United Nations

Working paper 4

Economic and Social Council

Distr.: General
18 April 2012

English only

Economic Commission for Europe

Conference of European Statisticians

UNECE-Eurostat Expert Group Meeting on Censuses Using Registers

Geneva, 22-23 May 2012
Session 1: Experiences with the use of registers in the censuses

A general approach to the importance and use of registers in the Spanish Census

Note by the National Statistics Institute of Spain[1]

Summary

2011 Spanish Census was based on a mixed approach of a big sample survey of 12% of the population (3M dwellings and 5.8M people) and use of administrative registers.

From a statistical point of view, a model like this sets outs several challenges. First of all, it should be carried out a study in depth of the different registers and features like completeness, accuracy and coverage should be analysed. On the other hand, we should define different strategies in order to link information from the survey with information from registers. Furthermore, other aspects like inconsistencies between different data sources and how detailed is the information we disseminate should be considered.

In this paper we will try to summarize all these points, stressing the importance of the strong and weak points of this methodology from our own experience.

[1] Prepared by Carmen Teijeiro Breijo, Jorge L. Vega Valle and Miguel A. Martínez Vidal.

I. Registers in Spain

1. As opposed to other countries where it exists one organization that gets together all the registers that exist in one country, in Spain there are several organizations (ministries, institutions, councils…) that administer different types of registers.

2. If we focus on the different sources that we used in 2011 Spanish Census we should start mentioning two of the most important ones: the population administrative register (PADRON) and the 2001 Population, Housing and Building Census.

3. PADRON is an administrative register, with its own legal regulation on how it should be managed. The operational part is based in a central database located within INE and every month all the municipalities (which are responsible for managing the registration of their inhabitants) send the changes occurred in their local registers. In addition INE, in a joint work with municipalities and other local authorities, try to improve PADRON’s quality deregistering duplicates or communicating changes of residence between municipalities.

4. On the other hand, one of the main limitations of PADRON is the lack of demographic topics we can find. Only very basic information like name, family name, date of birth, place of birth and citizenship make up the population register[2].

5. Nevertheless PADRON has some problems if we want to use it as a direct statistical data source. It is an administrative register and registration in it gives people fundamental rights like voting, access to health and education services, etc. So we can not decide to register or de-register a person in PADRON taking into account only statistical criteria. Because of this and in order to comply with the law, it can take several months to deregister a person who has moved abroad without saying anything to the municipality.

6. PADRON is an essential register for obtaining demographic features about people but although it contains territorial information of where people are living, this information is not organized from a computing point of view. In other words, PADRON is a register of people but not of dwellings. Because of this, the frame of buildings obtained from 2001 Census was another important pillar for 2011 project.

7. One of the most important outputs of 2001 Population, Housing and Building Census was that each person had a fixed code assigned to the dwelling they were living, so the linkage between the demographic and territorial information was possible.

8. We took advantage from that information obtained from 2001 Census and since them, we have been updating our dwelling frame using PADRON’s information with all the people that arrive or move to a dwelling whose code does not exist in our register. For 2011 Census we started from a frame that contained:

- Information from dwellings that existed in 2001 Population and Housing Census
- Information from dwellings that were occupied at least once between 2001 and 2011

9. Nevertheless our dwelling frame must be updated in order to include all those dwellings that have been built since 2001 and have never been occupied since them.

[2] It also contains a topic related with level of studies, but this information is not very accurate.

Because of this reason and also for collecting all the geographical coordinates of the buildings we carried out an almost exhaustive[3] 2011 Building Census.

10. Other important registers in 2011 Census project are national identity cards and resident card register (they contain important information to check the inhabitant identification), vital statistic bulletins (they allow us to check information from births, deaths and marriages), tax agency and social security.

11. However, registers should fulfil several conditions to be useful to Census purposes. Information from those registers should be updated frequently, coverage of registers should be good, they should refer to the same concepts that Census ask about, the breakdown of the information used in registers should be enough for Census dissemination and it should be possible to access to that register from a technical and legal point of view.

12. In the next table, we sum up what kind of registers have been useful for preparing 2011 Spanish Census:

Name of the register	**Responsible for administration**
Population register	INE
Census 2001	INE
National identity cards and resident cards	Home Office
Cadastre: Dwelling register	Treasury Department
Tax Agency	Treasury Department
Social Security	Ministry of Labour and Social Affairs
Vital statistic bulletins: Births, deaths and marriages	INE

Future perspective

13. Although INE has some experience with the usage of registers, there are still some difficulties in order to carry out a complete register based Census. Nevertheless some information from different registers could be used with a little amount of work both from INE and the body responsible for managing each register.

14. For example, information about consumption of electricity could be very useful to decide if we have to consider a dwelling as occupied the whole year or as a seasonal or empty one. Furthermore information from other registers should be improved and analysed more deeply in order to make the most of those registers. In this situation we can find different registers that have very rich information like registers related to educational attainment (both at university level or at school level) or registers that contain information about employed and unemployed people or also registers administrated by mutual insurance companies or social security.

[3] In a 20% of enumeration areas the information was obtained directly from our updated dwelling frame and coordinates of those dwellings were obtained from Cadastre’s information.

Name of the register	Responsible for administration
Consumption of electricity	Electrical companies
Education attainment (university)	Ministry of Education
Education attainment (school)	Ministry of Education
Employed/Unemployed people	Ministry of Labour and Social Affairs

II. Spanish methodology: the combination approach

14. Censuses are the biggest challenge that official statistical offices have to deal with. In fact, they are not only a data collection question but a logistic operation. To illustrate this, 2001 Spanish Census budget was more than 200 million of euros (today it would be more than 300 million), involved more than 40.000 people including enumerators, supervisors, etc. and more than 70.000 contracts were signed in order to make replacements when people of the staff decided not to continue with the job.

15. For the first time ever it exists an European Regulation according to the population and housing Census. This regulation provides the description of different sources of data that will be acceptable in the Census.

16. Population figure is one of the most remarkable result of a Census. Furthermore, other types of information are also very important depending on the interest of different users: demographic and social variables, dwellings and buildings information and many others.

17. Regarding to the first point, counting population, we must not do this without taking into account only the population administrative register. As we mentioned before PADRON should not be used without checking the information it contains. For this reason PADRON data are linked with other administrative sources, for example Social Security and Tax Agency in order to accumulate proofs of real residence in the country. With this approach we found that more of 95% of population is registered correctly and at the same time we have some information (based on sex, place and date of birth, place of residence and citizenship) to characterize people that are not registered properly.

18. So we decided not to dedicate Census field work to count the people that we have no doubts they were in the country and we concentrated our efforts in the people which had a doubtful residence in the country. This is the main reason why we did not carry out an exhaustive enumeration of people.

19. Moreover, there are also some administrative registers that can provide us information about dwellings and buildings. As we mentioned before PADRON only contains information of occupied dwellings and Cadastre has such a different structure of data, that our attempts of linking these two registers did not reach acceptable results. For this reason we decided to conduct an exhaustive enumeration of buildings and dwellings.

20. Taking into account all these considerations 2011 Spanish Census it is based on three main pillars:

- A preCensus file as a result of a linkage process between different administrative registers. It will be used to:
    - •Decide which population will be counted (or not).

  - •Characterize those groups of records that we do not know if should be counted or not.
  - •Be the initial frame of the Building Census.
  - •Provide data about, at least, some basic population topics: sex, age, citizenship and place of birth.
- A big survey (around 12% of population) to collect data about population and housing topics. This operation will be useful for:
  - •Estimating the size of those groups of records that we do not know if should be counted or not.
  - •Providing information about topics that are not included in the preCensus file.
- An exhaustive enumeration of buildings, that will be used:
  - •To obtain a complete frame of buildings (a basic necessity for future surveys).
  - •To get geographical coordinates of buildings.

21. 2011 Census information was collected using several channels: internet, mail and CAPI. First of all, a letter was sent to the occupied dwellings giving them only the opportunity to answer by internet. For those that did not use this option a second letter was sent including a paper questionnaire and both options were given: answering by internet or sending the questionnaire by mail. Finally those that did not answer were visited and collected the data using enumerators with hand held computers.

Advantages and disadvantages of this methodology

22. If we compare the methodology chosen by Spain in 2011 with a traditional Census, we can find several advantages.

23. The amount of people needed for collecting all the data (building, dwelling and population Census) was less than 5.000 people. The less amount of field staff we need, the easier it is for us to find that people and the more qualified they are. Moreover, the control of the total staff that took part of the project was much easier than in previous experiences and the training process was more effective.

24. The multichannel approach maximizes the possibilities of a household to answer to the Census. Internet is especially good with young people and also with people that want to collaborate whenever they want and wherever they are. Paper fits better, obviously, with people that do not have Internet connection at home and with people that are not fond of new technologies. Finally, enumerators with hand-held devices are suitable for those dwellings that are more difficult to collaborate and it ensures the total coverage of the field work.

25. Usage of registers allows an improvement in the quality of the Census information. Sometimes registers will be used instead of asking questions to the people. In other occasions registers will be used to confirm user’s response and in another ones register will be used in edit and imputation phases which provides more accurate information.

26. Another point that should be considered is that from an economical point of view this model is much less expensive than a traditional one. In an operation like this, almost 70% of the expenses are related to enumerators, so a reduction in the amount of enumerators means less money.

27. On the other hand another strong point of the 2011 Spanish Census is the collection of the building coordinates. This information is very important for different reasons:

- It will be possible to disseminate information in more visual formats like personalized maps.
- Users can ask questions to the system regarding not only to administrative boundaries but also with geographical points. It will be possible to paint an irregular polygon (that cover parts of some administrative areas) and the system will solve the Census query according to that polygon.

28. However a model like this is not perfect and it entails several inconveniences. Below, a short list of these problems are mentioned.

29. It is not easy to deal with the multichannel approach (Internet, paper, enumerators). It is possible to receive different answers from the same dwelling through several channels. And to make it even more difficult these answers may not be consistent among them. An algorithm that decides with answer is the most suitable should be implemented.

30. Furthermore, because of the sampling error this model has difficulties to solve detailed requests (that involve several Census topics) at a very low geographical level. Results would be so inaccurate that its publication would be counterproductive.

31. Additionally, the amount of time dedicated for collecting data was quite long (from September 2011 until April 2012) and the date of reference was in the middle of that period. It was necessary to carry out different adjustments in order to have information coherent with the date of reference.

32. Finally, a model like this had never been developed by our country and from a methodological point of view it has been a challenge. Sampling techniques and calibration adjustment will be used in order to produce consistent and accurate Census data.

III. Linkage between 2011 Census information and registers information

33. First of all, before starting with the linkage between 2011 Census information and information from registers, it should be mentioned how information from registers is linked.

34. Because of its quality, PADRON plays the role of the main backbone and information from other different registers will be linked to PADRON and will help us to check it. PADRON contains identification data from each person, but this information is stored in three different topics instead than one: number of national identification document (DNI), number of passport and number of residence card. Each person in PADRON will contain personal data in one of those three topics.

35. People in PADRON will have information about its personal identification number. Although it is not very frequent, sometimes we can find two or more people with the same or with a wrong personal identification number or even people without a personal identification number. In order to solve this problems and contrast PADRON information, a linkage between information from PADRON and information from Home Office[4] should be done.

[4] Home Office is the organization that assigns national identity cards, passports or residence cards to people.

36. In order to accept a linkage between PADRON and information from Home Office, it is demanded (on the first step) an exact coincidence of personal identification, name, family name, sex, date and place of birth, between these two sources. For those records where this coincidence is not perfect (maybe information from one of the two sources is wrong) probabilistic algorithms with an internal threshold or control parameter that measure similarities among people are executed.

37. Same type of processes are applied to other administrative sources in order to be sure that we have good quality identification numbers in all the registers involved and to increase the percentage of records linked.

38. On the other hand in 2011 Census questionnaire we do not ask people about their personal identification number (number of Spanish document, passport or residence card) so linkage between 2011 Census information and information from registers has been carried out on a different way.

Different techniques

39. First of all, for all users that respond to the Census information by internet or with a Census enumerator, an special software that checks if one person exists in our registers in that dwelling has been developed. This algorithm achieves percentages of linkage of almost 80% in most Spanish regions. A person may not be linked because of different reasons: maybe because the dwelling where the person answers to the Census is different to the place where the person is registered, maybe because the personal data that the person fills in the internet Census forms is slightly different to the information we have in our registers or maybe because between the date we generate the frame with the information and the date the person answers to the Census that person has moved to a different dwelling.

40. If a person has answered by Internet or with a Census enumerator but has not been linked on a first step or if a person has answered by paper, different techniques are implemented in order to increase the amount of linkages between Census information and information from registers.

41. One of the techniques consists on making a comparison between Census information and information from registers for different geographical levels: same dwelling, same enumeration area or same district. In order to accept a linkage as successful we demand to have equal or similar information in all the topics we compare: name, surname, date of birth, nationality...If one of those topics is different then the linkage is not accepted. Because a linkage can be accepted having similar, but not exact information in both sources (Census and registers) this technique takes up a lot of time.

42. Another technique that has been carried out consists on building a “super code” with some topics from the Census and registers (code of name + code of surname + date of birth + code of place of birth +...). If there is a perfect coincidence[5] between information from Census and information from registers then the linkage is accepted. This strategy has the advantage that it is very easy to implement and quite fast regarding to time computing and because of that this reason it is applied to all geographical levels including the whole country.

43. To conclude, for those registers that have not been linked yet, a probabilistic approach based on distances between two text fields that compares information from 2011 Census and information from PADRON and other registers has been implemented.

[5] According to this feature this technique is different to the previous one.

44. Results are very encouraging and although the probabilistic approach was not been executed yet, we are able to link around 97% of the registers that responded to the Census questionnaire by Internet and 92% of those that responded by paper.

IV. How we obtain the population figure

45. As we have mentioned before the preCensus file contains records from PADRON linked with other administrative registers. In order to obtain the population figure we will assign to each record a counting factor.

46. First of all, those records that exist in PADRON and have also been found in other registers, should be counted without any doubt. In this case we assign to these records a counting factor of 1.

47. Secondly if a record has been linked with some of the register of deaths provided by the Civil Register, we assign to that register a counting factor of 0.

48. Finally, if a record has not been linked with any of the administrative registers we should check if that record is marked with a warning[6]. In case the register is marked with a warning it will have a counting factor of "X", where X is a value greater than 0, not necessarily integer, that should be calculated using survey results. Otherwise, the register will have a counting factor of 1.

49. The aggregation of the counting factors will give us the population figure of a geographical area and results of the sample will be calibrated to this weighted frame.

Methodology for calculating X

50. The simplest way to calculate "X" would be to estimate the amount of "warning" records using the sample data. But it is impossible to do that because these records are defined by its administrative situation and this information, obviously, is not collected during the Census.

51. The strategy will consist on defining clusters based on social and demographic variables (for example, UE-citizens aged over 65 living in the Mediterranean region). In each cluster there will be some "warning" records and records with a counting factor of 1. Each "warning" record must only exist in one cluster. The amount of population included in clusters should not be very big (around 5% of total population) and there should be as many clusters as we need but all of them with small amount of population.

52. All the "warning" records included in a cluster will have the same counting factor "X". "X" will be calculated this way:

The estimated size of the cluster (obtained from the sample) minus the records included in the cluster with a counting factor of 1 divided by the number of "warning" records that belong to that cluster.

[6] This is a simplistic approximation of some of the different rules that assign a warning to a register:
-If a record belongs to a Spanish person older than 13 years old and has not a proper id number
-If a record has an identification number but it was not found in Police databases
-If a record that belongs to a foreigner remains unchanged for a long period of time in PADRON.

53. The most simple way to calculate this counting factor would be to estimate directly the size of the “warning” records from the sample results . But this is imposible because these “warning” records are defined in terms of their administrative situation and these are characterisitics that, obviously, are not investigated in a Census. The advantage is that these records are between 2 and 3% of total population.

54. So we need an alternative strategy. The first step is to create clusters defined in demographic terms that include the “warning” records. One example could be: “nationals of eurpean countries aged more than 65 years living in the mediterranean coast. Of course, in such cluster there will be “warning” records and records with CF=1. We can estimate the size of these clusters with data from the sample because they are defined by using variables investigated in Census.

55. We create so many cluster as we need, avoiding to create clusters with big population inside. We prefer to create more clusters but small ones. The final definitions of clusters could be made even after the data collection.

56. The CF for “warning” records of each cluster is given by the estimated size of the cluster minus the number of records included in the cluster with CF=1 (we are not looking for a CF for those records that have CF=1, only for that that have CF=X) divided by the initial number of “warning” records. This number could be less than 1 (if the cluster is over-represented in Padron) equal to 1 or greater than 1 (if the cluster is under-represented in Padron)

V. Conclusions

57. Spanish methodology has the advantage of making the most of the information contained in registers and disseminating detailed information based in a relatively big sample.

58. Moreover, it reduces the budget and the staff involved and it improves the quality of the field work.

59. Furthermore, this alternative saves the differences between administrative concepts and statistical variables.

60. On the other hand, it reduces the detail of information that can be disseminated if we compare it with previous Censuses. And of course, this methodology is completely new for Spain, so some unexpected difficulties must be faced.