 per cent enable sensitivity of 98 per cent and specificity of 95 per cent. Considering only apparent same-sex couples, approximately 150 undue corrections would be made in the Permanent Demographic Sample for almost 2,000 actual identified errors.

Table 2
Dictionary performance in identifying sex coding errors

		Persons in an apparent same-sex couple		
	All persons in a couple	*Total*	*Paper collection*	*Online collection*
Number	1 343 908	11 349	9 605	1 744
Proven errors	2 336	2 059	1 852	207
Optimal threshold (%)	41	41	41	51
Specificity (%)	99	98	98	99
Sensitivity (%)	89	95	95	95
Undue corrections	16 950	147	134	9
Valid corrections	2 088	1 951	1 754	197

Source: 2016 Permanent Demographic Sample research database, INSEE, unweighted data.

Scope: Persons over the age of 15 years living in a couple who participated in the census between 2010 and 2016.

V. The unexpected results of transferring the procedure from the Permanent Demographic Sample to the annual census survey and the need to adapt the procedure according to the collection mode

A. Direct transfer into the census leads to excessive corrections

29. Using the Permanent Demographic Sample allowed the correction procedure to be evaluated on a sample in which the coding errors were known. As the Permanent Demographic Sample data are taken from the census, this would seem to be the ideal way to predict the performance of the correction once applied to the entire census. Unfortunately, a direct transfer produced unexpected results: 12 per cent of persons who responded using the paper questionnaire in the 2017 annual census survey (irrespective of relationship status) had a first name almost always attributed to the other sex (Table 3), which is far too high a proportion compared to the 0.2 per cent of sex coding errors it was intended to identify. It emerged that more errors related to input of the first name than to sex: for cost reasons, the level of quality required from the service provider when entering the first names of persons who responded on paper is low, except for the forms to be included in the Permanent Demographic Sample.

Table 3
Comparison of the dictionary applied to the Permanent Demographic Sample and the 2017 annual census survey by collection mode

	All couples					*Same-sex couples*		
		2017 annual census survey – before consideration of collection mode		*2017 annual census survey – after consideration of collection mode*			*2017 annual census survey – after consideration of collection mode*	
Persons with this first name ...	*Permanent Demographic Sample*	*Paper*	*Online*	*Paper*	*Online*	*Permanent Demographic Sample*	*Paper*	*Online*
... are almost always of the same sex as the respondent (EI < 5 per cent)	93	72	95	41	97	79	32	84
... are mostly of the same sex as the respondent (EI between 5 and 40 per cent)	3	8	3	36	2	2	32	1
... include members of both sexes (EI between 40 and 90 per cent)	1	7	1	21	1	1	26	1
... are very rarely of the same sex as the respondent (EI between 90 and 95 per cent)	0	1	0	1	0	0	3	0
... are almost never of the same sex as the respondent (EI > 95 per cent)	0	12	2	1	0	18	8	13
Total	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

EI: Error indicator.

Interpretation: 32 per cent of persons in an apparent same-sex relationship who responded on paper have (according to the input of their individual census form) a first name shared almost always by persons of the same sex as those respondents: the dictionary indicates a proportion of less than 5 per cent of persons of the opposite sex for that name.

Scope: Persons in an apparent same-sex relationship, for the 2017 annual census survey, excluding non-respondent housing unit forms and households in which the sex of one of the partners has been imputed.

Sources: 2016 Permanent Demographic Sample research database, 2017 annual census survey, INSEE.

B. Differentiated dictionaries based on collection mode

30. The solution selected to overcome this difficulty was to incorporate the collection mode in dictionary construction. When this is done, the proportion of situations in which more than 95 per cent of persons with a given name are of a different sex to that reported by the respondent becomes very low even for paper collection. The drawback is the high proportion of intermediate variables, which requires specific handling in the processing operations. Since the quality, even for paper collection, is better in the Permanent Demographic Sample, it is not possible to use the sample to test this consideration of the lower quality of paper census collection.

31. However, the results for online collection are much closer to the Permanent Demographic Sample. The quality of sex coding error identification in the Permanent Demographic Sample can thus be considered representative of the quality of this processing operation on the online portion of the annual census survey. It is therefore proposed to use the online-appropriate dictionary and set the threshold at 51 per cent, the threshold which, for online collection with the Permanent Demographic Sample, maximizes sensitivity (95 per cent, meaning that 5 per cent of error-free cases are wrongly corrected) and specificity (99 per cent, meaning that 99 per cent of proven errors are identified). For a person in an apparent same-sex relationship, a correction will be made to the reported sex whenever the proportion of persons with the same name who are of the opposite sex is greater than or equal to 51 per cent.

C. For paper collection, the correspondence between sex and first name is less reliable

32. For paper collection in the annual census survey, the dictionary was constructed so as to avoid undue corrections. However, the error indicator frequently has an intermediate value: the first name criterion is less effective and, unlike with online collection, the result is extremely sensitive to the threshold selected. For example, with a threshold of 95 per cent, the sex of 8 per cent of individuals in apparent same-sex relationships who responded on paper would be corrected, compared to 13 per cent of online respondents. With a threshold of 40 per cent, the correction rates would change to 37 per cent of paper respondents and 14 per cent, almost the same level, for online respondents. For paper collection, intermediate values indicate that the name is likely to have been input incorrectly (some names are more susceptible to this than others). Other variables are then used to estimate whether or not the person is likely to be in an "actual" same-sex relationship. These variables are selected based on the Permanent Demographic Sample, which can be used to measure the proportion of actual sex coding errors for persons in apparent same-sex relationships responding to the census on paper.

33. Finally, if the proportion of persons with the same name who are of the opposite sex to the respondent is more than 90 per cent, the sex is corrected; if it is less than 20 per cent, the sex is not corrected. If it is between 20 per cent and 90 per cent, the first name is considered ambiguous and additional variables are used. For the 2017 annual census survey, three groups were selected; unmarried persons, for whom the error rates are lowest (10 per cent); married persons below the age of 50 years (23 per cent); married persons above the age of 50 years (32 per cent). Starting with the 2018 annual census survey, taking advantage of the redesigned housing unit form, the groups have been formed with consideration of the type of family (childless couple, family with children, reconstituted or not) and the processing has been improved by looking at the situation of the couple (and not the two partners separately).

VI. Application: how many same-sex couples are there in France and what is the trend?

34. It has been possible to estimate the numbers of apparent same-sex couples in annual census surveys since 2005 and in the Permanent Demographic Sample since 2010. They are steadily increasing. The break in time series in 2018 can be explained by a major decrease in sex coding errors following the redesign of the housing unit form.

35. For "actual" same-sex couples, the various estimations (procedure using first names for the 2017 and 2018 annual census surveys, extrapolation of individual errors to couples for the years 2010 to 2018 in the Permanent Demographic Sample) appear to be consistent with each other and with the family and housing survey data. The proportion of all persons cohabiting with a partner who are in a same-sex relationship seems to have grown steadily since 2010, increasing from 0.5 per cent to 0.8 per cent. With better characterization of the family situation, the proportions estimated in the 2011 family and housing survey and the 2018 annual census survey are slightly higher but indicate growth, from 0.6 per cent in 2011 to 0.9 per cent in 2018.

Table 4
Proportion of apparent same-sex couples and estimate of "actual" same-sex couples in different sources from 2005 to 2018

	Apparent same-sex couples			*Estimate of "actual" same-sex couples*		
Percentage	*Annual census survey*	*Permanent Demographic Sample*	*Family and housing survey*	*Annual census survey*	*Permanent Demographic Sample*	*Family and housing survey*
2005	0.6					
2006	0.7					
2007	0.7					
2008	0.7					
2009	0.8					
2010	0.8	0.8			0.5	
2011	0.8	0.8	0.9		0.5	0.6
2012	0.9	0.9			0.6	
2013	0.9	0.9			0.6	
2014	1.0	1.0			0.6	
2015	1.0	1.0			0.7	
2016	1.1	1.1			0.7	
2017	1.2	1.2		0.8	0.8	
2018	1.0/1.0*	1.0		0.8/0.9*	0.8	

Source: Annual census survey 2005 to 2018, 2016 Permanent Demographic Sample research database and additional data for 2017 and 2018, INSEE.

Scope: Persons living in a couple. For the annual census surveys, excludes non-respondent housing unit forms and missing responses for the sex of one partner.

* Based on new information following the redesign of the *housing unit* form: the scope is persons cohabiting with a partner according to the new family household analysis. Non-respondent housing unit sheets have been reincorporated, as have missing responses for sex (now non-existent on paper and very rare online).

VII. Conclusion

36. Based on this analysis and despite certain limitations, the application of the method using first names to the French case produces satisfactory results. The first study to be completed using this method, which is ongoing, on the characteristics of persons living in a couple with a person of the same sex in 2018 will be used to fine-tune the diagnostic process. The objective is then to incorporate it into the census processing sequences so that the quality of the sex variables included in the files for dissemination can be improved.

References

Algava E., and Hallepee S. (2018). Estimer les effectifs de couples de personnes de même sexe au recensement: expérimentation d'une solution de validation du sexe par le prénom. Document de travail F1807, INSEE.

Banens M., and Le Penven E. (2016). Les erreurs de sexe dans le recensement et leurs effets sur l'estimation des couples de même sexe. *Population* 71(1), pp. 135–148.

Bodier M. *et al*., coord. (2015). *Couples et familles*. Paris: Insee Références, 190 pp.

Breuil-Genier P., Buisson G., Robert-Bobée I., and Trabut L. (2016). 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*, No. 483–485, pp. 205–226.

Buisson G., and Lapinte A. (2013). Le couple dans tous ses états. Non-cohabitation, conjoints de même sexe, Pacs ... Insee Première No. 1432, INSEE.

Buisson G. (2017). La situation matrimoniale dans le recensement: impact de la refonte du questionnaire de 2015. Document de travail F1707, INSEE.

Durier S. (2018). L'échantillon démographique permanent a 50 ans: retours sur un dispositif statistique original. Paper presented at the Journées de méthodologie statistique. Paris, June.

Godinot A., and Durr J.-M. (2016). La rénovation du *Recensement de la population*. In *Économie et statistique*, No. 483–485, pp. 7–14.

Imbert C., Lelievre E., and Lessault D., dir. (2018). *La famille à distance*: *mobilités, territoires et liens familiaux*. Ined, Questions de population No. 2, 376 pp.

Kreider R., Bates N., and Yeris M.-G. (2017). Improving Measurement of Same-Sex Couple Households in Census Bureau Surveys: Results from Recent Tests. SEHSD Working Paper, 2017–28.

Kreider R., and Lofquist. D (2014). Matching Survey Data with Administrative Records to Evaluate Reports of Same-Sex Married Couple Households. SEHSD Working Paper, 2014–36.

Lathe H., Ménard F.-P., Martel L., and Hallman S. (2017). Same-sex couples in Canada in 2016. Census in Brief, Statistics Canada, No. 98-200-X2016007.

O'Connell M. and Feliz S. (2011). Same-Sex Couple Household Statistics from the 2010 Census. SEHSD Working Paper, 2011–26.

Rault W. (2016). Les mobilités sociales et géographiques des gays et des lesbiennes. Une approche à partir des femmes et des hommes en couple. *Sociologie*, 7(4), pp. 337–360.

Rault W. (2018). La distance, une composante plus fréquente des relations conjugales et familiales des gays et des lesbiennes? In *La famille à distance: mobilités, territoires et liens familiaux*, Imbert C., Lelievre E., and Lessault D., dir., Ined, Questions de population No. 2, 376 pp.

Rault W. (2017). Secteurs d'activités et professions des gays et des lesbiennes en couple: des positions moins genrées. *Population*, 2017/3 (Vol. 72), pp. 399–434.

Toulemon L., Vitrac J., and Cassan F. (2005). Le difficile comptage des couples homosexuels d'après l'enquête EHF. In *Histoires de familles, histoires familiales. Les résultats de l'enquête Famille de 1999*, Lefevre C., Fillon A., dir., Ined, Cahier No. 156, pp. 589–602.