

# Moving towards a register based census in Spain

Jorge L. Vega Valle\*, Antonio Argüeso Jiménez and Marina Pérez Julián  
*Socio-Demographic Directorate, National Statistics Institute, Madrid, Spain*

**Abstract.** Spain plans to carry out its first register-based census in the next 2021 round, becoming one of the biggest countries in the world using this approach.

Since 1996, when the Population Register was created, census methodologies have evolved enormously and Spain has vastly increased the use of administrative registers in official statistics. In the last 2011 census, it was too soon to completely rely on administrative data, so a combined census was performed.

Now, in 2018, 22 years after the creation of this register, INE faces this challenge with confidence. The availability of a variety of administrative registers and the fact that INE has access guaranteed by law to them, endorses this step forward.

There are still some difficulties to overcome and a great effort is being made today, collecting any suitable administrative source, processing and integrating data, developing new IT tools, and of course, evaluating the quality of the whole statistical product.

A complete population census test referenced to 2016 proved very satisfactory which enables for Spain planning to conduct a register-based census in 2021.

Keywords: Population and Housing Census, administrative registers, survey

## 1. Introduction

Once the 2011 Census was completed, the 2021 Population and Housing Census project was proposed to be based primarily on the treatment of administrative records as it was advanced in [1]. In order to make a final decision about the methodology to be used in 2021, a work based on information from different registers has been carried out. This exercise can be considered a pilot of the Population Census and it is referred to January 1st, 2016. Internally the information built is known as the 2016 preCensus File (2016-Pilot). Additionally, the first viability study of the Housing Census, also based on registers, has also been conducted.

Regarding the Population Census, the results of the feasibility analysis of the 2016-Pilot are conclusive: it is perfectly feasible to base Census only in the combination of administrative records. In certain and en-

closed cases it would be necessary to complete the information with statistical procedures of imputation, but not to a greater extent than what is used in traditional censuses. The minimum level of requirement,<sup>1</sup> which is the legal obligation established by the EU Regulation, is amply met for all the variables referring to the population.

According to the Housing Census, developments have been done, but its current situation is not so advanced. Anyway although it is proven that the EU regulation would also be complied with the strategy adopted, more efforts and time are needed to analyze the final quality of the product. In the least favorable scenario, the Housing Census would require a reduced and selective fieldwork operation in certain enumeration areas (those with poorest quality of information) of the national territory.

On the other hand, it is proposed to complete the 2021 Census project with a sociodemographic survey

\*Corresponding author: Jorge L. Vega Valle, Socio-Demographic Directorate, National Statistics Institute (INE), Madrid, Spain. E-mail: jorgeluis.vega.valle@ine.es.

<sup>1</sup>More details in: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0712>.

parallel to the census. The size of this survey could reach 1% of the total population and would allow to provide information useful for imputation of those variables that are not sufficiently well-covered by the Population and Housing Census. In addition, information of this survey has been largely demanded among users.

As a conclusion, 2021 Spanish Census can be considered closed from the methodological point of view and the main doubts are solved. However, during the following years, an intense work of refinement in all the variables, especially in the Housing part, is required. In addition, the methodology based on registers entails many other advantages, as it can be seen in detail in [2], such as the possibility of having information much more frequently, the increase in quality, the reduction of citizen's burden and even savings in long-term costs.

## 2. The population register (Padrón)

The main source as regards both population stocks and migration statistics in Spain is the population register, named Padrón in Spanish. Padrón is the official list of residents in each one of the 8,124 municipalities in Spain (as of 1<sup>st</sup> January 2018).

In Spain there are as many registers as municipalities. But there is a law, in force since 1996, integrating all these municipal lists into a single national database. There are also legal procedures to keep this database and the municipal files interconnected and updated on a monthly basis. This is made through the statistical office of Spain, INE. So, unlike other countries where the police or other administrative bodies are in charge of population registers, in the case of Spain, INE is the national institution that coordinates this single national population register.

Each month INE receives all the changes produced in every municipality. With this information INE performs validations and forwards these results to all the municipalities, to avoid duplications and also to include deaths, births or acquisition of Spanish citizenship that INE receives on a monthly basis from the Civil Register. Furthermore, all the consular offices of Spain throughout the world (around 250) are also connected to Padrón like the municipalities.

According to the law, there is no restriction for registering in Spain in terms of legal situation. All people living or willing to live in Spain, regardless of their legal situation, have the right (and obligation) to be registered.

For each person, Padrón contains these variables: gender, date of birth, place of birth (country, in the case of foreigners), nationality, educational attainment and national identity number. Foreigners in legal situation are recorded with the number of residence permit and people in irregular situation with the passport number.

Because of all the details mentioned above and also because the quality of this source has improved every year since the beginning, Padrón is the main element on which the Population Census is built. The 2016-Pilot is created using Padrón as of January 1<sup>st</sup> 2016 as starting point. This constitutes the backbone of the population file and dozens of administrative records of different nature have been linked to it in order to compose an individual information similar to the one that would be obtained with a questionnaire addressed to the households. The main data-linkage method is based on the equality of the national identity number but other types of techniques are also used (for example the creation of an internal "supercode" based on basic variables like name, sex and date of birth or probabilistic links).

## 3. Two questions that must be answered by a Census

Every census must provide information on two questions. First, it should provide information about the number of people living in the country with a high level of geographical detail. In addition, it should provide detailed characteristics (for example current activity status, year of arrival to the country or educational attainment) about all these people. From a graphic point of view, both questions can be seen as a large single file of two dimensions. The population figure will determine the number of records in the file (in vertical) and the number of variables will determine its amplitude (in horizontal).

### 3.1. *How many people live in Spain? The signs-of-life method*

The original purpose of the censuses was to count the population. In Spain this main objective is not so important anymore, because due to the existence of Padrón, the uncertainty in knowing the population figures during the intercensal periods has disappeared.

In the same way as in the 2011 Census, 2021 Census takes Padrón as its basic element. The next step involves performing several additional controls:

Table 1  
Graphical representation of Census information

Identification	Sex	Age	Place of birth	Legal marital status	Current activity status	Educational attainment	...	...	...
1									
2									
3									
...									
...									
...									

- Verification that births and deaths are up-to-date as of the reference date.
- Analysis of the expiration date of Padron foreigners (every 2 or 5 years depending on their particular situation, foreigners should update their presence in Padrón).
- Detection of “signs of life” of those people that appear in other files like Tax Agency or Social Security.
- Application of the 12-month criterion.<sup>2</sup>

Using all this information, a counting algorithm has been designed (still provisional and also susceptible of improvement) that provides a provisional population figure based on the signs-of-life method. As it can be seen in the following table, results obtained by the algorithm used in the 2016-Pilot are very similar to those provided by Padrón.

### 3.2. What are their characteristics?

If we take a look to the 2011 Spanish Population Census we can see that the individual questionnaire contained 22 questions, which can be grouped into eight blocks (basic demographic data, migrations, legal marital status and dwelling composition, studies, care provided, fertility, current activity status and mobility). Apart from that, the household questionnaire contained 7 questions about characteristics, facilities and equipment of the dwelling as well as the tenure status.

The main objective of the 2016-Pilot was precisely to try to replicate the 22 individual questions; that is, to investigate to what extent a register-based census could at least contain the same level of detail as the 2011 census. It must be taken into consideration that not all the questions of the 2011 Population Census questionnaire

have to be included in the 2021 census, in the same way that questions have been appearing and disappearing in each census along the years for many different reasons.

The first conclusion about the analysis of the 2016-Pilot is that, regarding to the population, all the variables that are compulsory according to the EU regulation<sup>3</sup> can be obtained from registers. In the following table can be schematically seen the main data source for some relevant variables.

There are other variables<sup>4</sup> that are not compulsory but have a great census tradition or big interest among the community of users; in almost all the cases it is possible to obtain this information from registers. More information about the technical specifications of the different variables can be found in [3]. In a very schematic way, below are the sources used in 2016-Pilot for the most representative variables of the Census, as well as some of their most relevant characteristics.

#### 3.2.1. Migration variables

Method and sources: Population register, but because it was not put into operation until 1996, it is also necessary to use the 2001 exhaustive Census.

General issues: In general, the quality of these variables is very high and all the information included in the 2011 Census questionnaire (years of arrival and previous places of residence) can be obtained from registers.

#### 3.2.2. Relationship between household members (and other derived variables)

Method and sources: The starting point is the Padrón, which contains information about the people that live in the same household. Information from the Tax Agency, Births and Marriages Bulletins, the Police

<sup>2</sup>According to the current European regulation, the following persons should be considered as usual residents in a geographical area: those who have lived in their place of usual residence for a continuous period of at least 12 months or those who arrived in the place of usual residence during the 12 months before the reference date with the intention of staying there for a least one year.

<sup>3</sup>More details in: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0543&rid=7>.

<sup>4</sup>For example some migration variables like year of arrival in the municipality or in the dwelling.

Table 2  
Population figures according to different methods

	2016-Pilot	Padrón	Difference (PCF- <i>Padrón</i> )
TOTAL	46,581,681	46,557,008	24,673
Spanish	41,972,389	41,938,427	33,962
Foreigners	4,609,292	4,618,581	-9,289

Table 3  
List of most important Census variables and main data source

Variable	Main data source
Sex, age, country of birth, country of citizenship, year of arrival in the country	<i>Padrón</i>
Legal marital status	Tax Agency, Vital statistics bulletins
Current activity status	Social security registers, unemployment register, Public Aids Database, Mutualities Registers, Register of Retired Civil Servants, ...
Occupation, Industry	Social security registers, Mutualities Registers
Education attainment	Ministry of education
Tenure status of households	Tax Agency
Useful floor space and dwellings by period of construction	Cadastré

Database (father's and mother's name are stored for every person) and previous Censuses are also used.

General issues: For each person it is analyzed who is his/her father, mother, spouse and other relative. Next, derived variables related to the families, nuclei and structure of the household will be generated. In general, the checks to detect the father and mother (common surnames and certain restrictions of age and sex) are much simpler than in the case of a couple. Regarding the couples, if they have a child living with them at home, they are much easier to detect. In other more complex cases (like cohabiting or same-sex couples without children), it will be necessary to perform an imputation model based on an external survey (and the detailed distribution of variables like age, sex or date of arrival to the dwelling) that assigns couples to some situations of people living together.

### 3.2.3. Educational attainment

Method and sources: *Padrón* contains information about this variable, but since its quality is not optimal it is necessary to use other sources like: several registers (diplomas, graduated people) from the Ministry of Education, information stored in the Unemployment Register, 2001 and 2011 Census. Information from previous censuses will only be used for people of a "certain age" that do not appear in the ministry records.

General issues: Information of this variable will only be provided for people aged 15 or more. A person will be assigned the highest educational level that has been observed in the different sources for this variable. Those residual cases that lack educational level and other cases where educational attainment was not assigned a single value, are imputed taking into account

the distributions according to sex, age and place of residence in an external survey.

### 3.2.4. Current activity status

Method and sources: This is one of the variables that involves a greater number of sources: Social Security Registers, Unemployment Registers, Public Aids Database, Mutualities Registers, Register of Retired Civil Servants, Registers of Students, Tax Agency information, 2001 and 2011 Census.

General issues: Information of this variable will only be provided for people aged 15 or more. Using all the sources mentioned above, it is normal to have, in some cases, conflicts among data sources. In order to solve this issue, priority rules based on the recommendations of the United Nations and the European Regulation for Censuses have been used. It is normal that some people (for example women aged 55 or more) do not appear in any of the sources used and they will be classified as "others" inside the group Outside of the labour force. The results obtained for this variable in the 2016-PFC are very similar to those one of the LFS although it is still necessary to refine the category of the unemployed so that it resembles the ILO recommendation.

## 4. A new quality framework

As it was advanced in [4], one of the main novelties of the 2021 Census will be the inclusion of a new mechanism that enables users to evaluate the quality of each of the census variables and which will also increase the amount and transparency of information disseminated by the INE. The idea is to create for each

variable (for example: legal marital status, educational level attained, etc.) a new one that would store information indicating the method or type of source used to provide the value for every person.

The procedure will consist in the creation (for each Census variable) of a new derived variable with several categories that will take into account various factors reflecting if it is a direct, indirect source or if the information has been imputed.

If we focus on the way we obtain the cell estimation, we will be able to quantify quality in a two dimensional basis: quality along a specific variable and quality in terms of each person.

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 the sources.

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

With this information our users will have more information available that will be useful to understand better the benefits of supporting the census information with administrative registers.

## 5. The Housing Census

The Housing Census consists, first of all, in an exhaustive quantification of all the dwellings (occupied and unoccupied), and secondly in a characterization of them and of the building where they are allocated. Similar to what has been described for the Population Census, it is planned to build an exhaustive microdata Housing file that contains all the dwellings and their characteristics, using the administrative sources available.

In the same way as Padron is the main source of information for the Population Census, we could say that Cadastre will be the main source of data for the Housing Census. It has the great advantage that contains a unique identifier for all dwellings (cadastral reference) and that all the information is georeferenced. On the other hand, the fact that it is used for tax purposes at municipal level, has as one of the main consequences that the quality of the information stored is very high.

One of the main drawbacks of Cadastre is that it is not linked to the Population Register, so in some cases it is not possible to determine which household of Padrón corresponds to each housing unit of Cadastre. On the other hand, Cadastre contains information on certain census variables like period of construction or useful floor space, but not all, as for example the variable tenure status of households.

During the following months an intense job of data-linkage between sources will be done, but any field-work operation is not foreseen. In the end, an integrated system where people from the Population Census are assigned to housing units belonging to the Housing Census should be available. The coherence between both products should be total.

In the same way as in the section of the population census, the situation for one of the most complicated variables of the next Housing Census is presented here schematically.

### 5.1. Tenure status of households

**Method and sources:** The main data sources are Tax Agency and Cadastre. In the Tax Agency all the declarants (around 65% of households) include the cadastral reference of their usual residence and also any other properties they own. On the other hand, Cadastre also contains information about the ownership of each housing unit.

**General issues:** Due to the fact that the proposed data sources are not totally exhaustive, it is necessary to carry out imputation based on a survey. Including certain improvements (for example information about rented dwellings that is stored in the declarations of the Tax Agency) in the data sources used would help to increase the quality of this variable.

## 6. Other elements

### 6.1. Inclusion of other variables

The use of administrative sources opens the door to the incorporation of new variables, either because of the availability of new sources, or because of the possibility to exploit them in another way.

A traditional Population and Housing Census uses the questionnaire as the instrument for collecting information. In this situation, it is very important to design a short and easy to understand questionnaire in order to achieve a good quality of response. However, in the

case of a register-based Census this limitation disappears; therefore, it can cover concepts that are difficult to include in a questionnaire or demanded by users. Some examples of the variables that could be incorporated in the 2021 Spanish Census are:

- Information about property of vehicles, from National department of traffic
- Fertility, number of children born in the previous years
- Data on other owned dwellings, based on Tax Agency information and Cadastre

On the other hand, one of the strengths of a register-based census, which has not been sufficiently exploited yet, is the possibility of incorporating new context variables, that do not refer to a concrete individual (both because of legal restrictions or because of quality issues for individual data), but to the average or total of his/her enumeration area. Some examples of this type of variables that could be incorporated are: income level, electrical consumption or presence of green areas.

## 6.2. *The survey*

Maybe one of the main points of improvement of the next Census is that some traditional variables cannot be produced from registers because they do not exist. The lack of certain information in relation to commuting (not mandatory but highly demanded) is an important weakness. Furthermore the need for more demographic characteristics like second generation of immigrants, knowledge of different languages, more reliable information about household members and better information about households and buildings is another handicap.

For all these reasons, and in order to complete the Census project, we are planning to conduct an ad-hoc survey that would target about 200,000 households (1%) and would answer these and other features (in a draft version: 12 questions about the household, 13 about the dwelling, 24 for each person and another 12 questions for people aged 16 or more). The editing and imputation of the information will be done through a software developed by INE and based on the Fellegi-Holt methodology.

The complete information about all the pieces that are part of the project can be consulted during the next years in [5].

## 7. Conclusions

A detailed analysis of the construction of variables for the 2016-Pilot allows us to conclude that it is already almost a complete Population Census. If we consider the planned improvements that it will be included in the future, all the required variables would be available. Analysis performed by INE indicates that conducting a register-based population census will provide census results with the same or higher quality compared to using traditional method. Spain would become one of the most populated countries (perhaps the largest) in the world with a register-based Census.

Regarding the Housing Census, the situation is not so conclusive but encouraging. The tasks for building the framework are still ongoing but the works done recently in certain provinces, where the percentage of data-linkage has been very satisfactory (in average more than 95%), allows us to state that a complete Housing Census based on registers will be also available with the same or higher quality than using traditional methods.

## References

- [1] Vega J, Argüeso A, Perez M. The 2021 population and housing Census in Spain: challenges and findings. ISI 2017, Marrakech (Morocco), 2017.
- [2] Register-based statistics in the Nordic countries. Review of the best practices with focus on population and social statistics. United Nations, New York and Geneva (2007). [http://www.unece.org/fileadmin/DAM/stats/publications/Register\\_based\\_statistics\\_in\\_Nordic\\_countries.pdf](http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf).
- [3] Conference of European Statisticians. Recommendations for the 2020 Censuses of Population and Housing. United Nations, New York and Geneva (2015). [http://www.unece.org/fileadmin/DAM/stats/publications/2015/ECECES41\\_EN.pdf](http://www.unece.org/fileadmin/DAM/stats/publications/2015/ECECES41_EN.pdf).
- [4] Vega J, Argüeso A. Spain 2021. Why will this Census have more quality than the previous one? European Conference on Quality in Official Statistics Q2016, Madrid (Spain), 2016.
- [5] 2021 Population and Housing Census. INE (Spain) <http://www.ine.es/censos2021/>.