Survey of Active Population

Technical report

Madrid, May 2002
Sample design area and evaluation of results

Index

Introduction

The Active Population Survey (APS), is a continuous survey intended to study the socioeconomic characteristics of the population, which has been carried out by the INE since 1964. Its design is framed within the General Population Survey (GPS).

Since its implementation it has undergone modifications in some aspects, always directed at improving the survey.

This reports objective is to cover the methodological aspects of the current design, as well as evaluating the data quality of the same.

The INE is grateful in advance for as many suggestions as can be presented for possible future improvements of the survey.

II. Design of the survey

1 Objectives

The APS main objective is to ascertain the economic activity of the country relative to its human component. It is oriented towards providing information on the main categories of population in relation to the labour market as well as obtaining classifications of these categories according to various variables.

Experience has demonstrated that the different statistical sources (Census, Wages Survey, Registered Unemployment, etc.) that provide information on these subjects are not adequate to satisfy the surveys objectives. In the concrete case of the Census for various reasons:

1) Its long interim period prevents knowing about the situation between Census'.

2) The Censal data is not sufficient in order to facilitate a detailed view of the labour situation.

3) The data is provided by the unassisted informant, therefore difficulties exist when the informant interprets the concepts used.

The need for a continuous survey, designed and conceived exclusively to obtain the populations degree of economic activity together with characteristics that are closely related, is thus justified.

The survey is designed to provide detailed results on a national level. Information is offered about the main characteristics broken down as much as the estimates variation coefficients for the Autonomous Communities and the provinces allow.

The definition of economically active population has been that accepted by the International Labour Organisation (ILO), according to which it is defined as the *group of persons, who in a given reference period, supply labour for the production of goods and services or who are available and attempt to incorporate themselves to the said production*.

The economically active population is made up of persons 16 years old and over who satisfy the necessary conditions in the reference week for their inclusion among the employed or unemployed persons according to the definitions given for the survey.

2 Survey's scope

The surveys scope is broken down in the following three sections:

2.1 POPULATION SCOPE

The Survey is directed at the population which reside in family dwellings, in other words, dwellings used all or part of the year as a habitual or permanent residence.

Collective dwellings, such as hospitals, hotels, barracks, convents, etc. are excluded from the investigation.

Those families that form an independent group and reside in these establishments, such as can occur with the centre directors, concierges and porters are included. In theory, only that population which lacks family residence is excluded from the survey definitively.

2.2 GEOGRAPHICAL SCOPE

The survey is carried out in the whole country.

2.3 TEMPORAL SCOPE

The APS is a continuous quarterly survey whose interviews are extended over the quarter's thirteen weeks.

As for the reference period we have to distinguish:

Reference period of the survey results: the quarter

The information reference period: has been adapted, as a general rule, the week before (from Monday thorough Sunday) the week of the interview. Said week is named *reference week* and all data should refer to it, save for the exceptions which appear in *Active Population Survey document. Survey description, definitions and instructions for completing the questionnaire*.

3 Surveys framework

To define the Surveys framework, it is necessary to begin with the administrative division of Spain which appears in the following form:

The whole Nation is divided into 17 Autonomous Communities, and at the same time in 50 provinces of which 47 are peninsular and 3 insular. The provinces are divided into municipalities and these into municipal districts.

This is the official administrative division. The INE, together with the Town Halls, then make a new sub-division of these districts into census sections.

The sections are used for all the work entrusted to the INE, in which an infra municipal division is necessary, amongst other needs for electoral purposes such as *electoral sections*, which according to Electoral Law require that each section include a maximum of 2,000 electors and a minimum of 500.

Therefore, the censal section can be considered a perfectly defined area, whose population size is limited by the conditions previously exposed.

The sectioning and it's enumeration varies considerably over time, due to which it is updated with reference to 1 January of each year, coinciding with the Electoral Census revision and in each Census or Register. On the one hand, there are sections that become unpopulated and it is necessary to join them with others. On the other hand, the contrary phenomenon also occurs, in other words, the sections grow until they exceed the population limits established and it is necessary to divide them. In all cases the sections selection probability is updated.

4 Design of the sample

4.1 SAMPLE TYPE. SAMPLE UNITS

The type of sample used is a two-stage sample with stratification of the first stage units.

The first stage units are made up of the censal sections. The sections sample remains fixed with the following exceptions:

a) When the results obtained in the Census cast slight variations in the populations structure which require a different fixing.

b) The available dwellings in the section are finished.

c) When in updating the selection probabilities its removal from the sample corresponds.

The second stage units are made up of the main family dwellings (permanently occupied) and fixed lodgings (shacks, caves, etc.). Second homes (occupied only part of the year) and those available for rent or sale are not considered, since they do not form a part of the population scope previously defined.

Within the second stage units no sub-sample is carried out, all the information is collected from all persons who habitually reside in them.

4.2 STRATIFICATION OF THE SAMPLE UNITS

The first stage units are stratified following a double criteria:

A.- Geographic criterion (of stratification)

The sections are grouped in strata by the province and type of municipality (according to geographic importance) to which they belong.

B.- Socioeconomic criterion (for stratification)

In agreement with the design of the GPS in each province the sections are grouped in *substrata* according to demographic importance of the municipality to which they belong and the socioeconomic category of the dwellings located there.

Strata

To attain the strata formation the following types of municipalities are considered:

1. Self represented. These are those which given their category within the province must always have samples in the sample.

Self represented municipalities are:

The province capital

Municipalities that have a certain number of inhabitants which in the proportional fixing within the province correspond to at least 12 sections in the sample.

There are no other similar municipalities with which to group municipalities having a highlighted demographic situation within the province, although proportionally they correspond to at least 12 sections in the sample.

2. Co-represented municipalities: These are municipalities within the same province that form part of a group of demographically similar municipalities and are represented in common.

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

Stratum 1. Province capital municipalities

Stratum 2. Self represented municipalities, important in relation to the capital

Stratum 3. Other self represented municipalities, important in relation to the capital or municipalities with more than 100.000 inhabitants

Stratum 4. Municipalities between 50.000 and 100.000 inhabitants

Stratum 5. Municipalities between 20.000 and 50.000 inhabitants

Stratum 6. Municipalities between 10.000 and 20.000 inhabitants

Stratum 7. Municipalities between 5.000 and 10.000 inhabitants

Stratum 8. Municipalities between 2.000 and 5.000 inhabitants

Stratum 9. Municipalities with less than 2.000 inhabitants

It has to be taken into account that given the different distribution in sizes of municipalities between the different provinces it has not been possible to carry out a uniform stratification for all of them. For example, in the province of Lugo there are only 10 municipalities with less than 2.000 inhabitants for which theoretical strata 8 and 9 have been grouped in stratum 8 which contains the municipalities with less than 5.000 inhabitants. Conversely, the province of Burgos has more than 350 municipalities of at least 2.000 inhabitants included in stratum 9 However theoretical strata 7 and 8 are grouped together into stratum 9, there being hardly any municipalities with between 2.000 and 5.000 inhabitants. Nevertheless, whenever it has been possible, a uniform stratification has been carried out for all provinces belonging to the same Autonomous Community .

The censuses and registers contribute the necessary information to update the stratification in each province by virtue of the population of municipalities.

Substrata

For the formation of substratum the socioeconomic category of the dwellings located in the section have to be taken into account.

The sections change substratum due to the variation of the population structure, for which the substratification is revised in each subcensus using the information which is provided on the characteristics intervening in the definition of socioeconomic category.

This information permits classifying the economically active population of the section in four large groups:

Group 1 (Farmers). Comprises the following categories

01. Agricultural business persons with employees

02. Agricultural business persons without employees

03. Members of agricultural co-operatives

04. Directors and Chiefs for agricultural operations

05. Rest of agricultural workers

Group 2 (Freelance workers). Covers the following categories:

06. Freelance professionals

07. Non-agricultural business persons with employees

08. Non-agricultural business persons without employees

09. Members of non-agricultural co-operatives

Group 3 (Managers, freelance professionals and administrative personnel):

10. Directors of non-agricultural companies and senior civil servants

11. Professionals employed by others12. Administrative, Sales and Services Department Chiefs for non-agricultural companies or Public Administration

13. Remaining administrative and sales personnel

18. Armed Forces Professionals

Group 4 (Remaining workers). Covers the following categories:

14. Rest of service personnel

15. Non-agricultural foreman and supervisors

16. Non-agricultural skilled workers

17. Non-agricultural unskilled workers

The **Non classifiable are not included in any group:**

Persons who seek employment for the first time.

Economically active persons who cannot be classified under any of the previous headings.

There are sixteen strata, fifteen of which are obtained by virtue of the population percentages of groups 1, 2, 3 and 4 and the sixteenth (substratum 0) is made up by those sections with a high percentage of inactive population.

The definition of the fifteen substrata is established according to: 1) there being a clear predomination of one of the four groups over the other three; 2) two predominate over the other two; 3) three groups predominate and 4) there is no clear predomination of any of the four groups. In some of the strata many of the various substrata may not exist.

The criterion used to determine if a section belongs to one or another sub-strata is the following:

The predominant group (the one with the largest percentage) we called importance group A. The following groups, according to decreasing percentages we named B, C and D respectively.

For example, if in a determined section the active population percentages are 40 percent of the group 3, 30 percent of group 2, 20 percent of group 1 and 10 percent of group 4 then A = 3, B = 2, C = 1 and D = 4.

Each section is assigned a four digit code as follows:

a) The first digit is:

1 if the group of importance A is group 1

2 if the group of importance A is group 2

3 if the group of importance A is group 3

4 if the group of importance A is group 4

b) To obtain the second digit we calculate the quotient:

$$\frac{PercentageGroupof\ \mathrm{Im}portanceB}{PercentageGroupof\ \mathrm{Im}portancceA}\ whichweshallcall\ \frac{B}{A}$$

The following cases may occur:

1st Case: If $\frac{B}{A} > 0{,}66$, the 2nd digit will be:

1 if the group of importance B is group 1

2 if the group of importance B is group 2

3 if the group of importance B is group 3

4 if the group of importance B is group 4

2nd Case: If $\frac{B}{A}$ is comprised between 0,33 and 0,66, the 2nd digit will be:

5 if the group of importance B is group 1

6 if the group of importance B is group 2

7 if the group of importance B is group 3

8 if the group of importance B is group 4

3rd Case: If $\frac{B}{A} < 0{,}33$ the 2nd digit will be zero in any case

c) to obtain the third digit we calculate the quotient:

$$\frac{PercentageGroupOf\ \mathrm{Im}portance\ C}{PercentageGroupOf\ \mathrm{Im}portanceA}$$

and we will then apply the same criterion explained in the previous section b).

d) To obtain the fourth digit we calculate the quotient:

$$\frac{Percentagegroupofimportance\ D}{Percentagegroupofimportance\ A}$$

The Substrata code cannot have digits repeated as can be easily verified (except zero).

Substrata zero code: As an exceptional case, if in a section it is verified that:

$$\frac{Population16yearorolderInactive}{Population16yearsorolderActive}>3$$

this sections substrata code is zero, without applying the socioeconomic category code assignment criteria , previously exposed.

In agreement with the exposed encoding, the substratification is carried out in the following manner:

Substrata O

Covers only code zero sections, in other words, those sections with a strong predominance of inactive population and, therefore, the socioeconomic category substratification criteria not been applied.

Substrata 1

Groups the sections predominantly from Group 1, in other words, those whose four digit code is one of the following:

1000 - 1600 - 1700 - 1800 - 1670 - 1760 - 1680 - 18601780 - 1870 - 1678 - 1687 - 1768 - 1786 - 1867 - 1876

Substrata 2

Groups the sections predominantly from Group 2, codes:

2000 - 2500 - 2700 - 2800 - 2570 - 2750 - 2580 - 2850 - 2780 - 2870 - 2578 - 2587 - 2758 - 2785 - 2857 - 2875

Substrata 3

Groups the sections predominantly from Group 3, codes:

3000 - 3500 - 3600 - 3800 - 3560 - 3650 - 3680 - 3860 - 3580 - 3850 - 3568 - 3586 - 3658 - 3685 - 3856 - 3865

Substrata 4

Groups the sections predominantly from Group 4, codes:

4000 - 4500 - 4600 - 4700 - 4560 - 4650 - 4670 - 47604570 - 4750 - 4567 - 4576 - 4756 - 4765 - 4657 - 4675

Substrata 12

Groups the sections predominantly from groups 1 and 2, codes:

1200 - 1270 - 1280 - 1278 - 1287 - 2100 - 2170 - 2180 - 2178 - 2187

Substrata 13

Groups the sections predominantly from groups 1 and 3, codes:

1300 - 1360 - 1380 - 1368 - 1386 - 3100 - 3160 - 3180 - 3168 - 3186

Substrata 14

Groups the sections predominantly from groups 1 and 4, codes:

1400 - 1460 - 1470 - 1467 - 1476 - 4100 - 4160 - 4170 - 4167 - 4176

Substrata 23

Groups the sections predominantly from groups 2 and 3, codes:

2300 - 2350 - 2380 - 2358 - 2385 - 3200 - 3250 - 3280 - 3258 - 3285

Substrata 24

Groups the sections predominantly from groups 2 and 4, codes:

2400 - 2450 - 2470 - 2457 - 2475 - 4200 - 4250 - 4270 - 4257 - 4275

Substrata 34

Groups the sections predominantly from groups 3 and 4, codes:

3400 - 3450 - 3460 - 3456 - 3465 - 4300 - 4350 - 4360 - 4356 - 4365

Substrata 123

Groups the sections predominantly from groups 1, 2 and 3, codes:

1230 - 1238 - 1320 - 1328 - 2130 - 2138 - 2310 - 2318 - 3120 - 3128 - 3210 - 3218

Substrata 124

Groups the mixed sections predominantly from groups 1, 2 and 4, codes:

1240 - 1247 - 1420 - 1427 - 2140 - 2147 - 2410 - 2417 - 4120 - 4127 - 4210 - 4217

Substrata 134

Groups the mixed sections predominantly from groups 1, 3 and 4, codes:

1340 - 1346 - 1430 - 1436 - 3140 - 3146 - 3410 - 4316 - 4130 - 4136 - 4310 - 3416

Substrata 234

Groups the mixed sections predominantly from groups 2, 3 and 4, codes:

2340 - 2345 - 2430 - 2435 - 3240 - 3245 - 3420 - 3425 - 4230 - 4235 - 4320 - 4325

Substrata 1234

Groups the mixed sections predominantly from groups 1, 2, 3 and 4, codes:

1234 - 1243 - 1324 - 1342 - 1423 - 1432 - 2134 - 2143 - 2314 - 2341 - 2413 - 2431 - 3124 - 3142 - 3214 - 3241 - 3412 - 3421 - 4123 - 4132 - 4213 - 4231 - 4312 - 4321

4.3 SIZE OF THE SAMPLE

For the determination of the number n of sections and m of dwellings per section of the sample a linear cost type function and of the expression of the variation coefficient for a proportion in the sampling of conglomerates with subsampling was the basis.

The following cost function is followed:

$$Q = n\,Q_S + n\,m\,Q_V \qquad \text{with} \qquad Q_S = Q_F + d\,Q_D$$

where:

Q = Total budget for paying the interviewers

Q_S = Primary unit cost (section)

Q_V = Final unit cost (dwelling)

n = Number of sections

m = Number of dwellings per section

Q_F = Fixed cost per section

Q_D = Daily cost for field work

d = Number of days necessary for field work

All the variables were known except n and m.

The variation coefficient for a proportion is given by

$$C^2(\hat{P}) = \frac{V(\hat{P})}{\hat{P}^2} = \frac{1-\hat{P}}{\hat{P}} \cdot \frac{1+\delta(m-1)}{n\,m} = \frac{1-\hat{P}}{\hat{P}}\,F(\delta, m, n)$$

being:

$$F(\delta, m, n) = \frac{1+\delta(m-1)}{n\,m}$$

and δ the interclass correlation coefficient, which has been calculated as 0,05 for the active population case.

The minimum of the expression $C^2(\hat{P})$ with respect to the variables m and n are obtained by calculating the minimum of the expression $F(\delta, m, n)$ which is independent of $\hat{P}$.

For different values of m compatible with field work,

m = 4, 6, 8, 10, 11, 14, 17, 18, 19, ...91, 100

and the corresponding values of n given by

$$n = \frac{Q}{Q_S + m\ Q_V}$$

different values are obtained for F (δ, m, n).

The minimum value of F (δ, m, n) with respect to m and n corresponded to m = 20 and n = 3.000.

Based on this result the sample is fixed at a total of 3.000 sections, investigating an average of 20 dwellings per section.

The sample has subsequently undergone various enlargements in order to achieve a greater representation in some Autonomous Communities and at the same time comply with the European Union requirements for the sample size for Employment Surveys. As of 1999 a section size of 3.484 and 18 dwellings per section is established, except in the provinces of Madrid, Barcelona, Sevilla, Valencia and Zaragoza where the number of interviews per section is 22.

4.4 FIXING

This section covers the criteria followed for the distribution of sample sections among provinces, within the province between strata and within these between substrata.

To obtain the fixing between provinces the following aspects were taken into account:

a) To dispose of a minimum sample size in each province which facilitates its estimates.

b) The national results must be as reliable as possible.

c) There must be an exact number of *Blocks* in each province. The block is defined as a set of sections that an interviewer must visit during a quarter. In this surveys specific case, the block includes 13 sections distributed one per week of the quarter. In order to make the three conditions compatible a compromised fixing has been accepted between the uniform and the proportional, based on grouping provinces of similar geographic importance and assigning from 3 to 12 blocks.

Within each province the fixing between strata is proportional to the size of each one of them, the strata have been potentiated for strata where the largest municipalities appear since it is expected that the majority of characteristics studied will be correlated to the cultural and socio-economic levels of the inhabitants and it is precisely in these strata where the scattering should be greatest and where the cost per interview the least generally.

Within the strata, the fixing between substrata is strictly proportional to size (measured in number of family dwellings).

In table 1 the distribution of the sample sections by provinces and strata is shown.

Chart 1

Distribution of the section samples by provinces and strata

Provinces	1	2	3	4	5	6	7	8	9	Total
1 Alava	30				3		6			39
2 Albacete	19				7		3	6	4	39
3 Alicante	19	13		6	20	7	7	3	3	78
4 Almeria	16				7	4	3	6	3	39
5 Avila	13						4	6	16	39
6 Badajoz	20				13	6	16	13	10	78
7 Baleares	39				19	13	10	10		91
8 Barcelona	61		36	16	20	10	7	3	3	156
9 Burgos	20				7	3	9			39
10 Cáceres	19				7	4	10	16	22	78
11 Cádiz	13	13	6	20	13	7	6			78
12 Castellón	26				20	10	6	7	9	78
13 Ciudad Real	13	9			13	13	16	7	7	78
14 Córdoba	33				16	7	13	9		78
15 Coruña (La)	22			13	6	17	14	6		78
16 Cuenca	10						7	6	16	39
17 Girona	16				17	13	9	13	10	78
18 Granada	29				13	9	10	10	7	78
19 Guadalajara	20					3		3	13	39
20 Guipúzcoa	26			6	13	20	7	6		78
21 Huelva	16					10	6	7		39
22 Huesca	13					10	6		10	39
23 Jaén	16	7			13	13	13	13	3	78
24 León	23	9				10	7	10	19	78
25 Lleida	16					3	3	7	10	39
26 Logroño	26					6	6	4	10	52
27 Lugo	13					9	7	10		39
28 Madrid	98		30	10	9	3	6			156
29 Málaga	36			10	16	7	9			78
30 Murcia	30	16		6	19	13	7			91
31 Navarra	32				4	10	6	13	13	78
32 Ourense	16					3	4	13	3	39
33 Oviedo	23	29		20	10	19	7	9		117
34 Palencia	20						3	6	10	39
35 Palmas (Las)	43			9	20	7	9	3		91
36 Pontevedra	10	26			13	16	10	3		78
37 Salamanca	20					3	3		13	39
38 S.Cruz Tenerife	23	13			13	10	13	6		78
39 Santander	29	10				13	10	10	6	78
40 Segovia	16						4	4	15	39
41 Sevilla	52			10	20	16	10	9		117
42 Soria	17						4	6	12	39
43 Tarragona	19	13			7	10	9	10	10	78
44 Teruel	10					4	9		16	39
45 Toledo	13	13					17	19	16	78
46 Valencia	45			10	26	13	10	7	6	117
47 Valladolid	36					3	6		7	52
48 Vizcaya	29	7		13	9	7	6	7		78
49 Zamora	16					4			19	39
50 Zaragoza	59					4	6		9	78
51 Ceuta	13									13
52 Melilla	13									13
TOTAL	1.305	178	72	149	393	369	373	306	339	3.484

4.5 SAMPLE SELECTION

The sample selection has been carried out in such a way that within each strata any family dwelling has the same probability of selection, in other words, that there are **self weighted samples within each strata**.

For this, the first stage units (censal sections) are selected with probability proportional to the number of main family dwellings, according to the data from the last Census or Register. Within each section selected in the first stage, a fixed number of family dwellings is selected with equal probability by means of a systematic sample with random start. For this survey it has been determined that 18 dwellings be selected per section (see section 4.3).

Therefore, the probability of selecting a dwelling I, belonging to section j of strata h, where K_h sections have been set would be

$$P(V_{ijh}) = P(S_{jh}) \times P(V_{ijh} / S_{jh}) = K_h \times \frac{V_{jh}}{V_h} \times \frac{18}{V_{jh}} = K_h \times \frac{18}{V_h}$$

where

$P(S_{jh})$ = Selection probability for section j of strata h

$P(V_{ijh}/S_{jh})$ = Probability of the selection of dwelling I conditioned for the selection of section j.

V_{jh} = Total dwellings in sample dwellings in section I, stratum h

V_h = Total dwellings from stratum h.

As can be seen, this probability does not depend either on I or j.

4.6 DISTRIBUTION OF THE SAMPLE IN TIME

The distribution of the sample is uniform over time.

Each period of the survey is a quarter. Each one of the sample sections visited is one of the 13 weeks in each quarter.

The total sample is divided into three representative independent subsamples, each one of them from the whole population.

4.7 ROTATION SHIFTS

As said in the previous paragraph each period of the survey is a quarter, this repeating successively.

The censal sections remain fixed in the sample indefinitely (save the exceptions that feature in section 4.1), however the family dwellings are renewed partially each quarter of the survey, in order to avoid the families tiring. This renewal is carried out for a sixth of the sections.

For these purposes the total sample is divided into six subsamples that are known as *rotation shifts*. Each section is identified by a five digit code. The last digit expresses the rotation shift it belongs to, numbered from 1 to 6.

Each quarter dwellings that belong to the sections of a determined rotation shift are renewed. Therefore each dwelling belongs to the sample for six consecutive quarters, at the end of which it comes out to be replaced by another from the same section.

Each quarter the rotation shift sections whose dwellings are interviewed for the last time are updated with the object of being able to incorporate those dwellings into the sample in the following period. This applies to both new construction and those which have been transformed into family dwellings, which when the last census or register was carried out did not exist or were unoccupied or destined for other purposes different from that of main dwelling.

These dwellings are incorporated into the sample with a probability which is the same as the original from the section dwellings.

Cycle is considered in the survey for each one of the periods. These are listed correlatively from the implementation period. By virtue of the cycle number the rotation shift corresponding to the sections whose dwellings are renewed in the survey can be determined by means of the following equation:

$$\text{Rotation shift} = (\dot{6} - \text{cycle number}) + 1$$

$\dot{6}$ expresses the multiple number of 6 nearest through excess to the cycle number

In this way for example, in the first quarter of 2002 the cycle number 118 was carried out in the survey. The multiple of 6 nearest due to excess is 120, where

$$TR = (120 - 118) + 1 = 3$$

This means therefore that in this quarter the dwellings from the sections that belong to rotation shift 3 were renewed.

4.8 ESTIMATORS

Up to 2001 **ratio estimators have been used** taking the demographic population projections elaborated by the INE as an auxiliary variable. The expression of the estimate of a determined characteristic Y in a survey quarter is the following:

$$\hat{Y} = \sum_{h} \frac{P_h}{p_h} \sum_{i=1}^{n_{th}} y_{hi} \qquad (1)$$

extending the sum h to the strata of a province, an Autonomous Community or the national total and where:

P_h : is the projection of the population that lives in family dwellings, in stratum h, referring to half of the quarter.

p_h : is the number of persons who live in sample dwellings, in stratum h, at the time of the interview.

n_h : is the number of dwellings in the sample section in stratum h.

y_{hi} : is the value of the characteristic investigated in dwelling I from stratum h.

From the first quarter of 2002 **Reweighting techniques** are applied to the estimators with the object of adjusting the estimates of the survey to the information coming from external sources.

The reweighting technique consists of the following:

$s = \{u_1 .... u_k ..... u_n\}$

Is considered a population $U = \{u_1 ..... u_N\}$ of which a sample is taken

The expression (1) can be written in the following manner:

$$\hat{Y} = \sum_{k \in s} d_k \, y_k$$

where:

y_k : Value of the investigated characteristic of the sample unit k.

d_k : Elevation factor of unit k obtained by means of the expression $\frac{P_h}{p_h}$, h being the strata to which the unit belongs.

$\sum_{k \in s}$: Sum extended to all sample units s.

J auxiliary variables are available whose values are known for the sample and whose totals are known for the population

$$X_j = \sum_{k \in U} x_{jk}$$

This is about finding a new estimator

$$\hat{Y}_w = \sum_{k \in s} w_k \, y_k$$

where the new weightings w_k fulfil the following conditions:

$\forall \; j = 1.....J$

- They are close to the initial weightings d_k
- They verify the balancing equation

  $$\sum_{k \in s} w_k \, x_{jk} = X_j$$

The problem is to find some values w_k that minimise the expression:

$$\sum_{k \in s} d_k \, G\left(\frac{w_k}{d_k}\right) \qquad \text{with the condition} \qquad \sum_{k \in s} w_k \, X_k = X$$

where:

G = Function of distance.

X = Dimension vector (J,1) with the total of auxiliary variables.

X_k = Dimension vector (J,1) with the value of auxiliary variables in sample unit k.

The solution of the problem depends on the distance function G that is used.

If the linear distance function of the argument is considered $z = \frac{w_k}{d_k}$:

$$G(z) = \frac{1}{2}(z-1)^2, \quad z \in R$$

the problem is solved by means of the use of Lagrange multipliers which facilitate obtaining a set of factors w_k which verify the balancing conditions and provide the same estimates as the generalised regression estimator.

In the particular case of the APS it has been opted to use the linear distance function but truncated (to avoid negative solutions of the equation system) with the object of getting the most out of the regression estimator properties, with small variation and minimal skew.

The following have been used as auxiliary variables:

- Population 16 years old and over by age groups and sex on an Autonomous Community level.
- Population of 16 years old and over by province.

In this manner, utilising the current estimators used in the APS, the population is correctly estimated by age group and sex.

For the practical solution to this problem the CALMAR (CALage sur MARges) software, programmed by the INSEE (National Statistics and Economic Studies Institute) of France, has been used.

5 Updates in the surveys framework

Continuous population variations either in their characteristics or in their spatial distribution require updates to be carried out in the framework that have repercussions on the sample structure.

In the APS framework the following three types of updates are considered:

Updates in the sections framework, a consequence of the modifications produced by various incidences such as partitions, fusions or variations of limits in the selected sections. In each one of these cases it is necessary to determine the probability of selection of the new sections as well as the number of interviews to carry out for the same.

Updates in the restricted and exclusive dwellings framework for the sample sections. This update as mentioned in section 4.7 has the objective of incorporating the main *tall* dwellings from the section into the list of dwellings.

General update relative to all the population sections and dwellings, in which the section selection probability is updated and is carried out periodically when the necessary information becomes available.

5.1 = TOTAL DWELLINGS IN SAMPLE DWELLINGS IN SECTION I, STRATUM H

The following cases are considered:

5.1.1 Division of sections

In the case of a section S in which the total growth of the number of main dwellings should be split into various parts S_1, S_2 ... S_K, either to form new sections or to incorporate into pre-existing ones.

The problem of establishing the selection probabilities of the new sections to get to know which will belong to the sample, as well as the number of dwellings to interview so that the sample continues to be selfweighted is put forward.

Two cases stand out:

A) Section S is broken down to form two or more complete sections.

In this case the following is done:

1) We call

V_S = Number of dwellings from section S according to the last census

V'_S = Number of dwellings from section S after updating it.

V_{Sj} = Number of dwellings from section j of section S according to data from the last census.

V'_{Sj} = Number of dwellings from part j from section S after updating it.

2) One of the new sections S_j is selected with probability proportional to its updated size V'_{Sj} / V'_S

3) The number of dwellings that must be surveyed is

$$m_j = 18 \frac{V'_s}{V_s}$$

which are systematically selected.

In this way the sample continues to be selfweighted.

B) Section S is broken down to be annexed to one or more existing sections.

In this case:

1) One of the fragments with probability proportional to its original size according to the last census V_{Sj} / V_S and the new section S'_j is selected where the said part is incorporated will be automatically selected.

2) The number of dwellings that have to be interviewed is given by

$$m_j = 18 \frac{V'_{S'_j}}{V_{S'_j}}$$

where

$V'_{S'j}$ = Number of main dwellings currently in the new section S'_j.

$V_{S'j}$ = Number of main dwellings that existed in the last census or register within the limits of the new section S'_j.

5.1.2 Joining of sections

Due to some sections becoming empty because of migratory and natural movements they are merged with another or others which in such a way that in the case of being selected there are units to be investigated.

The case of the merging of sections is a particular case of the partition studied in section 5.1.1.B.

Therefore is section S_j selected is merged with another to form the new section S this is automatically incorporated into the sample and the number of dwellings to interview is:

$$m = 18 \frac{V'_S}{V_S}$$

where

V'_S = Number of main dwellings currently in new section S

V_S = Number of main dwellings, according to the last census or register, within the limits of the new section S.

5.1.3 Variation of limits

This is the case of a section that is made up of fragments of two or more sections through readjustment of their limits.

For the calculation of the probability of selection this case may be considered as a two stage process: the first to partition each section and the second the appropriate fusion of the sections resulting from the partition.

In all cases detailed above the new sections are incorporated into the sample when due to *rotation shift* it corresponds to renewing the families in the sections affected by the said incidences.

5.2 RENEWAL OF THE SAMPLE AS A CONSEQUENCE OF UPDATING THE SELECTION PROBABILITIES

When information becomes available, either proceeding from the electoral files, Population Census or Register, the sections selection probabilities are updated and the number of interviews per section are adjusted to 18.

This procedure is carried out in such a manner that the selection probabilities of a section are proportional to the number of dwellings that each one has at that

moment. In theory, this could be accomplished starting from zero and selecting a new sample, but that would provoke a complete rupture with the previous sample, which is risky in the case of continuous surveys such as the APS. Therefore a procedure is decided upon which maintains the sample with minimum variations without distorting the selection probabilities which actually correspond.

This procedure, due to KISH (1971), is the following:

Let S be a section belonging to stratum h, selected in a census or register C, with probability

$$P_S = \frac{V_S^C}{V_h^C} = \frac{Dwellings in S according to census\ C}{Dwellings in stratah according to census\ C}$$

and suppose that in the following census or register C', there corresponds a probability of selection given by

$$P'_S = \frac{V_S^{C'}}{V_h^{C'}} = \frac{Dwellings in S according to census C'}{Dwellings in stratah according to census\ C'}$$

P_S is compared with P'_S with one of the following cases being possible:

1) If $P'_S > P_S$ the section S belongs in the sample with probability P'_S, since if it was selected with a probability P_S below that which actually corresponds to it, it would have been selected with even greater reason applying its actual probability P'_S.

2) If $P'_S < P_S$ the section belongs in the sample with probability P'_S/P_S and is removed from the section $1 - P'_S / P_S$.

This criteria motivates the exit of a certain number of sections from the sample. These will be replaced by other sections of the same stratum but selected from **those not belonging to the sample having increased in probability**.

With this criteria the scheme that the probability of a section belonging to the sample is that which actually corresponds to it, in other words, proportional to the actual dwellings is maintained.

III. Evaluation of the Quality of APS Data

1 Introduction

The errors that affect each survey may be divided into two main groups.

Sample errors, which come from obtaining the results on the characteristics of a population, based on information collected in a sample of the same.

Errors foreign to the sample, which are common to all statistical investigation, both if the information is collected by sample and if it is carried out via a census. These errors occur in any phase of the statistical process.

- Before taking data: due to deficiencies in the framework and insufficiencies in the definitions and questionnaires.

- During data collection: due to defects in the interviewers work and incorrect declaration on the part of the informants.

- After collection of the data: errors in filtering, coding, recording, tabulation and printing of results.

2 Sample errors

The sample errors of the estimates of some of the main characteristics investigated are calculated quarterly.

To obtain sample errors the *reiterated semisamples* method is used.

This procedure consists of obtaining successive semisamples of the initial sample. From each semi sample the estimate of the characteristic from which we want to obtain the sample error is calculated. Once all estimates are calculated with each one of the semi samples as well as the estimate with the complete sample, the estimator of variance is given by:

$$\hat{V}(\hat{Y}) = \frac{1}{r} \sum_{i=1}^{r} (\hat{Y}_i - \hat{Y})^2$$

where:

r : is the number of semisamples obtained, this is the number of reiterations

$\hat{Y}_i$: is the estimate obtained with reiteration I

For each reiteration the general estimate process is repeated, in other words, the reweighting technique using the software CALMAR is applied.

$\hat{Y}$: is the estimate based on the complete sample

In the case of the APS the number of reiterations that are used is 40: Forming them is done in the following way:

a) All the sections of each strata were grouped by pairs, making sure that the two sections of each pair belong to the same APS rotation shift.

b) The first section of each for 20 reiterations and the other section for the other 20 are assigned randomly.

In this way each reiteration is made up of a number of sections equivalent to 50 per cent of the sample (semisample) and each section appears in half of the reiterations.

The relative sample error as a percentage (variation coefficient), obtained in the following way is published in the tables:

$$CV(\hat{Y}) = \frac{\sqrt{\hat{V}(\hat{Y})}}{\hat{Y}} . 100$$

3 Errors outside the sample

The study of errors foreign to the sample presents numerous difficulties due to the great variety of causes as well as the hypothesis on which the technical models are based which in general are not fulfilled in reality thereby leading to approximate results.

In the APS the analysis of errors foreign to the sample is based on the mathematical model developed by the United States Census Office due to Hansen, Hurwitz and Bershad, and that, functionally, consists of repeating the survey interviews of the sample of dwellings originally selected. Later the data obtained on both occasions is compared in order to investigate the inconsistencies and quantify the errors by means of the application of various quality indexes.

Separate from the *repeated interview* a specific study of those selected units that are worthy of survey but who refused to supply the data requested.

For these units which refuse to respond a *refusal questionnaire* is completed, which collects a series of basic characteristics like sex, age and the relationship with main person of the person who refuses to be interviewed, as well as the main persons age, sex and level of studies completed, employment status and branch of activity.

3.1 EVALUATION SURVEY

This survey is carried out biannually, in other words, the sample selected extends over two consecutive quarters.

The comparison of results obtained from the evaluation survey (Repeated interview, RI) with those obtained in the original interview (OI) permits the evaluation of two main types of errors foreign to the sample.

a) Coverage errors, produced by the omission or by the erroneous inclusion of units in the original survey.

b) Content errors, that affect the characteristics researched in persons worthy of survey.

The field work is carried out by specialised agents who carry out the interview repeated 15 days after the original, referring to the data from both interviews over the same period of time.

Bearing in mind the double objective followed with the evaluation survey, evaluating the quality of the results and controlling the work of the agents who work on the APS, the section sample in which the second interview is carried out is subdivided into two subsamples.

Subsample A: For this subsamples selection 26 itineraries, of approximately 10 blocks each, are formed with all the survey blocks, except those corresponding to Ceuta and Melilla. In each of the 26 weeks which the semester is divided into, one of these itineraries is randomly selected and the corresponding section of each block is visited.

Each week, within subsample A some 10 sections are investigated corresponding to different blocks. Since the itinerary selection is irreplaceable and contains al the blocks which make up the APS, ie: the OI agents, by the end of the semester one section from each interviewer has been investigated.

Subsample B: There are blocks formed which cover the peninsular scope of the survey for the selection of subsample B, in other words, eliminating those corresponding to island provinces and to Ceuta and Melilla, 81 zones of approximately three blocks each. Each week of the semester one of them is selected randomly with replacement, except for the weekly selection those which have some block selected in subsample A, equally visiting one section of each of the three blocks from the selected zone.

In total approximately 340 sections are investigated each semester.

In the sections selected the interview is repeated for half of the dwellings surveyed in the OI.

3.2 COVERAGE ERRORS

With the comparison of the results obtained in both interviews indicators are obtained on the coverage of dwellings, that of persons as well as indicators on content errors.

Coverage of dwellings: dwellings which are worthy of survey in both interviews are obtained, those worthy of survey in RI and not in OI and viceversa.

Coverage of persons: to study errors in the coverage of persons. These are classified as:

- Comparable persons, these are those that both agents have considered worthy of survey.

- Omitted persons are those whose data has been collected by the RI agent because they are considered worthy of survey, but on which no information exists in the OI.

- Persons erroneously included which feature in the OI but not in the RI due to the repeated interview agent considering they were not worthy of survey.

3.3 CONTENT ERRORS

The data on content errors is based on the information supplied in the two interviews by the persons classified as worthy of survey.

Thus for a characteristic C with the modalities M_1, ..., M_k, a person worthy of survey could be included in a table with the following format:

O.I / R.I.	Total	M_1	M_2	. . .	M_j	. . .	M_k
Everybody	**n**	**$n_{.1}$**	**$n_{.2}$**	**. . .**	**$n_{.j}$**	**. . .**	**$n_{.k}$**
M_1	**$n_{1.}$**	n_{11}	n_{12}	. . .	n_{1j}	. . .	n_{1k}
M_2	**$n_{2.}$**	n_{21}	n_{22}	. . .	n_{2j}	. . .	n_{2k}
. .	**. . .**	. . .	. . .	. . .	. . .	. . .	. . .
. .	**. . .**	. . .	. . .	. . .	. . .	. . .	. . .
. .	**. . .**	. . .	. . .	. . .	. . .	. . .	. . .
M_i	**$n_{i.}$**	n_{i1}	n_{i2}	. . .	n_{ij}	. . .	n_{ik}
. .	**. . .**	. . .	. . .	. . .	. . .	. . .	. . .
. .	**. . .**	. . .	. . .	. . .	. . .	. . .	. . .
. .	**. . .**	. . .	. . .	. . .	. . .	. . .	. . .
M_k	**$n_{k.}$**	n_{k1}	n_{k2}	. . .	n_{kj}	. . .	n_{kk}

n_{ij} represents the number of persons classified in modality M_I according to the RI and in M_j according to the OI.

The main diagonal (n_{ii}) represents the number of persons who have been identically classified in both interviews.

For each modality M_I of characteristic C the following table may be obtained:

R.I. \ O.I	With Modality M_I	Without Modality M_I	Total
With Modality M_I	a	B	a + b
Without Modality M_I	c	D	c + d
TOTAL	a + c	b + d	n

Comparing with the previous table we have the following equivalences:

$a = n_{ii}$ number of persons classified under M_I in both interviews.

$b = n_{I.} - n_{ii}$ number of persons classified under M_I in RI and differently in OI.

$c = n_{.I} - n_{ii}$ number of persons classified under M_I in OI and differently in RI.

$d = n - n_{I.} - n_{.I} + n_{ii}$ number of persons classified other than under M_I in both interviews.

$n = a + b + c + d$ is the total of persons who have been classified in both interviews with respect to characteristic C studied.

Based on these reduced tables the following quality indicators for M_I are defined:

a) Percentage of identically classified

$$P.I.C.(M_i) = \frac{a}{a + b} \times 100 = \frac{n_{ii}}{n_{i.}} \times 100$$

Varies between zero and a hundred. This is an indicator of the response stability. Its optimum value (100) expresses that all persons belonging - according to the RI - to the modality M_I will be classified in the same way in the OI.

b) Net index of change

$$I.C.N.(M_i) = \frac{c - b}{a + b} \times 100 = \frac{n_{.i} - n_{i.}}{n_{i.}} \times 100$$

Can be positive ($c > b$ or $n_{.I} > n_{I.}$) or negative ($b > c$ or $n_{I.} > n_{.I}$). This is an indicator of the lack of response, expressed as a percentage of the number of persons belonging to M_I according to the RI.

c) Net rate of difference

$$T.D.N.(M_i) = \frac{c - b}{n} \times 100 = \frac{n_{.i} - n_{i.}}{n} \times 100$$

Similar to the previous, but in this case it is a percentage with respect to the total of persons who have been classified in both interviews with respect to the reference characteristic.

d) Gross index of change

$$I.C.B.(M_i) = \frac{c + b}{a + b} \times 100 = \frac{n_{.i} + n_{i.} - 2n_{ii}}{n_{i.}} \times 100$$

Can be null or positive. It is an indicator of the variance of response.

e) Gross rate of difference

$$T.D.B.(M_i) = \frac{c + b}{n} \times 100 = \frac{n_{.i} + n_{i.} - 2n_{ii}}{n} \times 100$$

To compare the general quality of the different qualities evaluated the **global consistency index** is used, obtained from the table in which all modalities of characteristic C appear. This is defined as

$$I.C.G.(C) = \frac{\sum_{i=1}^{k} n_{ii}}{n} \times 100$$

The value which expresses the lack of error is 100.