United Nations ECE/CES/GE.41/2016/6

Economic and Social Council

Distr.: General
14 July 2016

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Eighteenth Meeting
Geneva, 28 - 30 September 2016
Item 7 of the provisional agenda
Possible uses of new data sources (e.g. "Big Data") for censuses

Three examples of innovative data sources in 2021 Spanish Census

Note by the National Statistics Institute, Spain [1]

Summary

Spain plans to conduct the first register-based census of its history in 2021. The main source of data that will make this possible is the population register (PADRON) that has been running since 1996.

Many data sources have been investigated during the last years in order to succeed in our purposes. Among all, this paper will concentrate on the possibilities which arise with the use of three modern data sources at very detailed geographical level: tax-data, power consumption and mobile phone information.

A description of each one of these data sources and real examples of information which can be obtained will also be included.

Keywords: Register-based census; data sources.

[1] Prepared by Jorge L. Vega,. Antonio Argüeso and Carmen Teijeiro.

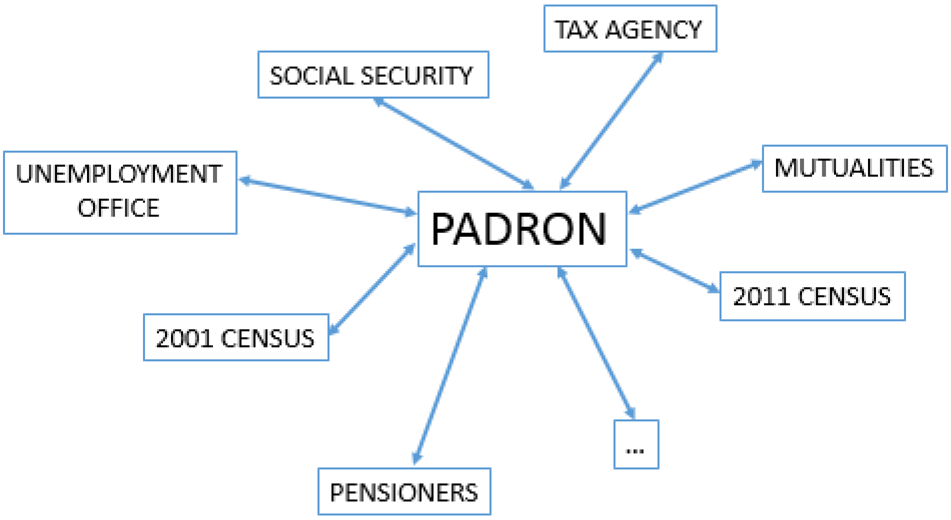
I. Introduction

1. Following the methodological evolution of previous editions[2], in which an increasingly intensive use of administrative data has been made, Spain expects to undertake a mainly register-based Census in 2021.

2. In order to successfully complete a project of this kind it is vital to have a good population register for linking with other data sources. In this sense, Spain is in a very good position, because it has available one of the world's best population registries (the "Padrón") which has been successfully working since 1996. It is subjected to constant quality improvements by all the agents involved.

3. Padron, as well as including numerical identifiers to facilitate links with other data sources, it also includes full names and dates of birth which are very useful both for establishing links when the numerical codes do not produce a satisfactory result and also for obtaining census information directly (e.g. for establishing family relationships).

4. Spain foresees linking multiple data sources with the Padron to obtain the different census variables: Social security register, unemployed people register, pensioners registers, contracts register, civil servants register, educational level register, foreigners register, vital statistics bulletins, previous census information, etc.



5. Internal research about available data sources carried out that in some cases, access to information with considerable potential and of very interesting content is also available, but its publication is not possible at personal level. The reasons may vary, from the safeguarding of confidentiality to difficulties in linking information at microdata level (but because of this reason, this does not mean that the data source is not useful for Census purposes).

6. In this paper we will see how these two theoretical limitations, far from excluding the use of these data sources, achieve a final result far richer than traditional Censuses. Three specific examples that are expected to be used in Spain in the run up to the next Census will be shown.

[2] An exhaustive Census was carried out in 2001 using population register information. In 2011 the use of registers increased and a field operation reaching 8% of the population was carried out.

II. Source examples

A. Level of income

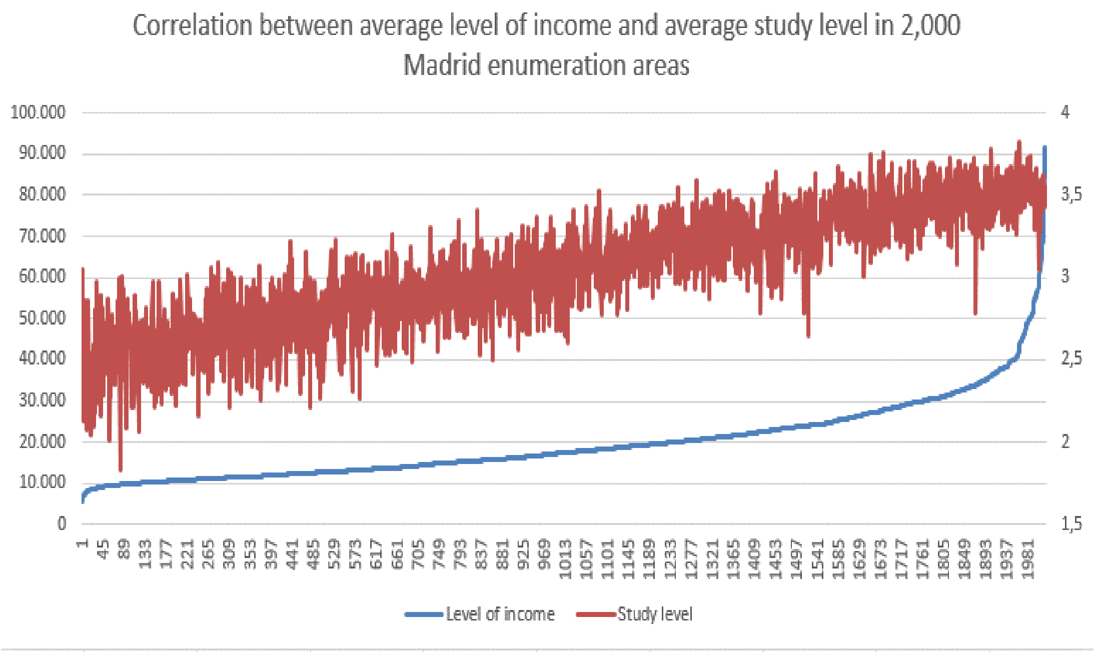
7. In Spain, the statistical collaboration between institutions is particularly good. A clear example of how this works is in the various exchanges between the INE and the Tax Office that allow the INE to have income information available at enumeration area[3] level (other statistics could also be available such as the percentiles of income level or the percentage of people living below the poverty threshold of the section).

8. For confidentiality reasons it is not possible to provide people's income levels individually, but the fact that we have this information at enumeration area level in no way invalidates the use of this source.

9. This information at enumeration area level is very rich. If we relate it to other census variables it opens the door to a number of possibilities that will surely be of great interest both to specialist researchers and to the general public.

10. For example, it can be seen whether or not there is a correlation between the income level of a section and the educational level of the people living in it. Do the sections with lower unemployment have a higher income level? Is there a connection between the income level of a section and the average household size? And with the average age of the section?

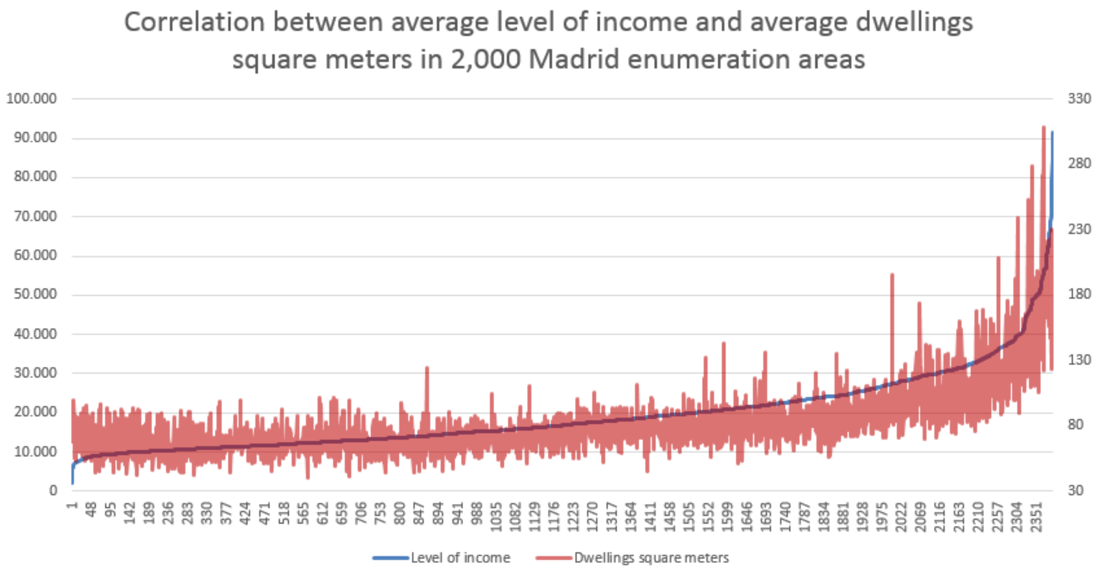
Figure 1



11. Correlation between the average level of income level (left side) and the average study level (right side) in 2,000 sections of the Madrid city area. Enumeration areas are in ascending order according to their average level of income.

[3] Spain, due to electoral reasons, is divided into 36,000 homogenous enumeration areas.

Figure 2



12. Correlation between the average level of income level (left side) and the average dwellings square meters (right side) in 2,000 sections of the Madrid city area. Enumeration areas are in ascending order according to their average level of income.

13. We are confident that making all this information available for the first time in one place to all users would be welcome by our users, not just for all the analyses that could be carried out but also for the depth of detail (both at a geographical level and regarding the number of variables) that could be reached.

B. Electricity consumption

14. One of the most representative characteristics of the Census is that it provides information on the types of dwellings (occupied, seasonal and empty) located throughout the entire country. To get this information the census enumerators carry out meticulous fieldwork which undoubtedly has a subjective element.

15. Among the different data sources that are being researched in the run up to the next Census, is the electricity consumption. This information[4], suitably treated, will allow us to obtain a new classification of households depending on consumption levels.

16. The strategy to be followed involves, as a first step, coordinating the territorial information from Cadaster, previous Censuses and addresses from the Population register. In this way we can produce a housing framework and have available both the amount of dwellings (VT) in a region and their location.

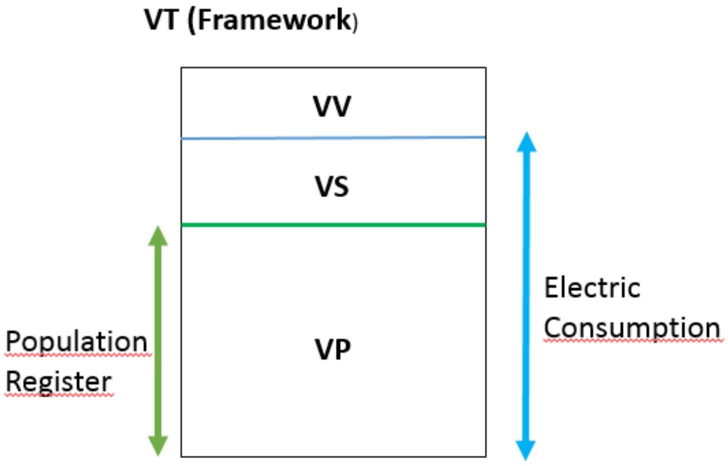
17. The Population register will allow us to determine which dwellings within the framework we classify as occupied dwellings (VP). Separately, the electricity consumption information[5] will be used to determine which dwellings have been occupied for all or at

[4] The electricity companies are obliged to submit information to the Tax Office showing the amount of electricity consumed at a given address in one year.

[5] In reality, electricity Consumption also provides information on some of the empty dwellings (those that have an electricity meter), but we can filter them by establishing a consumption threshold.

least part of the year (VP+VS). Finally, using the differences it is possible to obtain the empty dwellings (VV) since the equation VT = VP + VS + VV must be met.

18. The following graphic summarizes this:



19. Although at first sight this theory seems straightforward, linking data with territorial information is always complicated and the results are less satisfactory than those obtained from cross-referencing information on people.

20. So far, some tests have been carried out and the comparisons made between the information from the 2011 Census and that from electricity consumption have proved quite satisfactory with those obtained from the latter seeming, if anything, more credible.

21. It is not easy to use electricity consumption information at individual level because of the difficulties during the linkage process but this does not invalidate to use this information to provide some indicators in enumeration areas.

22. In principle the intention is to calculate new indicators at inframunicipal level, such as the percentage of dwellings with a consumption at or lower than the percentile 10, the percentage of homes with consumption between the percentile 10 and 25, etc.

C. Mobile telephones

23. The information related to mobility variables (place of work/study, linked population, etc.) has been the most requested in the latest Census. There are multiple uses for this information: organizing the communications that exist between different localities, knowing the real weight of population that a geographical area supports, determining "job markets", etc.

24. Mobile telephone companies have at their disposal a large quantity of information of this type that can prove very interesting for the purposes of the next Census, because every time we make a call, send a message or connect to the network we leave a trace on the aerial nearest to where we are located.

25. If we consider the three examples explained in this paper, this is the most innovative. Nowadays, we are trying to obtain mobile telephones information but, up to now, there is no guarantee in this issue. For the moment, all the information we have is based on a pilot project carried out with only one company.

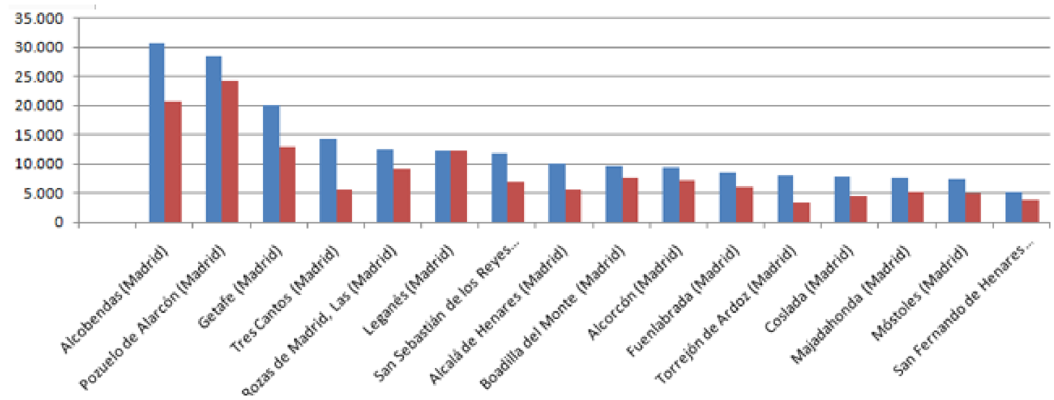
26. In the first place, using information supplied by the operators, the INE will determine in which enumeration area a person lives and also which is the target enumeration area (for work or study reasons) of each mobile telephone.

27. These are some specific products at enumeration area section level that could be construed using this information:

- Product 1: Matrix of daily movements between enumeration areas for a specific day of the year (e.g. one day in November). The box "row A-column B" of the matrix would show the number of people who, on average, move from area A to area B for a number of hours above a certain threshold.

Figure 3

Comparison of destination for population living in Madrid. 2011 Census vs mobile telephone data (pilot study)



- Product 2: Matrix of population in enumeration areas at different times in a given day (e.g. one day in November). This product can provide something that has already been produced by other organizations and that is very much in demand by users i.e. figures for the daytime and night-time population of an area.

- Product 3: Matrix of linked population for a given amount of days under observation. The box “row A-column B" will contain the number of people who live in cell A but can be found for a certain number of hours in cell B. The idea is to calculate this matrix for about 10 specific days of the year: 5 days in July/August and another 5 days in the rest of the year (Easter Week, Christmas, etc.)

28. One finer point that has to be considered with these products is that the procedure as it is described above measures movements of mobile telephones, but some correction factor could be applied to make it a measure of population.

III. Conclusions

29. The fact that both the population register and other sources containing basic census data have improved in quality and coverage in recent years will allow Spain to carry out a register-based Census in 2021.

30. The philosophy of using administrative registers allows much greater depth than traditional census questionnaires where the list of census variables was limited by the physical space on the question paper. Another advantage of this strategy is the possibility of publishing information periodically as considered appropriate, and the ability to tailor it to requirements at any time.

31. With this new method, it is possible that even those data sources with missing data linkage at microdata level or those that cannot be published for individuals due to confidentiality reasons but which offer interesting content, could also form part of the Census content and thus enrich the quantity of information made available to users.