VI. Design of the sample

A Introduction

Following the guidelines given by the European Union Statistical Office (EUROSTAT) a representative probabilistic sample of the private household population has been selected. This has been obtained as a subsample of the list of dwellings used in one of the latest household surveys carried out by the INE. This has facilitated making an updated directory available as well as some of the characteristics of these dwellings which have been used in the sample selection process.

This framework was made up by 8.000 census sections of which 999 were selected by probability.

In turn the sample framework originally used was the relationship of censal sections from the Population and Housing Census 1991.

For practical reasons a small group of provinces has been deleted from the sample. These are now represented by other provinces with similar characteristics and which remain in the same autonomous community.

Given the special difficulties presented by a panel investigation and with the object of achieving the minimal sample size required by EUROSTAT, all the incidences that occurred during the field work relating to refusals, absences and non-surveyable dwellings have been replaced.

Therefore apart from the sample dwellings, a sample of reserve dwellings was selected which was used in the replacement of those dwellings that presented some type of incidence.

The replacement was controlled in the sense that each dwelling was replaced by another with the same characteristics. The control variable used was household size

The most important characteristics of the sample design are detailed below.

B Type of sample and stratification

A stratified two stage sample has been used among first stage units.

Given that being able to facilitate certain classifications on a NUT2 level is among panel objectives (autonomous communities), an independent sample of each one of them has been selected.

The first stage units were the census sections which are considered groups of dwellings.

Although not calculated for this survey, the effect of the design obtained in other surveys that use similar designs, for characteristics related to the activity presented in 20% of the population, is around 2.

The second stage units are family dwellings. Subsample was not carried out for these dwellings and all households and persons who are household members in agreement with the definitions given for the survey, were investigated.

The stratification variable is the size of the municipality to which the section

belongs.

The following types of municipalities are considered to attain the strata formation:

1. **Autorepresented municipalities**: These are those municipalities, which given their category within the province, should always have sections in the sample.

Self represented municipalities are:

- The province capital
- Municipalities of similar or greater importance than the province capital
- For municipalities which have a notable demographic situation within the province there are no other similar ones with which to group them.

2.- **Corepresented municipalities**: Those which form part of a group of municipalities within the same province which are demographically similar and which are represented in common.

In agreement with this classification, the theoretical strata considered generally meet the following concepts:

Stratum 1: Province capital municipality

Stratum 2: Self-represented municipalities of similar or greater importance than the province capital (for example Gijón or Vigo)

Stratum 3: Self-represented municipalities important in relation to the capital or municipalities with more than 100,000 inhabitants.

Stratum 4: Municipalities with between 50,000 and 100,000 inhabitants

Stratum 5: Municipalities with between 20,000 and 50,000 inhabitants

Stratum 6: Municipalities with between 10,000 and 20,000 inhabitants

Stratum 7: Municipalities with between 5,000 and 10,000 inhabitants

Stratum 8: Municipalities with between 2,000 and 5,000 inhabitants

Stratum 9: Municipalities with less than 2.000 inhabitants

It has to be borne in mind that given the different distribution of municipality size among the different provinces it is not efficient to carry out uniform stratification for all of them. However, we have tried to carry out uniform stratification for all provinces which belong to the same autonomous community. So for example, in Andalucía and Canarias there are hardly any municipalities with less than 2000 inhabitants. For this reason the lesser stratum in these municipalities corresponds to **municipalities with less than 5000 inhabitants**, including the theoretical strata 8 and 9 in just one stratum. Conversely, in Castilla and Aragón municipalities with less than 2000 inhabitants take on a lot of importance and there are hardly any intermediate municipalities with between 5000 and 10,000 inhabitants. The theoretical strata 7 and 8 have been grouped in these communities thereby creating a municipality stratum with between 2000 and 10,000 inhabitants and fully maintaining stratum 9.

C Sample size and distribution

Sample size on a global level, for all European Union countries is 76,500 countries which means a total of approximately 155,000 persons interviewed.

The distribution of the sample among the different countries has been carried

out by bearing in mind the size and characteristics of the country aside from the need for information.

Spain has been assigned a sample of 8000 households.

A total of 8 dwellings has been set per section, for a final sample of 999 sections distributed over the whole country.

In order to achieve the objectives of this survey, a minimal sample of 400 households for each NUT1 (main regions) and 200 households for each NUT2 (autonomous community) is required for which reason the setting of 999 sections among autonomous communities has been carried out by assigning part of the sections uniformly and the rest proportionally.

The distribution of the sample of sections appears in table I.

Table I. Distribution of the sample

NUT1	NUT2	SECTIONS
1. Northwest	Galicia	63
	Asturias	40
	Cantabria	32
2. Northeast	País Vasco	54
	Navarra	32
	Rioja	29
	Aragón	42
3. Madrid	Madrid	108
4. Centre	Castilla-León	61
	Castilla-La Mancha	48
	Extremadura	40
5. East	Cataluña	113
	Comunidad Valenciana	80
	Baleares	35
6. South	Andalucía	124
	Murcia	40
	Ceuta-Melilla	-
7. Canarias	Canarias	58

The setting for each community was done first between provinces and subsequently between strata, following the setting criteria proportional to the population, although strata containing municipalities of greater size have been potentiated due to these being the ones with greater variability characteristics and also where a greater number of incidences are presented.

D Selection of the sample

The selection of primary units, within the general sample from which it has been obtained, has been carried out in each stratum with probability proportional to size.

The selection of the sample of dwellings in each section has been carried out by means of a systematic sample with random start.

E Estimators

The estimators used in each survey are corrected expansion indicators with information facilitated by external sources in order to adjust the distributions obtained by means of population distributions.
To calculate estimates a process that facilitates obtaining the final weight or weighting of each household in four steps is followed. Each one is applied after the other and this facilitates the contribution of each one of them to the final weight. (The last two steps were carried out by EUROSTAT).

These factors are calculated standardised in such a way that comparison between themselves of all the households in the sample is facilitated.

Each person within the household has the same elevation factor.

The steps to follow are the following:

Step 1. **Design factor**

This factor is the inverse of the probability of selecting the household.

In agreement with the sample selection procedure, the probability of a dwelling i in a stratum h may be calculated approximately by means of the expression:

$$P(V_{ih}) = \frac{8\,n_h}{V_h}$$

where:

n_h = number of sections set in the survey in stratum h

V_h = Total dwellings according to census in stratum h

The standardised design factor has the expression:

$$W_{ih}^{(1)} = \frac{\frac{V_h}{8\,n_h} \cdot m}{\sum_h \frac{V_h \cdot m_h}{8\,n_h}}$$

where:

m_h = effective sample of households in stratum h

$$m = \sum_h m_h$$

Step 2. **Adjustment factor for lack of response**

This factor tries to cover the effect of the different response rates that may be achieved in different parts of the sample.

The standardised expression for this factor is:

$$W_{ih}^{(2)} = \frac{m_h' \cdot m}{m_h \cdot m'}$$

where:

m_h' theoretical sample of dwellings in stratum h

m_h = effective sample of dwellings fully interviewed in the stratum

$m = \sum m_h$ = total muestra efectiva

$m' = \sum_h m_h'$ = *total muestra teórica* 4

Step 3. **Factor to correct the distribution of households**

This is introduced in order to adjust the distribution of estimated households based on the sample, after applying the weightings obtained in steps 1 and 2 for the population distribution of households, ascertained by means of external sources.

This information should be reliable and updated and may come from a census, survey or any other source.

The adjustment may be carried out for different characteristics.

The control characteristics used were:
- Distribution of households according to number of active persons
- Distribution of households according to number of persons who live there

Both have been obtain from Economically Active Population estimates in the fourth quarter of 1994.

The general setting out of the problem is as follows:

Let:

$\hat{P}_k$ = estimated proportion of households that have modality k of a determined classification characteristic (k = 1...K) where these modalities are mutually exclusive.

$$\hat{P}_k = \frac{\sum_i d_{io}\ y_{ik}}{\hat{Y}}$$

$\hat{Y}$ = Total households in the population, estimated from the sample.

$$\hat{Y} = \sum_i \sum_{k=1}^{K} d_{io}\ y_{ik}$$

d_{io} = Weight applied to the household after applying steps 1 and 2.

y_{ik} = Variable that takes values 1 or 0 according to whether household i has modality k of the characteristic.

$\sum_i$ is extended to all households in the sample.

P_k = Population proportion in modality k of the characteristic used for the adjustment.

The objective is to find a new weighting d_i ($d_i = f_i.d_{io}$) in such a way that the following is verified:

$$\hat{P}_k = \frac{\sum_i d_i \ y_{ik}}{\sum_i \sum_{k=1}^{k} d_i \ y_{ik}} = P_k$$

with the condition that the gap between d_i and d_{io} is minimal.

There are various solutions to the problem set out above that depend on the distance function chosen.

In our case the methodology developed by the Statistics Institute of France, adopted by EUROSTAT for this survey, has been applied,

The procedure facilitates obtaining factor f_i, which once this has been standardised and following previous notation we shall call W_i (3).

Step 4. **Factor to correct the distribution of persons**

This factor is introduced in order to adjust the distribution of persons in the sample to the distribution of persons in the population for certain significant demographic characteristics.

The characteristics used for the adjustment are age and sex, the population distribution used being the demographic projection of population referring to half of the fourth quarter of 1994.

The approach is analogous to that set out in step 3.

Let:

$\hat{P}_c$ = Proportion of persons with category c estimated from the sample (c = 1....C)

$$\hat{P}_c = \sum_i \frac{d_{i1} \ x_{ic}}{\hat{X}}$$

where:

x_{ic} = Number of persons who belong to category c of the adjustment variable (age groups and sex) in household i.

$\hat{X}$ = Total persons estimated in the sample

$$\hat{X} = \sum_{i} \sum_{c=1}^{C} d_i 1 \; x_{ic}$$

d_{i1} = Weight assigned to the household after applying the three previous steps

P_c = Proportion of category c obtained from the Demographic Population Projections.

The objective, analogous to step 3 is to adjust the weight of household d_{i1} by factor g_i in such a way that in applying this corrected weight $d_{i1}.g_i$, the estimated distribution of the sample coincides with the distribution given by the population projection obtained from external sources.

The adjustment is made in such a way that factor g_i approximates to 1 as much as possible, by applying the method of general minimal squares which assign the same factor to all household members.

We call this standardised factor $W_i^{(4)}$ following the nomenclature used in previous phases.

Final weighting

Each household is assigned a global adjustment factor obtained as the product of factors calculated in each one of the previous steps.

$$W_{ih}^{(final)} = W_{ih}^{(1)} \; x \; W_{ih}^{(2)} \; x \; W_i^{(3)} \; x \; W_i^{(4)}$$

This final weight is standardised, in other words the average for everybody is equal to 1.

Expansion factor

The previous weightings are valid for the estimation of averages and proportions.

In order to estimate the total of the characteristic investigated, the quotient between the total population and the corresponding sample population P/p, we multiply by the standardised sample weightings. With the totals factor the result is: $P/p.W_{ih}$.

F Estimators for cycles subsequent to the first one

The calculation of the weightings for the transversal and longitudinal analysis in panel type surveys over time, involves a series of complex techniques.

The initial weightings assigned to sample persons in cycle 1 have to be adjusted in order to reflect the changes in the study population as well as the evolution of the sample over time.

These adjusted weightings are what we will call **basic weightings**. In principle, the data from the original sample with the basic weightings may be used for transversal and longitudinal analysis. However, and given that the panel also contains information on non-sample persons who have entered the survey due

to living in a household with one or more sample persons, an approximation will be used to include information facilitated by these non-sample persons in the analysis.

The weightings necessary for specific analysis are obtained by means of simple transformation of the basic weights.

F.1 CALCULATION OF BASIC WEIGHTINGS

a) **Initial weight**

In cycle 1, for each household i and each household member k, the same weight is assigned which in the previous section we called $W_{ih}^{(final)}$.

Let us call this factor **initial weight** which from now we will note by

$$u_i^{(1)} = u_{ik}^{(1)}$$

defined for all households in the sample and for all sample persons.

b) **Basic weight**

In each cycle t, this weight is defined for all sample persons. *Sample persons* are considered those who formed part of the sample in cycle 1 plus the children born to a sample woman while still belonging to the sample.

The basic weight of a person in cycle t is obtained by the adjustment of its basic weight assigned in cycle t-1, in other words

$$u_{ik}^{(t)} = u_{ik}^{(t-1)} \cdot f_{ik}^{(t-1 \to t)}$$

To calculate the adjustment factor $f_{ik}^{(t-1 \to t)}$ the following factors that cover changes that have occurred in the population and the sample between cycles t-1 and t have to be taken into account:

1º) **Adjustment of lack of response**
The weights from cycle t-1 are multiplied by a factor inversely proportional to the probability of the persons collaborating in cycle t, having collaborated in cycle t-1
This probability is determined based on certain characteristics of households or persons.

2º) **Adjustment with external sources**
In each cycle, the weights should be adjusted so that the estimated data based on the sample are consistent with the distribution of the population according to different demographic characteristics.

3º) **Children born between t-1 and t** by a sample woman receive the basic weight of their mother.
Children born by a mother who is not a sample person are not considered sample persons.

These basic weights are defined on a person but not household level.

All the non-sample persons have a basic weight equal to zero.

F.2 ESTIMATORS FOR TRANSVERSAL ANALYSIS

Each household i is assigned a weight obtained as the average of the basic weights of its adult members, in other words:

$$w_i^{(t)} = \frac{1}{s+n} \sum_{k=1}^{s} u_{ik}^{(t)}$$

where:

$u_{ik}^{(t)}$ = basic weight assigned to adult sample k in household i, in cycle t

s = the number of sample adults

n = the number of non-sample adults

In the same way that households interviewed in cycle 1 weighted by their initial weights $u_i^{(1)}$ provide a representative sample of households in the period when the panel was started, the households interviewed in any cycle subsequent to t, weighted by weights $w_i^{(t)}$ provide a representative sample in time t.

For the transversal analysis of persons, for all sample and non-sample members, adults and children, the household weight calculated previously is assigned, in other words

$$W_{i,j}^{(t)} = \mathrm{W}_{\mathrm{i}}^{(\mathrm{t})}$$

Everybody interviewed in cycle t weighted with their **shared weight** $w_{ij}^{(t)}$ are, for the objective of representing the transversal population in period t, equivalent to sample persons interviewed in t weighted with their basic weights or $_{Ik}^{(t)}$.

Using the first one instead of the last one facilitates an extension of the sample base due to the inclusion in the analysis of non-sample persons who live with sample persons.

F.3 LONGITUDINAL ANALYSIS OF PERSONS

The longitudinal analysis has to be specified for a time interval and limited to individuals that belong to the sample during said interval.

The transformation of basic weights is similar to that indicated in the transversal analysis, in other words considering the time interval elapsed between t and T.

$$w_{ij}^{(t\to T)} = \frac{1}{s^{(T)} + n^{(t\to T)}} \sum_{k=1}^{S^{(T)}} u_{ik}^{(T)}$$

where:

$u_{ik}^{(T)}$ = Basic weight assigned to adult sample k from household i in cycle T

$s^{(T)}$ = Total sample adults in T

$n^{(t \to T)}$ Number of non-sample adults in T and which formed part of the sample in t or before.

All household persons have the same weight.