A population census based on registers and a "10% survey". Methodological challenges and conclusions.

Antonio Argüeso*
National Statistics Institute, Madrid, Spain antonio.argueso.jimenez@ine.es

Jorge L. Vega
National Statistics Institute, Madrid, Spain jorgeluis.vega.valle@ine.es

Abstract

The population and housing census 2011 in Spain has followed a new approach. Instead of an exhaustive door-to-door enumeration of the population, it has been conceived as a combination of two elements: registers and survey. The overall cost of the census, 85 million euros, is 20% the one of a classical census. We can benefit from the existence of a Population Register of high quality in Spain. Taking this register as its backbone, a *census file* was created using all available administrative registers (tax collection agency, social security, vital statistics…). These registers were used to provide, for every person, a measure of *proof of residence*. But some 2.2% of the population remained as being of *doubtful residence* in the registers, because that there was no sufficient evidence in registers to count them. These people were grouped into homogeneous clusters (region, sex, age, citizenship…) and were counted in the census using weights (called *count factors*) obtained from the survey. This survey was actually the second element of the census. It consisted on a fieldwork operation including a classical building census, that enabled geo-referencing every building and a sampling survey addressed to 10% of the population, in order not only to get those weights but also, and mainly, to provide characteristics of persons and dwellings. This "10% survey" was collected using a sequential multi-channel method, thus promoting the Internet as first option (38% of questionnaires were collected through this channel). Eventually, the census, as a product, is the mixture of two components: a *weighted census file* containing around 47 million registers but only with a few variables for every person (those contained in the population register) and a file containing many variables, the ones collected in the questionnaires, but only for 4.2 million people. This system has similarities with censuses based in long and short forms. But providing consistency between detailed data coming from the survey and the main figures coming from the weighted census file has also proved to be a challenge. Some inconsistencies show up and consequently, more than ever, the dissemination and explanation to users becomes crucial.

Key Words: administrative registers, multi-channel surveys, population register.

1. The population and Housing Census 2011. General Approach

The 2011 Population and Housing Census in Spain combined the following elements:

- A "**Weighted Census File**": it was created using available administrative registers, taking the Population Register as the basic element of its structure.

- A fieldwork operation including two components:
  - A comprehensive **Building Census** that enabled geo-referencing all buildings and ascertaining their characteristics.
  - A large **sampling survey**, aimed at a relatively high percentage of population, to get the rest of characteristics of persons and dwellings.

The following are some of the key aspects in the strategy of the 2011 Census:

- The population figures were obtained counting records (persons) contained in the *Weighted Census File (WCF).* It is "weighted" because a very small percentage (2.2 %) of people are counted using weights called *count factors,* obtained from the survey.

- The theoretical sampling rate of the survey was approximately 12% of population. The sampling rate was not uniform, as a minimum amount of information has to be provided for all municipalities. It implies that a minimum sample size has to be allocated for every municipality: there are 8,000 of them in Spain, and their population is ranging from 4 people to 3 million people (almost 5,000 municipalities having less than 1,000 inhabitants). Thus, the sampling rate ranged from 100% -exhaustive census- for municipalities below 100 inhabitants to 9% for big cities.

- The Buildings Census was based on a "pre-census file", made fundamentally from a combination between the data from the 2001 Census, the Population Register and the Cadastre. Using this pre-census file as an input, in 20% of territory there was no need to perform any fieldwork as the information was considered to have enough quality. For the rest of territory a complete enumeration of buildings was made. As a result of this process GPS coordinates were given to every building (containing dwellings) in Spain.

In section two we describe briefly how the WCF is created and how population figures are obtained; in section three the second pillar of the system, the fieldwork operation, is described. Section four contains a flash explanation about how data from the two sources are combined to provide all census information and finally in section five some conclusions about the overall operation are drawn.

2. The Weighted Census File (WCF)

The first step in 2011 Census is to build the so-called *pre-census file*. The population register in Spain provides each year official population figures at the municipal level. Created in 1996, it incorporated all residents in each municipality regardless of their legal status. Therefore, its purpose is to collect resident population, with the same definition of the census. But because of its nature, as it is an administrative file, it may contain errors, mainly because it does not adequately reflect emigration (some people leaving the country are not properly deleted from the file). **The (registered) population of Spain as of January 1st 2012 was 47,265,321 inhabitants, being 5.736.258 of them of foreign nationality (12.1%).**

In order to get population figures for the census the strategy followed was to divide the population register into three groups: *sure population* (about which there is enough evidence to consider them "without any doubt" resident population), *errors* (there is enough evidence to delete those records because they are wrongly included, mainly deaths not incorporated to the population register) and *doubtful population* (there is no sufficient evidence whether to count them or not). The doubtful records were counted therefore as resident population but weighted using the so-called *count factors*. These *count factors* reflect the uncertainty assigned to this population and are calculated from by the survey using the procedure explained below.

The population register was crossed with various administrative records, especially Social Security and the Tax Collection Agency in order to obtain further evidence on the residence of people included in it rather than simply being registered. Also some records were added, mainly recent births registered in vital statistics but still not reflected in the

population register. The *pre-census file* is the result of this process, with roughly 47.4 million records.

Around 15 different criteria by which a record could be considered doubtful were defined. Eventually, 97.7% of the pre-census file records were considered *sure* (count factor equal to 1), 0.1% were considered errors (count factor 0, then those records were deleted) and for 2.2% of records the evidences were inconclusive, so they were considered *doubtful* (and the count factor had to be assigned according to the results of the survey). The number of doubtful records was 1,040,000 people, 87% of them being foreigners.

2.1 Partitioning the pre-census file into classes and assigning Count Factors.

In order to assign count factors, these doubtful registers were grouped into classes. The whole pre-census file was partitioned into classes (groups of people defined by demographic characteristics). These characteristics determining the classes are age, nationality and province of residence (there are 52 provinces in Spain). Each class may contain sure and doubtful records. If a class does not contain any doubtful record, then no estimation is needed, and the population for that class is the one coming from the pre-census file.

All classes containing doubtful records were configured to include at least 1,000 of them. Classes with less than 1,000 doubtful records were grouped in order to reach at least that threshold. The grouping strategy was, firstly wider age groups, secondly wider territory selection and lastly grouping nationalities.

724 clusters or classes were made, with at least 1,000 doubtful records in each. As mentioned, a survey of 12% of the population was carried out (percentage of data finally collected was 10%). The population from the survey is grouped into the same 724 classes according to the same characteristics (age, territory, nationality). After that, the sample and the pre-census file records are linked at the individual level to determine which records from the sample had been previously classified as doubtful in the pre-census file.

In order to calculate these count factors we need to estimate the proportion of *sure population* in the survey. In a given class i of the pre-census file, the total number of records, T_i can be split into the number of records of sure population S_i and the doubtful ones D_i, Thus, the proportion of sure population in the pre-census file, for this class i, P_i, is:

$$P_i = \frac{S_i}{S_i + D_i} = \frac{S_i}{T_i}$$

Considering that a part of this doubtful population should not be counted, we can use the data of the survey to estimate the actual value of this proportion, $\hat{p}_i$ as: $\hat{p}_i = \frac{\hat{s}_i}{\hat{t}_i}$

Where $\hat{t}_i$ is the estimation of population of class i and $\hat{s}_i$ is the estimation of sure records. We prefer not to estimate doubtful population but total and sure population only; by doing so, all people that are not included in the population register but filling a questionnaire will be counted in $\hat{t}_i$ but not in $\hat{s}_i$, so as to avoid under-estimation of doubtful people. Then, the number of doubtful people actually estimated in the survey for class i is: $\hat{d}_i = \hat{t}_i - \hat{s}_i$.

We consider then $\hat{p}_i$ as our estimation of the actual proportion of doubtful population. We assign the same count factor (CF_i) to all doubtful records in class i so that the estimated population for class i will be: $\hat{T}_i = S_i + CF_i * D_i$

Then, $$\hat{P}_i = \frac{S_i}{\hat{T}_i} = \frac{S_i}{S_i + (CF_i * D_i)} = \hat{p}_i = \frac{\hat{s}_i}{\hat{t}_i}$$

Thus, in every class i, the count factor CF_i assigned to all doubtful records, can be calculated as:

$$CF_i = \frac{S_i(\frac{\hat{t}_i}{\hat{s}_i} - 1)}{D_i} = \frac{\hat{d}_i / \hat{s}_i}{D_i / S_i}$$

Where S_i and D_i come from the pre-census file and $\hat{s}_i, \hat{d}_i$ are derived from the survey. The count factor is then the ratio of doubtful records in the survey compared to the one calculated in the pre-census file, and it that may be greater, less than or equal 1 because there might be non-registered people filling the survey questionnaire.

With those count factors incorporated to every doubtful record, the pre-census file becomes the Weighted Census File (WCF) and allows obtaining census population figures from it. The population of a given geographic area, T_i, is obtained as the sum of count factors of WCF records in that area (given that CF=1 for all sure records).

This procedure provides statistical figures for the census, but of course it cannot be used for administrative purposes to update the population register, since it is not be possible to determine which individual citizen should be counted and which one should not.

2.3 Doubtful population and census results. Some figures

As stated before the number of doubtful records were 1,040,000 in round numbers. The average count factor for all classes was 0.424, which means that these 1.04 million doubtful records are counted as a population of approximately 440,000 people.

The most prominent class in terms of size of doubtful population is British nationals aged 60-64 in the province of Alicante, consisting of about 19,300 people, of which 5,960 (over 30%) are considered doubtful. With the results of the survey the count factor for this class was 0.36 and the population was reduced to 15,485. The same happens with German citizens also in Alicante. It has to be considered that many Germans and Britons go to this place in the Mediterranean coast, to live after their retirement. This result is then not surprising.

In 22 classes the CF was higher than 1, meaning that these populations are considered to be sub-registered. It is the case of nationals of Pakistan aged between 25 and 30, with a count factor 1.67 but affecting only 1,055 doubtful records that are consequently counted as a population of 1,762.

As a result of these calculations, the population of Spain reaches 46,815,916 inhabitants as of November 1st 2011, some 450,000 inhabitants below the one provided by the Population Register (reference periods differ in two months which is almost negligible). Number of Spaniards hardly change (the census accounts for 30,000 more people due to

the aforementioned births included in vital statistics) but in the case of foreigners the census decreases the figures in 480,000, to a total number of 5,252,473 (11.2 % of the population of Spain, 8% less foreigners than those considered in the population register).

3. The fieldwork operation: the *Building Census* and the *Households Survey*

In October 2011, about 2.2 million letters were sent to households asking to fill in the census questionnaire online. The procedure followed for data collection was a sequential multi-channel survey, so that respondents were asked to fill the questionnaire only on the Internet as first option. For those not answering, reminders were sent where they were also given the possibility to fill it in paper or to call a help-line (not actually promoted as an option to fill the questionnaire). Only after several reminders, a sub-sample (of around 50%) of those households not cooperating so far received a visit of INE's staff (and a face-to-face CAPI interview was made then).
In parallel with the first stages of data collection 4,000 enumerators and 900 group managers were contracted. Their first duty was to collect the Building Census and after finishing it they interviewed the sub-sample of households mentioned before.
During the Building Census some new households were also selected: a sample of those living in buildings not previously contained in the directory and consequently with zero probability of being selected otherwise (it was called the supplementary sample).

In terms of human resources, the Building Census was made with almost 5,000 people in 2,5 months and the (rest of) Household Survey with the same 5,000 people for one month, right after finishing the Building Census. Including training, enumerators were hired for some 4 months. It sums up to 16 % the amount of human resources needed for the 2001 census (43,000 people between enumerators and group managers, for three months).

The effective sample size was 4.2 million people, corresponding to 1.65 million occupied dwellings (some information was also collected for 600,000 empty or secondary dwellings).

The results by channel were 38% collected through the Internet (CAWI), 52% in paper and 10% via face-to-face CAPI interview.

4. Combining results from a file and a survey: calibration and consistency

As explained before, the population figures are not directly drawn form the survey but from the WCF. This file contains only 4 variables for every record (or person): age, sex, nationality, place of birth. Thus, for a given region or municipality, a table like *population by sex, age group and nationality* is obtained from the WCF. But a table like *population by age group and household size* is not, since the WCF does not contain anything about households. This table must be generated from data collected in the survey.

To minimize inconsistency between these two sources, the results of the Household Survey were massively calibrated to the ones of the WCF. For example for a small municipality of 10,000 inhabitants calibration is made by sex and (8) age groups, and by sex and nationality (only Spaniards/foreigners). It means that 20 figures are calibrated. But inconsistencies show beyond those calibrated data. In the example above, for this (not so) small municipality, if we were retrieving from any of the two sources (WCF or the Household survey) a table like population by sex (2), age (20) and nationality (5), it would provide (2*20*5=) 200 data, 20 of them would be calibrated and consequently exactly the same, but the other 180 of them could differ. Inconsistencies show immediately and the impact of it is difficult to minimise since we do not publish a limited

set of tables but a *Datawarehouse system* where users may create an almost unlimited number of tables (of course with the boundaries of confidentiality and quality of the request).

This *Datawarehouse System* (still under preparation) is a crucial element of the new census. As the survey is geo-referenced (because all the buildings are), the system will allow performing queries that go beyond administrative boundaries. The user will have the possibility to get information like how many people live at less than 1 km away from the coastline or within a given irregular polygon drawn by the user (if enough households are selected to permit the query). The potential usefulness of census data increases exponentially. All these data will always come from the survey, never from the WCF, as it is the survey the source that is geo-referenced.

5. Some conclusions

In our opinion the quality of figures coming from a census like this is much higher than the one of a traditional door-to-door enumeration. In a country like Spain, where there is a Population Register fully consolidated and where there are some important registers like those mentioned before, the amount of information that we have about people is enormous and it is worth using it for counting people, instead of knocking on every door. And this is even clearer when a pre-census process like the one we did provides a kind of certainty of residence for 98% of population.

In terms of money saved, we cannot have a perfect comparison between 2001 and 2011 censuses notably because the 43,000 agents contracted in 2001 had working conditions not applicable today in the Spanish public administration (they were paid in terms of *pay per questionnaire collected*). But in rough terms we can say that a classical census in 2011 would have cost something around 500 to 550 million euros and the actual cost of it has been 85 million.

The weakest point of the fieldwork operation is the timing of the process. In a sequential multi-channel data collection it takes a lot of time to collect all data. Some households that received a first communication from INE in the last week of September 2011 were finally visited by an agent in March 2012, after four or five letters. Data collection spreads in 6 months, though 90% of data are collected in three. It makes it more difficult to state that the reference day is November 1st. It is, in terms of counting births, deaths and so on, but not so much in terms of characteristics of the population.

As regards calibration of data, we have to recognise that there is not a perfect solution for inconsistencies. Figures not calibrated are different using different sources (the WCF or the household Survey) and both sets of figures will have to be or have already been published. The best thing we can do is, in our opinion, be transparent and clear when explaining the methodology of the census.

References:

Population and Housing Census 2011. INE.
www.ine.es/jaxi/menu.do;jsessionid=82DAB3C1D2D6066AE9D26E022A0A6B1D.jaxi03?type=pcaxis&path=%2Ft20%2Fe242&file=inebase&L=1