**Assessing quality in a register-based census**

Marina Pérez Julián (\*), Head of Unit in the Census Department, marina.perez.julian@ine.es

Cristina Casaseca, Head of Unit in the Census Department,

Cristina.casaseca.polo@ine.es

Antonio Argüeso Jiménez, Director of Sociodemographic Statistics Department,

antonio.argueso.jimenez@ine.es

**Abstract**

*Spain is going to conduct for the first time a Census which will be largely based on registers. It will be built around the Population Register and will integrate administrative data from many sources.*

*In order to help users understand the production process and better interpret the results for each variable, another categorical variable measuring the quality of the information will be disseminated.*

*These variables provide a powerful way for assessing quality in the census data across columns (variables) and rows (people and households). As the process for elaborating the census 2021 is ongoing, the intermediate results of this analysis is also helpful in focusing our efforts. In this paper we present some preliminary results based on the first general test of integration of sources we have done so far, the so-called pre-census file 2016.*

**Keywords:** Quality, register-based census, integration of data sources

**1. Introduction**

Spain is already preparing its first Population and Housing Census largely based on administrative registers for 2021 following the path of other countries in Europe that have proved this methodology to be solid and consistent.

Spain is in a good position to change the Census methodology mainly due to three factors: firstly, the National Statistical Office in Spain (INE) has direct access to the Population Register, secondly the Law on Public Statistical Services (1989) empowers INE to compile information from the statistical services of ministerial departments and other bodies of the administration and in third place, the existence of a large collection of administrative records whose information is feasible for statistical use which is at INE’s disposal.

Although the work to carry out the biggest census in terms of population with this methodology for the first time is huge, it is well worth the effort.

The financial savings in the short and long term are obvious but this is not the main reason: some of the clear advantages are related with Eurostat intentions beyond 2021, shortening the production time of the Census and having census data continuously available, preventing users from having to wait 10 years for the next Census round.

By using administrative data, Spain will meet the main principles of statistical production in the code of good practice (relevance, precision, punctuality, coherence, comparability and access to the information). It will reduce bias by avoiding asking for information with questionnaires at the same time as it reduces the response burden. Comparability over time and other countries is guaranteed thanks to the use of consolidated administrative sources and the existence of a Commission Implementing Regulation (EU) 2017/881 of 23 May 2017 implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses.

The first general test of integration of sources has already been carried out in 2017, resulting in the so-called pre-census population file 2016 (PCF-2016), whose results will be shown in this paper.

Section two in this document briefly explains the construction of the pre-census file. Section three introduces the tool for measuring quality. Then the example of the legal marital status (LMS) estimation by administrative registers if taken as an example in section four. Section five includes a very preliminary “row” analysis based on signs of presence. Finally, in the sixth section, some conclusions are drawn.

**2. The pre-census file**

The Spanish population Census file will be built around the Population Register (PR), which 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.

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

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

**Figure 1. The population census final product.**



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

In order to help users understand the production process and better interpret the results we will provide extra information (metadata) of the sources or methods used for each cell value. A detailed breakdown of this extra information is provided in the following section. The pre-census file for 2016 (PCF-2016) has already been built following this strategy.

**3. Assigning a type of source value to every census cell**

For each census variable value (for example: legal marital status, maximum educational level attained, etc.) there will always be another one indicating the method or type of source used to provide the value for each person.

**Figure 2. The population census final product with metadata for each cell value.**



It must be noted that the methodology, peculiarities of the phenomenon being measured and the existing information vary for the different Census variables. For instance, 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. In other cases, there is always information direct from the Population Register (sex, age, place of residence….).

The procedure consists in the creation of a linked variable with provisionally five categories that will take into account various factors: the nature itself of the source, the reference date of the source, whether the source is specific for the phenomenon being estimated or not, and the many probabilistic or deterministic method used to estimate the value.The initial proposal of categories for all the variables is the following:

**Table 2. Initial proposal of categories of the type of source value**

|  |
| --- |
| 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. |

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. Though it may look simple, it is a very powerful tool for the quality measurement of the Census and enables it to be quantified in a bi-dimensional basis: quality in terms of each register (person), and quality along a specific variable.

An analysis by columns (variables) across people allows us to detect for every variable involved what the percentage of records provided by different sources or methods and the percentage of imputed records is. This information helps us to detect the quality in terms of the construction of the variable and sources. It must be underlined that the fact of obtaining information with the same method (for example deterministic imputation) for different variables may have different interpretations depending on each individual and variable. For example, for a 18 year old person estimating his/her LMS as single if there is no other evidence in administrative records may result very accurate, while if we estimate deterministically his/her occupation the precision may well be reduced.

And analysis by rows (people) allows us to identify those records with the poorest quality level: i.e., those that have missing values or imputed information in most 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.

It must be noted that, although for each cell value a quality indicator is assigned, this information will not be available due to confidentiality issues. Quality reports explaining how each variable is constructed and aggregated quality results would be released together with the anonymised microdata and tables (approximately 10% of the whole Census data).

However, the expert user would be able to work with anonymised microdata samples where these quality indicators are included so he/she would be able to perform a better analyse of the Census information.

**4. A practical example with the legal marital status**

According to the Commission Implementing Regulation (EU) 2017/881 of 23 May 2017 implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses the legal marital status must be estimated for the whole population.

Unfortunately, the PR does not contain this information and there is no unique source of information for every resident in Spain. Therefore, to reach a complete coverage of the Spanish population it is necessary to look at different registers. The ones used for this purpose for the PCF-2016 are: Public Aid Database, Marriage Bulletins from the Civil Registry Office, Divorce Sentences from the Civil Registry Office, Central Register of Foreigners, Tax Agency files, 2001 and 2011 Censuses.

Given that the possibility exists of a person being found in different registers at the same time, a set of priorities is established for those cases where a person’s legal marital status can be obtained from different data sources, and with sometimes conflicting results.

The priority rules are raised with the simple philosophy of discarding those persons whose information has already been found in a more reliable source than the remaining ones. Since the LMS estimation takes into account seven different sources of information, first we show the algorithm to understand the methodology adopted and afterwards we carry out the column analysis to evaluate the quality of the methodology.

The algorithm to estimate the variable is easily resumed with a set of consecutive questions checked for each person in the PR. If the answer to each question is positive, then a value for the legal marital status and type-of-source value is given. If the answer is negative, then the next question applies. For the non-assigned cases there is a final question that imputes the value depending on their age, information in past censuses, and number of members in the household. There is no probabilistic imputation for this variable, although some persons remain unassigned.

The registers whose information was taken from the 2001 and 2011 Censuses are categorised as ‘CS’, the Tax Agency Files from 2011 and the historical legal marital status found in the Tax Agency Personal Data Base are coded as ‘DSN’, all the others are direct sources up-to-date. Those registers whose value is determined in the end by rules combining their age, legal marital status in previous censuses and members in the household are coded as ‘DI’.

**Figure 3. Decision algorithm to estimate the legal marital status and type of source value.**



In the pre-census file 2016, the distribution of type of sources for the LMS estimates is:

**Table 3. Quality in the construction of the legal marital status.**

|  |  |
| --- | --- |
| **Type of source** | **%** |
| **Total** | **100.0%** |
| DS: Information provided by direct sources up-to-date. | 57.8% |
| DSN: Information provided by direct sources but not up-to-date. | 9.5% |
| CS: Past Censuses information. | 2.6% |
| PI: Probabilistic imputation. | 0.0% |
| DI: Deterministic imputation. | 24.5% |
| Not assigned | 5.5% |

The results for the LMS are optimistic as 57.8% of the legal population marital status was estimated by an up-to-date direct source. The fact of estimating a proportion of 24.5 of the population by deterministic imputation does not mean the quality is poor as it includes, for example, all people below legal age for marriage). In the case of this variable, the majority of singles are not in any register by default so taking into account other evidence and the age groups is a good method for capturing them. The information from past censuses is only used in 2.6% of the cases. We could then say that around 12.1% (2.6%+9.5%) of the cases could be improved by taking direct up-to-date sources covering the whole territory. Finally, 5.5% of the cases were not yet assigned but there are still some sources not included in the PCF-2016 that will certainly be used before 2021.

In the future, those remaining cases with no LMS assigned may be subject to probabilistic imputation taking into account some classic variables such as nationality, age according to already existing surveys.

If we focus now in the LMS values in the PCF-2016, 42.5% of the population was single, 41.7% married, 6.0 widow and 4.3% divorced. As we have previously seen 5.5% of the population remains unassigned after having checked each register and carrying out the mentioned imputations.

**Table 4. Legal marital status results for the pre-census file 2016 and comparison with the Continuous Household Survey 2016.**

|  |  |  |  |
| --- | --- | --- | --- |
|   | **% Pre-census file 2016** | **% CHS 2016 (\*)** | **Differences CHS vs PCF** |
| **Total population** | **100.0%** | **100.0%** |  |
| Single | 42.5% | 44.4% | 1,9% |
| Married | 41.7% | 43.7% | 2,0% |
| Widow | 6.0% | 6.3% | 0,3% |
| Divorced | 4.3% | 5.6% | 1,3% |
| Not assigned | 5.5% |  | -5.5% |

The results show close similarity with the Continuous Household Survey (CHS) for 2016 (a continuous survey that collects around 55,000 questionnaires every year). The better the coverage of sources, the fewer the non-assigned cases. This would result in lower gaps for the LMS categories which are currently explained by the existence of non-assigned cases.

**5. Analysis by rows**

The complete analysis checking the quality of all the persons in the matrix is currently ongoing.

The first analysis in terms of signs of life, checking which registers remain not found, gives a good approximation of part of the row analysis and have enlightened the production process detecting specific problematic profiles in the population.

To determine the population figures at the PCF reference date (1st January 2016) we consider the whole Population Register and prove if each person is still living in the country by checking if the register has not been deleted due to expiration dates, especially for the foreign people. In addition to this, we use a ‘signs of life’ method to have evidence of the current residence in Spain of each person in the PR.

For the ‘signs of life’ method four important economical sources are used: The Tax Agency Files, the Social Security Files, the Unemployment Office Database and the Public Aids Database.

After cleaning the PR file with those people that are not currently living in Spain at the reference date, we proceed to analyse those registers that were not found in any administrative source.

In the PFC 2016, only 2.1% of the whole population was living in a household were not a single member was found in an administrative register.

The distribution of the missing persons is explanatory. The lack of administrative information is concentrated in the foreign people (10.5% of the foreigners living in Spain are not found in any register while only 1.1% of the Spaniards remain not found). If we analyse the foreign people that were not found in the administrative sources, there is a clear distribution of nationalities (37.1% of English inhabitants are not found, 31.4 of Russians, 30.5% of Germans followed by the Nordic Countries). Other nationalities such as Americans, Moroccans are well collected in our registers.

A residual percentage is associated with children with no identification number.

There is a higher concentration of not found persons along the Mediterranean coast and the Baleares and Canary Islands, which are typical touristic areas in Spain. In some provinces (Alicante, Malaga) 1 out of 4 foreigners living in the region is not found in any administrative register.

In the following map, we represent all geographical areas according to the proportion of non-found persons. The darker regions represent those municipalities with more missing information in registers.

**Cartogram 1. Geographical distribution of non-found persons in administrative registers.**



If we focus with more geographical detail to the Census geographical areas of some of these provinces with the highest proportions of not found persons it is obvious that the problem is concentrated in specific municipalities with high presence of foreign inhabitants registering more than 30% of the population as not found.

**Figure 4. Geographical distribution of non-found persons in administrative registers in Alicante.**



Thanks to this preliminary analysis the chances are that the lack of presence in the administrative registers might be conditioned by some factors apart from the nationality and tax scheme applicable out of Spain, such as age of the reason of residing in Spain (holidays, retirement, work…).

This motivates the possibility of carrying out some survey tests to analyse whether by a direct interview the information for specific persons could be completed, or otherwise to interview the totality of specific geographical regions where the presence of problematic population is high.

**6. Conclusions**

The access and content of the existing information in the administrative sources in Spain meets the requirements to produce the best Census product for 2021.

The quality focus for the next Census described in this paper has two key uses: in first place, the measurement of its quality for users and secondly, to enlighten the production process so that we can detect some weaknesses in estimates and therefore better calculate the data.

We must help users evaluate the quality of the new paradigm by providing information about the sources and methodology used to obtain the data. All 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 will help them better understand the benefits of supporting the census information with administrative registers.

The pre-census population file 2016 quality analysis showed the big advances made and good results obtained and also pointed out the weaknesses and strengths of each variable estimation and lack of information within specific profiles of the population. Thanks to it we will continue moving in the right direction in order to have the best census product for 2021 in terms of quality.

We are working in possible refinements of the quality measurement variables for a better understanding of users and comparability among different census variables.

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