United Nations ECE/CES/GE.41/2013/21

Economic and Social Council

Distr.: General
3 September 2013

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Fifteenth Meeting
Geneva, 30 September – 3 October 2013
Item 2 of the provisional agenda
Census methodology

How to deal with two different sources: registers and survey?[1]

Note by the National Statistics Institute, Spain

Summary

The 2011 Spanish population and housing Census has been carried out combining two main sources: registers and a survey. Registers have several advantages such as completeness (most of the times) or being used for a long period of time with continuous improvements and updates, but they also have some problems like not being able to detect some situations or having only very basic variables. On the other hand, the survey has the advantage of having as many variables as needed and sometimes information is more updated in the survey than in registers. However other problems like universality of information, non response or invalid values have to be solved. If a user makes a request that deals only with register variables, there will be no problems of inconsistencies and the information will be obtained directly from the register. On the other hand, if a user makes a request that deals both with register and survey variables, calibration procedures should be applied in order to minimize differences between these two sources. Of course, it will not be possible to calibrate all the breakdowns of the different variables and some differences, that users will have to accept, will show up.

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

I. General ideas about the 2011 Spanish Census

1. 2011 Population Census combined these two elements:

2. *Registers.* The most important register used was the "Population Register", but registers from other sources, such as "Tax Agency", "Social Security" or "Vital Statistics" were also used.

3. In Spain there is a Population Register (PADRON), that has been operating for a long period of time. Since 1996, PADRON works on a continuous way with the collaboration of more than 8,000 municipalities, and checks, updates and improvements are incorporated every month.

4. The Population Register contains the list of all people registered as usual residents in every municipality, regardless their legal situation. PADRON data are, in general, very reliable and updated. It is the backbone of the whole census operation.

5. In order to assess its quality and to provide more accurate population figures it has been linked with other registers mentioned before. Information from other registers was used only to "assess the likelihood of being resident". For example, a person that exists in all the different registers has a higher likelihood value than a person that only exists in the Population Register.

6. PADRON contains several core variables from a demographical point of view like: sex, date of birth (age), place of previous residence, place of birth and citizenship for all the population.

7. Concerning the Census point of view, these features are not enough for the Census purposes, so the necessity of using other kind of sources like the Survey is absolutely clear.

8. Apart from that, PADRON has difficulties to count correctly some situations. For example, if a person moves abroad the country, it takes several months to detect that situation and to delete that information from the population register. This point will try to be solved using the survey.

9. *Survey.* A big sample survey with information of 4,108,000 people (aprox. 9% population of Spain).

10. The survey contains information from all the Census variables needed (legal marital status, migration, education, economy, place of work/study, relation between household members…).

11. Survey information is not always perfect. Some types of problems like invalid values, inconsistencies among different variables or non-response are quite frequent in the survey and have to be solved using editing and imputation (both deterministic and random) techniques. These problems, take different and specific types of procedures to be solved, sometimes require long periods of computing.

12. Both Registers and the Survey have advantages and disadvantages, but the combination of these two sources allows to produce a good quality product and more reliable Census figures. On the other hand, and as opposed to the previous Census, this method will not produce such detailed results. If we consider a request that deals with an area of the same population in both Censuses, the amount of variables that could be disseminated or the level of detail of the breakdowns involved in this Census cannot be so huge than in the previous one. Nevertheless, the collection of geographic coordinates for all the buildings will make it possible to disseminate Census information not only for administrative division, but also from a more natural point of view. Users will be able to

make their own requests drawing the geographical area in which they have interest on (for example: "how many people live within 200 meters from a given point").

II. Procedure used for obtaining population figures and detailed Census information

13. As stated before, the PADRON is almost enough to provide good population figures but it was crossed with many other sources to gather more information of "likelihood of residence". As a result of this process the whole population is divided into three categories depending on the evidences we have to count each person or not according to the information we have in other types of registers (Social Security, Tax Collection Agency, Vital Statistics Bulletins...):

(a) Group 1: "Sure" population: 46,372,000 records (97.2% of the PADRON). There is enough evidence to consider them "without any doubt". These records will have a "count factor" of 1.

(b) Group 2: Errors: 292,000 records (0.6% of the PADRON). There is enough evidence to delete these records because they are wrongly included. Most of these records exist in the Population Register but do not count as active. These records will have a "count factor" of 0.

(c) Group 3: Doubtful population: 1,046,000 records (2.2% of the PADRON). 15 different criteria were defined in order to consider a record as doubtful. If a record belongs to this group, there is not enough evidence to count it or not. A huge amount of the records that belong to this group (87%) are foreigners. These records will have a count factor higher than 0, but its concrete value will be assigned using the information from the Census survey.

14. In this process also some 38,000 births, still not included in PADRON, but available from vital statistics are added with a count factor equal to one. Eventually, the file constructed from PADRON adding those count factors and births is named WEIGHTED CENSUS FILE (WCF). For a 2.2% of records this count factor is still to be assigned.

15. The final size of the WCF was 47,711,000 registers.

16. In order to assign the concrete count factor, those 1,046,000 doubtful records (and the rest of the records of the WCF) were divided on 724 classes[2] (each one of at least 1,000 doubtful records) depending on demographical variables like age, place of residence, citizenship...

17. The proportion of sure records (Pi) is estimated[3] through the proportion of sure records in the survey ($\hat{p}_i$).

[2] For example, one class could be German people between 70 and 74 years old (both included) living in Alicante. Another example could be Spanish people between 0 and 4 years old (both included) living in Madrid region.

[3] For more details about how the counting factors are calculated, take a look to "A population census based on registers and a 10% survey. Methodological challenges and conclusions" (Argüeso), ISI – Hong Kong, August 2013.

18. Proportion of sure records in the population:

$$P_i = \frac{S_i}{S_i + D_i} = \frac{S_i(group1)}{T_i(group1 + group3)}$$

where:

- S_i is the amount of sure records in the population that belong to class i.

- D_i is the amount of doubtful records in the population that belong to class i

19. Proportion of sure records in the survey:

$$\hat{p}_i = \frac{\hat{s}_i}{\hat{s}_i + \hat{d}_i} = \frac{\hat{s}_i}{\hat{t}_i}$$

where:

- $\hat{s}_i$ is the amount of sure records in the survey that belong to class i.

- $\hat{d}_i$ is the amount of doubtful records in the survey that belong to class i

20. We will assign the same count factor CF_i to all the doubtful records that belong to class i. The value of CF_i can be calculated as:

$$CF_i = \frac{\frac{\hat{d}_i}{\hat{s}_i}}{\frac{D_i}{S_i}}$$

where S_i and D_i are obtained from the WCF and $\hat{s}_i$ and $\hat{d}_i$ are obtained from the survey.

21. As mentioned above, the amount of doubtful records was 1,046,000. The average count factor for all the doubtful records was 0.424, and these doubtful records were counted as 443,000 people. So, the final population figure was 46,815,916 (46,372,000 from group 1 + 443,000 from group 3)

$$Census\_population = \sum_{i=1}^{47,711,000} Count\_factor_i$$

22. Information from the WCF is used to disseminate the population figure of each region and basic Census information, with "PADRON variables" (place of residence, sex, age, place of birth and citizenship) and only those variables. The procedure to obtain every given subpopulation is always the same: the sum of counting factors of those records in the WCF.

23. On the other hand, information not included in the WCF or the combination of these variables with those included in the WCF is obtained[4] from the Census survey. Each one of

[4] For example, information like population by sex and age will be obtained from the WCF, while population by age and occupation will be obtained from the Census survey.

the 4,108,567 records of the Census survey has a sampling weight[5] and, the procedure to calculate any subtotal is quite similar to the previous one but instead of adding counting factors, adding sampling weights.

24. Because its limited size the survey may not provide all the possible combinations of age and sex or the different types of citizenships (these are just two examples) that exist in a concrete geographic area.

25. Moreover, because of being an independent operation, some inconsistencies between the WCF and the survey will raise. For example values that do not exist in the WCF inside a concrete geographic area, could appear in the Census survey or vice versa (for example, people of a given citizenship).

26. In order to minimize differences between these two sources, calibration techniques have been used.

III. Calibration technique

27. In general, the calibration approach to estimation for finite populations consists of a computation of weights that incorporate specified auxiliary information and are restrained by calibration equations.

28. From the mathematical point of view, the problem can be seen as a system of J[6] equations and n[7] unknown quantities, where J is supposed to be much smaller than n, what it means an indeterminate system of linear equations, in other words, with infinite solutions. In order to find a solution to this problem, it can be looked at as an optimization problem and it can be solved using Lagrange multipliers method.

29. This procedure adjusts the sampling weights by multipliers known as calibration factors that make the estimates agree with known totals. The resulting weights[8] are called calibration weights or final estimation weights.

30. This calibration method has been developed by Deville, Särndal and Sautory (1993) in SAS statistical software with the name of CALMAR (CALage sur MARges), and it is possible to download it from National Statistics Institute of France (INSEE)[9].

31. Nowadays CALMAR is being used by several public statistical offices and private research centers all around the world. CALMAR has succeeded because of its simplicity and also because it is possible to use it both with quantitative or qualitative variables.

32. INE-Spain used the CALMAR program in order to make basic demographic figures from the WCF and the Census survey coherent at a given level. Inputs are:

(a) Information from the WCF plays the role of being the "population frame" with N elements, and the Census survey plays the role of being the sample with n elements.

[5] The population figure is 46,815,916 and the size of the survey is 4,108,567 records, so the average value of the sampling weight is a bit higher than 11. All the people living in the same dwelling have the same sampling weight value.

[6] *J* is the amount of auxiliary variables which have known population totals

[7] *n* is the size of the sample

[8] Inverse inclusion probability weights are, by definition, greater than or equal to unity. Calibrated weights, on the other hand, are not necessarily greater than or equal to unity, unless special care is taken in the computation to obtain this property.

[9] www.insee.fr

(b) Detailed information by sex, age, citizenship and place of residence from the WCF are our J auxiliary variables and their values are perfectly known.

(c) The sampling factors (weights) of the Census survey are the parameters that CALMAR program should adjust in order to reflect the same information (or at least as similar as possible) than the auxiliary variables in the WCF.

33. Calibration process is done for each one of the 8,116 Spanish municipalities (and also for the administrative divisions that contain the municipality, like NUTS2 or NUTS3 regions).

34. In the most populated municipalities an additional calibration adjustment by district (administrative division; depending on the size of the municipality, one municipality can be divided in one or more districts10) is carried out.

35. Table 1 contains the detail of the variables (sex, age and citizenship) that are involved in the calibration process depending on the size of the place of residence (district, LAU2, NUTS3 or NUTS2 regions)

Table 1. Variables involved in the calibration process by the size of the place of residence

		SEX(2)				*SEX(2)*		*Total calibration figures*[11]
Group	*Population*	*Only SEX*	*x AGE.L*[12]	*x AGE.M*[13]	*X AGE.H*[14]	*x COC.L*[15]	*x COC.M*[16]	
1	Less[17] than 50							1
2	51-200	2						3
3	201-500		6					7
4	501-2,000		6					7
5	2,001-10,000			16				17
6	10,001-20,000			16		4		21
7	20,001-50,000			16		4		21
8	50,001-100,000				32	4		37
9	100,001-250,000				32		10	43
10	More than 250,000				32		10	43

36. In order to obtain the most accurate figures, a two step calibration procedure is carried out:

37. Step 1: Information from the sample of every municipality is adjusted to the auxiliary variables that have been established within that group, according to the table above.

[10] For example the municipality of Madrid is divided into 21 different districts, Barcelona in 10, Sevilla in 11, Valencia in 19...

[11] Total population is always calibrated.

[12] AGE.L (3 different values): 0-15, 16-64, +64

[13] AGE.M (8 different values): 0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, +70

[14] AGE.H (16 different values): 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, +74

[15] COC.L (2 different values): Spanish, Not Spanish

[16] COC.M (5 different values): Four most frequent citizenships and a residual category for other citizenships

[17] Nothing else than total population is calibrated within this group.

38. Step 2: The total amount of municipalities that belong to one group[18] (according to its population) and the same NUTS3 region are adjusted to the highest level of calibration. This level of calibration (usually SEX x AGE.H and SEX x COC.M) is established by the most populated municipality in the NUTS3 region.

39. Because of this, the NUTS3 region is also calibrated to the same level than the most populated municipality in that NUTS3 region. Most of the NUTS3 regions will allow 43 different figures to be calibrated (SEX x AGE.H and SEX x COC.M), but there are some NUTS3 regions with less population, like for example Huesca, that will allow only 21 different figures to be calibrated (SEX x AGE.M and SEX x COC.L).

40. The application of calibration has to reach a trade-off with the number of variables that are involved in the process. If several variables are included, then the information that is not calibrated can produce "quite strange" results, because of the amount of constraints that should be met. Nevertheless, it is important to include as many variables as possible, in order to guarantee the highest coherence between the WCF and the Census survey. It is not an easy task, to reach the equilibrium point mentioned before.

A. Two examples of calibration

1. Example 1: Almazan (a municipality of Soria between 2,001 and 10,000 inhabitants)

41. Almazan contains 5,955 records (people) in the WCF (5,857 of those records are sure population – group 1, 47 are errors – group 2 and 51 are doubtful population – group 3) and 618 (10.4% of the WCF) records in the Census survey, so the population is the sum of count factors: 5,882.17.

42. According to table 1, 17 different figures should be calibrated in this municipality. It is checked that if a figure is calibrated, it gives the same result in the WCF than in the Census survey, but if it is not calibrated, differences appear. For example, in the WCF we can find information from 32 different citizenships and a total amount of foreign people of 708.04 while in the Census survey only 11 different citizenships and 656.08 foreign people are found (table 2).

Table 2. Example of calibration number 1: Almazan

	WCF	*Census survey*
Total population (Calibrated)	5,882.17	5,882.17
Male population (Calibrated)	3,016,09	3,016,09
Population between 51 and 55 years old (Not calibrated)	405.85	405.20
Foreign population (Not calibrated)	708.04	656.08

2. Example 2: Almeria (a NUTS3 Spanish region with more than 250,000 inhabitants)

43. Almeria contains 712,627 records in the WCF (677,045 of those records are sure population – group 1, 7,191 are errors – group 2 and 28,391 are doubtful population – group 3) and 54,981 (7.72% of the WCF) records in the Census survey. The population (the sum of count factors) is 688,736 inhabitants.

[18] There are some NUTS3 regions with very few municipalities that belong to one group. In this cases, an association of municipalities from different groups was carried out.

44. According to table 1, 43 different figures should be calibrated in this NUTS3 region. Just like in the previous example, those figures that are calibrated, are the same in WCF and in the Census survey, and in those that are not calibrated there exist some differences (table 3).

Table 3. Example of calibration number 2: Almeria

	WCF	*Census survey*
Total population (Calibrated)	688,735.62	688,735.62
Foreign population (Calibrated)	137,214.23	137,214.23
Population with 29 years old (Not calibrated)	11,728.62	11,675.26
Population from France (Not calibrated)	1,627.08	1,724.34

IV. Conclusions

45. The 2011 Spanish Census model tries to take advantage from the quality of existing Registers (availability of information for the whole population) and Survey properties (availability of as many variables as needed).

46. However, we should be able integrate both sources to provide a single set of figures in order to disseminate good quality data. Because of that, information from the Survey (in principle, with less quality) should, whenever possible, be calibrated to information from the Register (known information and with more quality). This is not always possible. And if we want to published detailed figures by nationality, for instance, we have to assume that these results provided by the WCF will not be the same as those provided by the survey except in those cases where figures are calibrated.

47. Total calibration is obviously not possible and an equilibrium with the figures that will be calibrated should be achieved. Information that is not calibrated will not be consistent between the two sources and it will be a challenge to explain our users on a transparent way the reason of those differences. Users will have to live with differences (especially if requests involve detailed breakdowns of basic demographic variables).