Seminar on Registers in Statistics - methodology and quality
21 - 23 May, 2007 Helsinki

Use of administrative information for the construction of the frame in the Spanish household surveys.

Juana Porras, Montserrat Herrador

Instituto Nacional de Estadística (INE). Madrid.Spain.

E-mail: juporras@ine.es, herrador@ine.es

Chapter 1. Introduction.

The frame in a sampling survey is the list of units from where the sample is going to be selected, together with all the additional available information for each of these sampling units, which may be used for improving the efficiency of the sampling design.

The frame must be an exact image of the population that is going to be investigated, the target population of the survey. When this requirement is not met, considerable biases may be introduced in the selection of the sample with the resultant impact on estimations. The frame is therefore one of the essential elements in a survey.

Two important aspects in relation to the frame are the sources used for obtaining it and the procedures needed to keep it up to date. If the target population is a highly dynamic population, streamlined procedures have to be designed for its updating, so that the sample selected at any time may be a satisfactory representation of the population.

In the following section it is explained how the frame was formed traditionally and then how administrative information is used today for constructing and improving household survey frames.

Chapter 2. The population census in the construction of the frame.

The sampling procedure used in household surveys is a stratified two-stage cluster design. In spite of presenting a certain loss of precision in relation to simple random sampling, this technique is commonplace in official statistics due to the fact that it represents a reduction in the costs of the investigation and facilitates frame updating.

The frame compiled for use in household surveys is a double frame formed by:
1. The list of census sections (first-stage units)
2. The list of the dwellings in each census section(second-stage units)

along with all auxiliary information available on the sampling units

Traditionally, about 3000 census sections were selected by INE and order to be used in all the household surveys carried out by INE. This frame of primary and secondary units has been periodically updated.

This frame is formed on the basis of census data. On the one hand, this information is fundamental for preparing the list of dwellings from which the sample is going to be collected for the surveys, and on the other, the census is the only source that provides detailed auxiliary information of the census sections, used as primary sampling units. In this respect we should point out that the more quality auxiliay information there is available, especially in variables correlated with the study variables, the greater the assurance will be that the sampling designs may be efficient.

Auxiliary information is used both in the stratification and substratification and in the construction of estimators in the use of Calibration Techniques.

The first-stage units are stratified according to the size of the municipality to which they belong. Within each stratum, the PSU's are grouped in substratum.

The substratification process makes use of the following variables:

- Activity status
- Occupation and branch of activity
- Nationality (proportion of foreigners)
- Highest level of education completed
- Age and sex groups
- Socio-economic condition
- Income variables:

    a) Average of taxable income per dwelling with declaring persons.
    b) Proportion of farm self-employment income over the total income of the section.
    c) Proportion of property income over the total income of the section.

The taxable income variables were supplied at section level by the Spanish Tax Agency (AEAT), after an *ad hoc* loading procedure on the Spanish Population Register, according to a recently initiated line of cooperation between AEAT and INE, without precedent and with a great potential impact on the national statistical system.

For the construction of the substrata techniques based on Cluster Analysis are used in order to take full advantage of all the information available for each census section. These techniques are used by the statistical offices of such countries as Canada (Statistics Canada) or the USA (Census Bureau).

With regard to the updating of the frame, its use by the Labour Force Survey gives us an advantage in order to maintain an ongoing updating process both of the first-stage units and of the second-stage units **in the part of the census sections that are included in the sample**. This represents approximately ten percent of the primary unit frame.

The second-stage unit frame is made up of family dwellings classified in the census either as occupied dwellings or as empty dwellings. In order that this frame may be representative in any inter-census period, probability of belonging to the sample has to be attributed to the population that did not belong to the census section at the time of the census and has been selected for the sample. Updating consists of visiting empty dwellings and any other census unit, (business premises, newly built dwellings, etc.), to see whether their situation has changed and, if so, include it in the frame. This updating process is carried out every year and a half.

These updating processes ensure that the frame is in line with the real situation at all times, taking into account population dynamism.

Chapter 3. The Spanish Population Register: Useful administrative information for the construction of the population frame.

The SPR is governed by Law 4/1996, of 10 January, in modification of Law 7/1985, of 2 April, regulating the Basis of Local Government, in relation to the Municipal Register, and its regulatory development, approved by Royal Decree 2612/1996, of 20 December, amending the Regulations on Population and Territorial Demarcation of Local Authorities.

The SPR is the administrative record in which the residents of the municipality are set down. Its data provide evidence of residence in the municipality and of having permanent address there.

The Local Council is responsible for its construction, maintenance, review and safekeeping, in accordance with the regulations approved jointly by the Ministry of the Economy and Finance and the Ministry for Public Administrations at the proposal of the Registration Council. The Review of the SPR is obtained with reference to 1 January every year. Previously the register was reviewed in the years ending in 1 and 6, and now it is maintained continuously.

On a mandatory basis the entry in the SPR contains only the following data of each person:

a) Full name
b) Sex
c) Permanent address
d) Nationality
e) Place and date of birth
f) Spanish National Identity Card number or, in the case of foreigners, the document in lieu of this.

Article 16.3 of Law 4/1996 regulating the Basis of Local Government governs the use of the Register for statistical purposes: “The Spanish Population Register may also be used for compiling official statistics subject to statistical secrecy, in the terms specified in Law 12/1989, of 9 May, on the Public Statistical Function and in the statistics laws of the autonomous communities with powers in this matter“.

The existence of this register and the appropriateness of its use for statistical purposes have provided the possibility of having a new alternative source for preparing a new survey frame. However, bearing in mind the existence of two different types of survey aimed at the population, namely continuous surveys and structural surveys, hitherto INE has only used the SPR as a sampling frame for surveys of a structural type.

The SPR is the list of the inhabitants of Spain, which permits its use both as a frame of persons, facilitating their direct selection, and as a frame of dwellings on the basis of an *ad hoc* utilisation performed on it. This utilisation consists of obtaining dwellings as a set of persons registered at the same postal address.

Being a **live record of the population,** one of the main advantages is that it makes a permanently updated frame of areas available quickly and economically, on which the two-stage sampling design ,used in the INE household surveys, can be applied.

However, with regard to the current situation, there are some problems in the use of the SPR as a frame of household surveys which should be pointed out:

The SPR is a register of inhabitants. As mentioned above, dwellings are usually the second stage sampling units through which population is investigated. An exploitation in order to obtain dwellings raises difficulties due to the lack of standardisation in the postal addresses and errors in the processing of municipal register page numbering. However, a tailor-made program is being applied to obtain dwellings and results in the field are quite satisfactory, even though it is not without errors (the existence of duplicated dwellings).

Being a public document, the SPR offers limited auxiliary information: the only information that it provides is the total number of persons per household, age and sex structure, and number of foreigners. This information is important, but compared with what is obtained in censuses it is very limited and does not permit the formation of good substrata. However it does permit, and in fact this has already been done, the analysis of certain features in the selected dwellings that have presented an incidence, such as refusal, non contact, in the surveys. Experience shows that the results based in questionnaires especially designed to perform this type of analysis, are very limited due to the existence of a high lack of response.

Registration in a municipality may be associated with a series of rights and obligations, which gives rise to the existence of people who are intentionally registered incorrectly. From the statistical point of view, this leads to two kinds of problems: on the one hand, to the **selection of misclassify dwellings** since dwellings classified as principal ones, are in fact empty dwellings and, on the other hand, persons who theoretically are living in these misclassify dwellings, do not have probability of being selected.
The updating, that are carried out on the frame based on census, eliminate this last problem because dwellings classified as empty are investigated to check whether they are really empty or not. It is in these dwellings where the incorrectly registered population may be located.

Chapter 4. Combination of administrative records and their use in sample design.

INE and AEAT have concluded an agreement by which a stable framework of cooperation is established between the two institutions in the scope of information exchange for statistical and taxation purposes.

By this agreement, AEAT provides tax information to be used for efficiency improvement in stratification process. In this respect, AEAT supplies to INE aggregated tax information associated with different types of territorial units, the most desegregated being that supplied at census section level, the primary sampling unit in the population surveys.

The aggregated tax information is obtained matching the information from SPR and from the AEAT file by the use of the ID Card Number of the persons registered in the SPR. The main aim is to obtain a classification of census sections according to the level and structure of declared income, aggregated for all the residents in the section.

The descriptive variables of the features of the section have been used for the formation of strata and substrata, in order to improve the efficiency of the selection and estimation processes in sampling surveys. The more homogeneous the above-mentioned section groupings are with regard to variables that are correlated with the survey target variables, the more efficient the sample of households resident in the sections will be. In general, the structure and level of per capita income indicators in the sections present correlations with a wide variety of social characteristics of the households that are studied in official INE surveys.

The INE experience in the utilisation of the information from STA has been as follows:

- **Sampling design of the Family Financial Survey**, carried out by the Bank of Spain in 2002 and 2005.

In the 2002 survey the criteria of stratification was the income levels of the census sections. In this sample, bearing in mind the actual objectives of the survey, there was an overrepresentation of richest families, considering these to be the families that are included in Tax on Wealth file and, furthermore, in this group the wealth intervals were considerate differently, since it was expected that as wealth level rose the non-response ratio would be higher. Consequently, in order to preserve the strict confidentiality rules, the sample of 8000 families was selected by means of the exchange of anonymity files between both institutions.

In 2005 the sample used was made up of the 5200 families that took part in the 2002 survey, to which a new sample was added in order to complete the total of 8000 families again. In the selection of this sample a similar process was followed to that of the earlier survey, taking into account the new size allocated in each stratum obtained on the basis of distribution of the population according to 2003 income levels.

- **Substratification of the primary sampling units in the frame for household surveys.**

Different income level variables of the primary sampling units, obtained from tax information of 2003 were used as substratification variables for the new sample designs of the Labour Force Survey 2005 and the Continuous Family Budget Survey 2006, respectively.

Taking into account the aims of both surveys, the stratification variables used in each of them were different. Cluster Analysis Techniques were applied at different levels of aggregation for their construction.

Chapter 5. Conclusions.

- The existence of the SPR has allowed the possibility of using an updated sampling frame different to the one that has been traditionally used in household surveys.

- Like any other administrative source, the SPR requires adaptations for its statistical use. In particular, it is essential to incorporate a code that allows the identification of dwellings and persons who live in and, in that way, the SPR can be used both as a frame of persons and of dwellings, without any kind of error. However, it will be important to study the procedure in order to be able to include all the auxiliary information possible in it so as not to forgo efficiency in the sample designs.

- The use of auxiliary information from administrative sources enables the improvement of the efficiency of household surveys. In this sense tax information from AEAT has been used.

- To afford the above three points, INE is developing a new project, Estudio Demográfico Longitudinal (EDL) in order to construct a frame to be used in household surveys. This frame will be, not only, a list a units. It will have incorporated all kind of social and demographic data. The project is going to be prepared at the end of 2008, and will be used in the next census 2011.

Madrid, March - 2007