United Nations ECE/CES/GE.41/2017/12

Economic and Social Council

Distr.: General
17 July 2017

Original: English
English and Russian only

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Nineteenth Meeting
Geneva, 4–6 October 2017
Item 4 of the provisional agenda
Evaluating the census and measuring data quality

A procedure for assessing quality of variables in a register-based census

Note by the National Statistics Institute of Spain[1]

Summary

Spain is going to conduct for the first time a register based Census which will improve the quality of the operation.

The amount of administrative data available, type of information included on them and accessibility to the registers guarantee all together the quality of the Spanish Census. Besides, with the aim of being more transparent and provide valuable information for users we consider to disseminate for each particular variable another categorical variable that measures the quality of the data source that provided the information.

The final aim of this new approach is to make a step forward in some elements associated with the quality of the new Census regarding the last Census 2011 round.

[1] Prepared by Marina Pérez, Jorge Vega and Antonio Argüeso

GE.17-11987(E)
1711987

I. Introduction to the new census round

1. Since the traditional Census in 1991, in the latest editions Spain has carried out the Census using different methodologies according to the needs and circumstances. A combined one (traditional and registers) in 2001 and another combined one in 2011 based on a survey and registers.

2. 2021 Census will not be an exception and continuing with the step taken forward by other countries, Spain will join the list of countries with a register-based Census in the next round.

3. Today, the Spanish situation enables this change due to different reasons.

4. Firstly, Spain has a high quality Population Register (PR) called 'Padrón' since 1996 which is key in order to carry out a register based Census as the PR is the backbone of any register-based census.

5. Unlike other countries where the Police or other administrative bodies are in charge of population registers, in the case of Spain, the National Statistical Office (INE) is the national institution that coordinates this single national population register.

6. Every person residing in Spain must be registered in the municipality where he/she usually2 lives. The creation, maintenance, revision and safekeeping of every population register is carried out by each municipality and INE is responsible for the coordination and integration of all these municipal PRs into a single national population database.

7. The fact that INE has direct access to the PR makes the Spanish situation very different from other countries. It makes things easier to produce other statistical operations such as the Census.

8. This system has been working since 1996 and continuous improvements have been made during the last years.

9. Secondly, thanks to the Law on the Public Statistical Services (BOE 1989), INE is able to compile information from the statistical services of Ministerial departments and other bodies of the Administration. It provides Spain with the maximum possible guarantees for a successful operation in 2021.

10. There are at our disposal a large collection of administrative records that enable a qualitative step forward in quality compared to the traditional way of collecting information through household questionnaires. Many of the problems in previous Census editions will be overcome.

11. Currently we have gathered information from many public organisms such as the Tax Office, the Social Security, the Ministries and the Cadastre that produce an increasing amount of data extremely useful for the Population and Housing Census.

A. How will the 2021 Census be built?

12. Spain will build the Census file around the PR. The PR contains only a few variables for each person: name and surname, identification number (ID card number or passport), sex, date and place of birth, citizenship, current address, and educational attainment.

[2] If a person lives in several municipalities, he/she must register in that one where they live most days during a year.

Contrary to other countries[3] in Spain we have direct access to personal information which allows to improve data-linkage percentages among data sources.

13. The idea is to integrate the data coming from different sources to add variables to each register in the PR to complete the information.

14. For each Census variable not included in the PR (for example, legal marital status) different sources of information may be considered and different methodological approaches must be adopted. Some of the variables are easily estimated from records, but in other cases the challenge is much greater.

15. The population census final product can be considered as a matrix of (approx.) 47 million rows (people) with a few columns with the PR information and many other additional columns for the rest of variables based on the integrated information from sources.

		Information already in PR			*Information to be added by record linkage with data sources*					
	Identification	*Sex*	*Age*	*Nationality*						
1		…	…	…	…	…	…	…	…	…
2		…	…	…	…	…	…	…	…	…
3		…	…	…	…	…	…	…	…	…
4		…	…	…	…	…	…	…	…	…
.		…	…	…	…	…	…	…	…	…
.		…	…	…	…	…	…	…	…	…
.		…	…	…	…	…	…	…	…	…
.		…	…	…	…	…	…	…	…	…
.		…	…	…	…	…	…	…	…	…
.		…	…	…	…	…	…	…	…	…
.		…	…	…	…	…	…	…	…	…
47M		…	…	…	…	…	…	…	…	…

16. A similar approach will be made for the housing census based on a directory of dwellings and adding existing information in administrative sources.

17. With this new methodology we expect to produce a high quality census with all the pros related to this change.

[3] The most common situation in other NSIs is having access only to a PIN number but not to personal information such as name and surname.

18. The advantages of carrying out a register based census are well known by statisticians and affect all of the key points of the ESS guidelines for the implementation of the ESS Quality[4].

19. The production time will be shortened, and in line with the intentions of Eurostat beyond 2021 for the post-Census (to have census data continuously available, not having to wait a 10 years period of time). It avoids the bias due to forms, provides small area data and achieves a lower response burden. It provides information for the whole population in the majority of census variables enabling policy makers and all users to use it for their purposes. And last but not least, this new approach reduces drastically the costs (the Census in 2001 cost nearly €200m and the one in 2011 almost €85m). The financial savings will be maintained in the long term too.

20. In order to help users to better understand how the census is built and evaluate its quality we will provide extra information of the sources used to each estimation.

21. Section two explains briefly the procedure of assigning a type-of-source value to every census cell that will help us to measure quality afterwards. In section three two examples of analysis by columns are presented to better understand the possibilities of creating variables with extra information with the data source. In the following section, the analysis by rows is explained. In section five, we specify the information that will be made available for users and finally in the sixth section, conclusions are drawn.

II. Assigning a 'type-of-source' value to every census cell

22. There are two main reasons why we believe this new focus will improve quality. The first one is that users will be able to compare with other existing data and the second one is that we have created as a novelty a mechanism that enables users to evaluate the quality of each of the census variables and that will also enable INE to make better decisions.

23. For the matrix of the population census containing several billion cells, we will include for each census variables another one with metadata related with the source or method used to obtain each cell value.

24. The idea is that for each variable value (for example, legal marital status, maximum educational level attained, etc.) there would always be another one indicating the method or type of source used to provide the value for every person.

25. It must be underlined that for each Census variable we have a specific methodology in order to obtain each value that depends on the existing information and the peculiarities of the phenomenon implied. For example, if for a specific variable we find a large number of missing values in official registers, we may impute this data with deterministic or probabilistic models depending on the case.

[4] http://ec.europa.eu/eurostat/documents/64157/4373903/02-ESS-Quality-and-performance-Indicators-2014.pdf/5c996003-b770-4a7c-9c2f-bf733e6b1f31

	sex	*SEX type of source*	*age*	*AGE type of source*	*lms*	*LMS type of source*	*cas*	*CAS type of source*	*occ*	*OCC type of source*	*sie*	*SIE type of source*	*ind*	*IND type of source*	*edu*	*EDU type of source*
1	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
2	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
3	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
4	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
.	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
.	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
.	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
.	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
.	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
.	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
.	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
47M	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

A. How it will be made?

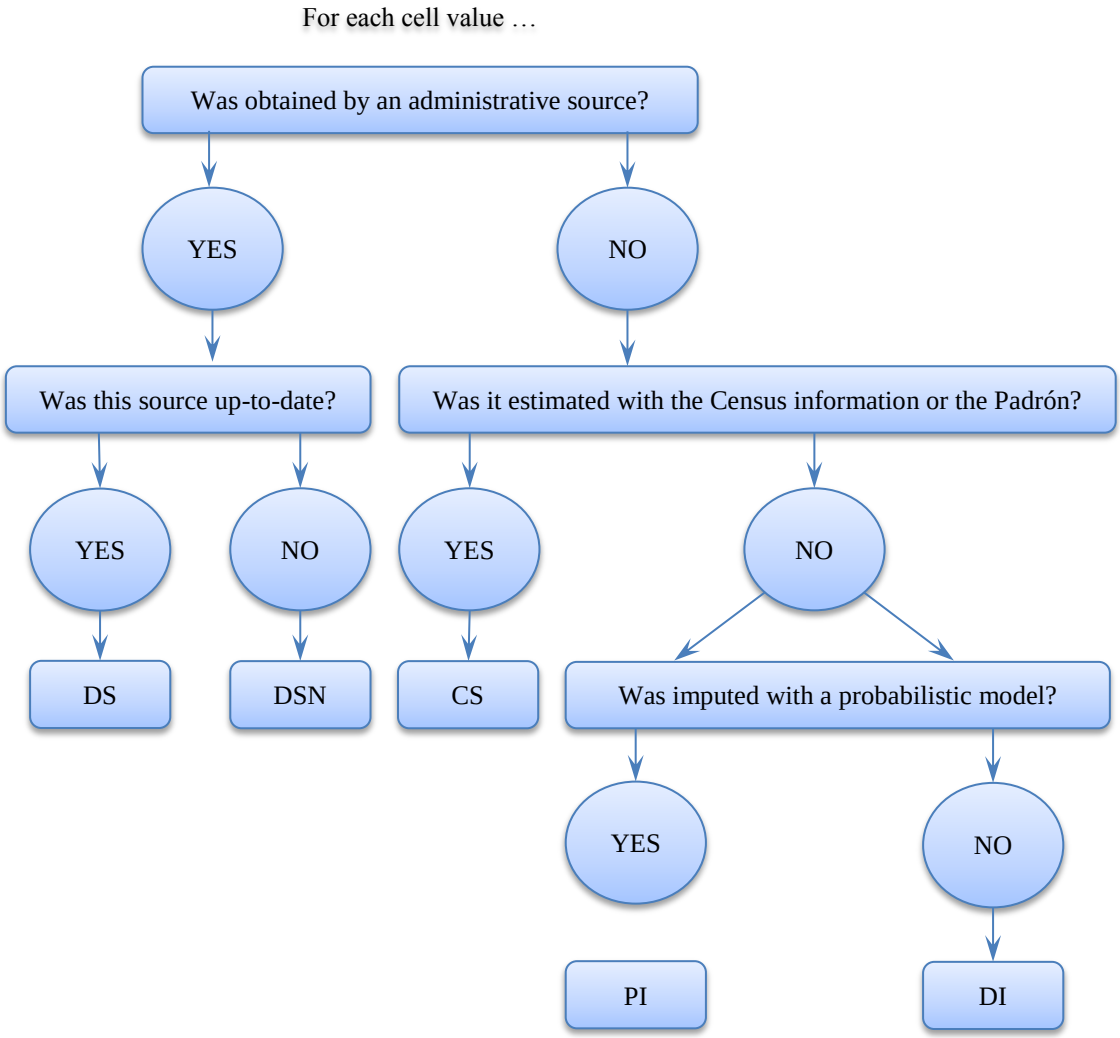
26. The procedure consists in the creation of a linked variable with five categories that will take into account various factors: the reference date of the source, the nature itself of the source whether it is specific for the phenomenon to estimate or not, and if any probabilistic or deterministic method was used to estimate the value.

27. The initial proposal of categories for all the variables is the following:

- DS: Information provided by direct sources up-to-date;
- DSN: Information provided by direct sources but not up-to-date;
- CS: Past Censuses information;
- PI: Probabilistic imputation;
- DI: Deterministic imputation.

28. Depending on the methodology and the nature of the variable itself the different registers will fall in different categories, and as a consequence, there might be categories with no observations.

Diagram 1
Categories of registers



29. Although this approach may seem relatively simple, it provides a very powerful tool for the quality measurement of the Census and it helps us to refine our methodology internally.

30. By considering the way we obtain each cell estimation, we will be able to quantify quality in a bidimensional basis: quality in terms of each register, and quality along a specific variable.

31. An analysis by columns (variables) across people, allows us to detect for every variable involved what is the percentage of records provided by different sources or methods and the percentage of imputed records. This information helps us to detect the quality of sources.

32. If we focus our research on rows (people) we can identify those records with the poorest quality level: those that have missing values or imputed information in most of variables. It is very plausible to find profiles of people with missing information difficult to estimate by administrative records such as foreigners or people living in deprived areas.

33. This meta-information could be used internally to conduct a specific survey in those geographic areas or to specific population groups with lower data quality or missing information in order to finally improve the whole Census data quality. In such cases, then the value of the quality variable related will be 'Direct Source'.

34. To better understand the proposal of quality measuring, we will now look to some examples of this row – column analysis.

III. Some examples of analysis by columns

A. Place of usual residence one year prior to the census (ROY)

35. Currently this variable is expected to be estimated with two sources: the PR, which includes information about changes of residence and dates related to them and the information of 2001 Census.

36. The PR was initiated in 1996 and it is a longitudinal register, which contains information about places of residence and dates of arrival since 1996 so it is an incalculable source of information in order to estimate migration variables such as the place of usual residence one year prior to the census.

37. Nevertheless, for all the register inhabitants the year of arrival to Spain is at maximum 1996 as there is no information for previous years, therefore there is no information for residence changes before 1996.

38. This lack of information in Padrón is completed with the use of the 2001 Census, which was exhaustive and includes information of places of residence before 1996.

39. In this case, there is no probabilistic imputation and so the only possible values of this variable are the following:

PLACE OF USUAL RESIDENCE ONE YEAR PRIOR TO THE CENSUS QUALITY MEASUREMENT

Deterministic imputation model	→	DI
Census2001 Census2011 sample	→	CS
Padrón	→	DS

40. The results for the whole population with reference date 1-January-2016 are:

ROY AT NUT S2 LEVEL	*%*
TOTAL	100.0%
Padrón	75.8%
2011/2001 Censuses	24.1%
Direct assignation (deterministic based on information in Padrón)	0.1%

ROY QUALITY	*%*
TOTAL	100.0%
DS	75.8%
NDS	0.0%
CS	24.1%
PI	0.0%
DI	0.1%

41. The direct assignation is based on a large group of deterministic rules thanks to the high correlation among all the residence variables included in Padrón and the last censuses.

42. Some consistency rules must be always complied, and in case of divergence of information among sources there is a deterministic decision. The coherence between dates and places of residence is checked altogether.

43. For example, if the dates of change of residence are missing in the PR and only the date of birth is registered, it is assumed that there has been no change of residence since then, so the previous place of residence is the same as the first one.

44. It must be underlined that the fact of obtaining information from a past Census does not imply a lower level of quality. In this case it completes the coverage of information and it also improves the residence situations registered in the Padrón.

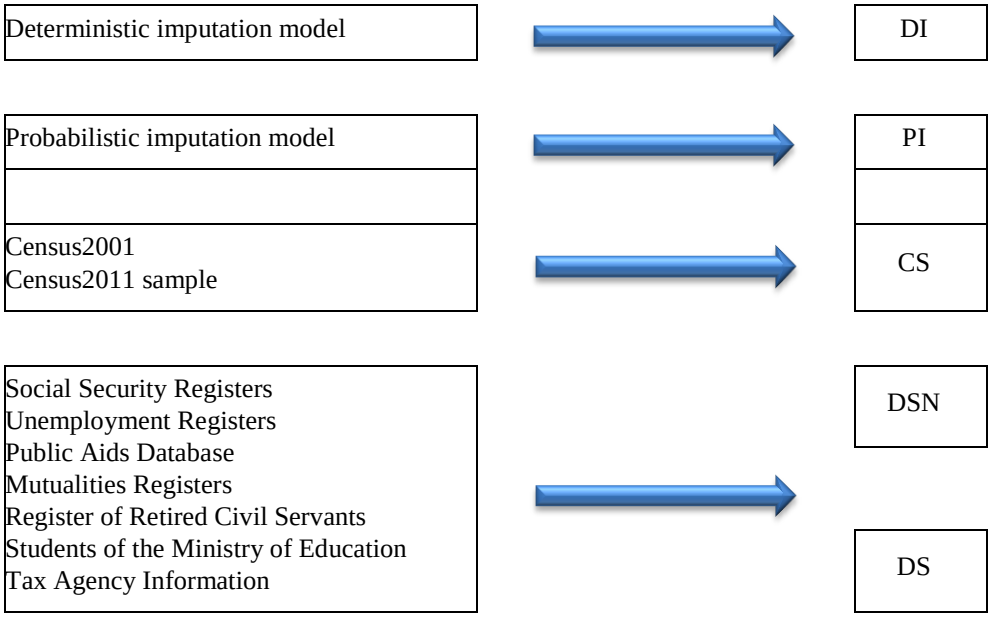
B. Current activity status (CAS)

45. For the estimation of this variable a long collection of sources is needed.

46. Today, apart from the Population Register and past Censuses information, all these different up-to-date sources available at the moment to estimate the current activity status: Social Security Registers, unemployment registers, public aids database, some registers with specific information for mutualities[5], the register of Retired Civil Servants, the register of students of the Ministry of Education and Tax Agency information.

[5] Mutualities are specific Social Security systems for specific population groups, such as Civil Servants, Judges, Armed Forces.
A civil servant or public servant is a person in the public sector employed for a government department or agency. The extent of civil servants of a state as part of the 'civil service' varies from country to country. In Spain, both workers in the field of public administration and in 'non-departmental public bodies' are considered civil servants.

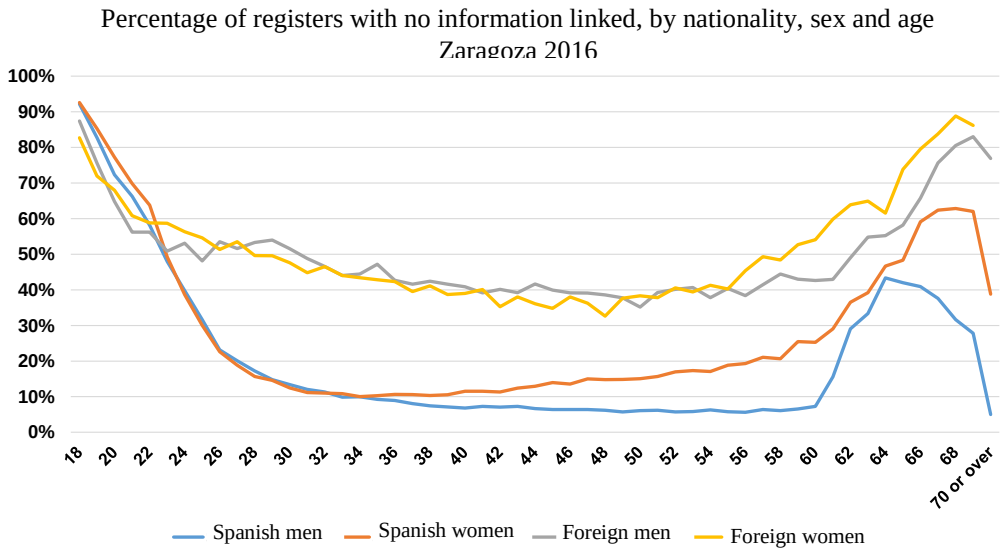
CURRENT ACTIVITY STATUS QUALITY MEASUREMENT



47. For this variable, the production process is not yet finished. For example, considering a particular geographical area in Spain (NUTS Level2), the 13.3% of the population has not be found in any register. These people are very likely inactive. The 86.7 of the information provides from a direct source. For those not linked cases, there is a high proportion of foreigners and people not in intermediate ages (youngest and oldest ones).

PERCENTAGE OF REGISTERS BY SOURCE OF INFORMATION TO ESTIMATE CAS

CAS Source of information	*%*
No information	13.3
Social Security registers	42.7
Public Aids register	17.3
Unemployment registers	8.9
Tax Office	7.9
Students	5.9
Civil service staff	2.4
Justice Staff	1.4
Armed forces staff	0.2
Retired civil servants	0.1



48. A high proportion of these not found persons will actually be inactive, but in other cases they could be employed in some specific jobs normally in the informal economy or not registered in fact such as the specific case of the domestic work.

49. As the possibility of finding them in any register or sources of information available is discarded, the possibilities are to impute deterministically or probabilistically whether they are inactive, unemployed or employed.

50. For the probabilistic imputation the intention is to consider external high quality sources such as the Labour Force Survey and build regression models. There is evidence that for some subpopulations we have difficulties to get their labour market information, such as foreigners or inactive persons.

51. These two examples illustrate how the assignment of a type of source improves the process and therefore the quality in the Census. When we carry out the 'column analysis', it is necessary to check the distributions for each variable with other existing information such as surveys to detect possible bias in data.

52. For each variable, coherence both at macro and micro level is examined. This contrasts of distributions and analysis of differences ensures the quality of the estimations algorithms.

53. For instance, in the case of the current activity status, the distribution is compared with other official figures published such as the LFS ones or the Unemployment figures.

IV. Analysis by rows

54. The mechanism of measuring each cell value quality opens the door to a row analysis, where the quality of the information we have for each person in the matrix can be quantified.

55. There might be specific profiles of people especially difficult to obtain their information, and also they could be residing in specific geographical areas.

56. For example, foreign retired people that come to the coast to live whose information is not available in administrative records since they don't pay taxes in Spain and so on.

57. The possible scenario of carrying out an ad-hoc survey in order to capture the information for specific subpopulations which are especially difficult to estimate by administrative sources may be evaluated by the help of this row analysis.

58. If a special subpopulation has a worrying percentage of missing values for many variables or the quality of the information is really low, INE may take the decision to conduct a survey in a geographical area to set this information directly from questionnaires rather than only provide the administrative data estimations.

59. In that case, the category of the type of source used may be considered as 'Direct source'.

60. For the moment, the integration of information has not finished yet so this decision would be evaluated in the following years.

61. It should be noted also, that some problems detected in the analysis by rows of the not linked registers depend on the 'skeleton' of persons in the PR. Although the PR is an excellent population database, there are some persons that are in fact not living in Spain anymore and should not be counted as total resident population. Consequently, the amount of not linked registers will be reduced, leading to better results in the proportion of linkage.

V. How will users see the extra information

62. The whole Census microdata will not be published due to confidentiality issues.

63. A selection of the Census population, between the 5% and the 10% of the total will be anonymised and published.

64. The most innovative aspect of this 2021 Census round is that for all the selected microdata uploaded in the web, every "type of source" variable will be also published. In this way, expert users will be able to evaluate in a better way the quality of the Census product.

65. Also, referring to the whole population, INE will publish a detailed quality report that will be released with many tables for each Census variable, focusing on the distribution of the type of source involved and further disaggregation by other interesting population characteristics such as geographical area of residence, sex, age, etc. This aggregated analysis will be made for each variable covering the whole population and housing Census.

VI. Conclusions and some final remarks

66. Spain has taken the challenge to conduct for first time the Census by linking administrative sources following the path of other countries in Europe that have proved this methodology to be solid and consistent. The access and content of the existing information in the administrative sources in Spain meets the requirements to produce the best Census product.

67. The quality focus for the next Census described in this paper will have two key uses: in first place, the measurement of its quality for users and secondly, it will enlighten the production process to help us detect some weaknesses in estimates and so better calculate the data.

68. We must help users to evaluate the quality of the new paradigm by providing information about the sources and methodology used to obtain the data. With this extra information of detailed tables for each variable specifying the distribution of the type of source used and methodology followed, and anonymised data with the meta information of type of source involved to estimate each value, we will help them to better understand the benefits of supporting the census information with administrative registers.

69. We believe that this procedure of assessing quality will receive a positive response from the users community.