INE
INSTITUTO NACIONAL DE ESTADISTICA

Economically Active Population Survey

Design of the Survey and Assessment of the Quality of the Information.

Technical Report

Madrid, October 2009
Revised edition: April 2012
Department for Sample Design and Result Assessment

Index

I.- Introduction

The Economically Active Population Survey (APS) is a continuous survey aimed at revealing the socio-economic conditions of the population, and has been carried out by the INE since 1964. The design of the survey is framed in the General Population Survey (GPS).

Certain aspects have been modified since its creation with a view to aiding the completion of the survey.

This report aims to collect the methodological aspects of the current design, as well as to assess the quality of the data contained within.

The INE welcomes any suggestions that could improve future editions of the survey.

II.- Survey Design

1 Objectives

The APS aims to identify the country's economic activity, analysing the human component. The design focuses on providing information on the main population categories in relation to the labour market as well as obtaining classifications of these categories according to different variables.

The different statistical sources (Census, Wage Surveys, Registered unemployment, etc.) that provide information on these issues cannot satisfy the objectives set out for this survey due to different reasons.

The Census is inappropriate because:

1) Its delayed periodicity prevents identifying the situation in inter-census periods.

2) Census data do not provide a detailed vision of the employment situation.

3) Data are obtained by autolisting, therefore in many cases the informant encounters difficulties when interpreting the concepts used.

The Industrial and Wages Surveys only provide information on the paid population.

Registered Unemployment and Social Security Registration do not provide homogeneous series, since the legal regulation that governs them is variable. Furthermore, they do not garner information on many of the variables investigated in the survey.

These elements justify the need for a continuous survey, designed and conceived expressly to identify the population's level of economic activity, alongside other characteristics that are closely linked to said activity.

The survey has been designed to provide detailed results on a national level. As regards Autonomous Communities and provinces, information is featured on the main characteristics, using the breakdown level that can be achieved using the estimators' variation coefficients.

This survey follows the definition for Economically Active Population agreed by the International Labour Organisation (ILO), establishing it as the *set of persons who, during an established reference period, furnish the supply of labour for the production of goods and economic services or are available to do so and carry out actions to incorporate themselves into said prod*uction.

In line with the aforementioned, the survey considers the economically active population to be composed by persons aged 16 years of age and over who in the reference week fulfil the conditions required to be included among employed or unemployed persons as defined for the survey.

2 Scope of the Survey

The survey covers a scope that can be broken down into these three sections:

2.1 POPULATION SCOPE

The Survey focuses on the population resident in main family dwellings, in other words, dwellings used all or most of the year as the regular or permanent residence.

The survey excludes so-called *group dwellings*, i.e. for example, hospitals, hotels, barracks, convents, etc.

It does include families that reside in these establishments but form an independent group, as can be the case of the managers of the centres, or the caretakers and porters. Theoretically, the survey only excludes populations lacking a family dwelling, which only represents 0.6 per cent of the total population according to the data from the 2001 Census.

2.2 GEOGRAPHICAL SCOPE

The survey is carried out in the whole country.

2.3 TIME SCOPE

The APS is a continuous survey that is performed every three months, with interviews undertaken during the thirteen weeks of each quarter.

It is necessary to make a difference between:

Reference period for the results of the Survey: quarter.

Reference period for the information collected: as a rule, this refers to the week (from Monday to Sunday) prior to the date the interview takes place. This week is called the *reference week* and all data should refer to this period, apart from the exceptions contained in the document entitled *Economically Active Population Survey. Survey description, definitions and instructions for completing the questionnaire*.

3 Survey Framework

In order to define the Survey framework, it is necessary to consider Spain's administrative division, which is as follows:

The country is divided into 17 Autonomous Communities and two autonomous cities, that compose the NUTS 2 (Nomenclature of Territorial Units for Statistics) endorsed by the European Parliament. These Autonomous Communities are subdivided into 50 provinces (NUTS 3), of which 47 are peninsular and 3 are insular. Provinces are subdivided into municipalities and the latter are subdivided into municipal districts.

Considering the aforementioned, the INE and the Councils subdivide these municipal districts into census sections.

These sections are used in all tasks undertaken by the INE that require infra-municipal division, among others for electoral purposes like *electoral sections*. In accordance with the Electoral Law, this requires that each section includes a maximum of 2,000 voters and a minimum of 500.

Therefore, the census section could be considered a geographical area with perfectly defined limits, whose population size is limited by the aforementioned conditions.

The section and number varies considerably over time and, therefore, it is updated on January 1st each year, coinciding with the revision of the Electoral Census and in each Census or Register. On the one hand, some sections have become uninhabited and have to be merged with others, and on the other, the opposite also occurs. That is to say, some sections have grown to the point that they exceed the established population limits and have to be divided.

4 Sample Design

4.1 TYPE OF SAMPLING. SAMPLING UNITS

The survey uses a two-stage sampling with first stage unit stratification.

First stage units are composed by **census sections**. The section sample always remains permanent with the following exceptions:

a) Sections in which all surveyable dwellings have been visited are removed from the sample.

b) When during the process for updating the sections (see point 5) some of the sections have to be removed from the sample, either due to probabilistic calculations, or due to changes in the allocation by strata.

In all cases, the sections that are removed from the sample are replaced by other sections selected randomly.

Second stage units are composed by main family dwellings (permanently inhabited) and permanent accommodations (shacks, caves, etc.). Secondary dwellings (inhabited only part of the year), or those that are for rent or sale, are not considered surveyable units as they are not part of the aforementioned population scope.

Sub-sampling is not carried out in second stage units, information is collected on all persons who regularly live in the same.

4.2 STRATIFICATION OF SAMPLING UNITS

First stage units are stratified following a double criteria:

A. Geographical criterion (for stratification)

Sections are grouped in strata in each province, in accordance with the demographic relevance of the municipality they belong to.

B. Socio-economic criterion (for substratification)

Census sections are grouped in substrata in each strata, according to the socio-economic conditions of the same.

4.2.1 Strata

The following types of municipalities are considered to establish strata formation:

1. Self-represented municipalities: Municipalities that, given their category in the province, must always have sections in the sample.

Self-represented municipalities are:

The province capital.

Municipalities that given the number of inhabitants are allocated at least 12 sections in the sample given the proportional allocation in the province.

Municipalities that have a notable demographic situation in the province, when there are no other similar municipalities to group them with, although proportionally they are allocated less than 12 sections in the sample.

2. Co-represented 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 line with this classification, in general, the theoretical strata considered correspond to these concepts:

Stratum 1: Province capital municipality.

Stratum 2: Self-represented municipalities, important areas in comparison with the capital.

Stratum 3: Other self-represented municipalities, important areas compared 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 under 2,000 inhabitants.

It is important to consider that given the different sizes of the municipalities in the different provinces, not all stratification is uniform. For example, in the province of Alicante, strata 9 disappears as there is not enough population to compose it, and therefore municipalities with less than 2,000 inhabitants are included in strata 8. Conversely, the province of Burgos has over 350 municipalities with less than 2,000 inhabitants, included in stratum 9. Nevertheless, stratum 7 and 8 are grouped in stratum 7 since there are hardly any municipalities with between 2,000 and 5,000 inhabitants.

Every ten years, the definition of the strata in each province is updated using the information from the Population Censuses.

4.2.2. Substrata

Two groups of sections are considered when creating the substrata in each stratum:

a- Sections from strata 7, 8 and 9. This group of sections from small municipalities presents a relatively low variability compared to the objective variables, which is in any case well explained by the territory they belong to. Therefore, as their substrata they take the region (LAU1-Local Administrative Units) of the municipality they belong to. Consequently, by doing so, as well as distributing the sample in homogeneous groups, the sample representation of the territory will allow the survey to obtain more broken down estimates in the future.

b- Other sections. These sections are grouped in their strata by implementing cluster analysis techniques. In this case, as they are larger municipalities, which more or less practically guarantee the sample representation of the region (LAU-1) they belong to, priority has been given to the use of the auxiliary information available to form homogeneous groups of sections and, therefore, improve the precision of the estimates.

The auxiliary information used to perform the analysis in this second group of sections stems from the 2001 Census and the State Tax Administration Agency

(AEAT). The characteristics selected are the most correlated to the variables under study in the Economically Active Population Survey.

The following auxiliary variables are used at the section level:

Percentages of unemployed persons

Percentage of inactive persons

Percentage of employed persons

Percentage of foreigners

Percentage of persons between 0 and 19 years old

Percentage of persons between 15 and 24 years old

Percentage of persons aged 65 years old or more

Percentage of persons with level of studies, 1, 2 or 3 according to the classification of the 2001 Census, i.e. illiterates, without studies or first grade level of studies

Percentage of persons with level of studies 4, 5, 6 or 7, i.e. SOE, GBE, secondary education diploma, VT

Percentage of persons with level of studies 8, 9 or 10, i.e. diploma, graduate degree or Ph.D.

The survey also considers the 18 variables relating to employed population percentages in the section according to their socio-economic condition, which are classified as follows:

01 Agricultural businesspersons with wage earners

02 Agricultural businesspersons without wage earners

03 Members of agricultural cooperatives

04 Directors and heads of farms

05 Other agricultural workers

06 Professionals, technicians and the like that exercise their activity as freelance workers

07 Non agricultural businesspersons with wage earners

08 Non agricultural businesspersons without wage earners

09 Members of non-agricultural cooperatives

10 Directors and managers of non agricultural establishments and senior civil servants

11 Professionals, technicians and the like that exercise their activity employed by others

12 Heads of administrative, commercial or non-agricultural corporate services departments

13 Other administrative and sales personnel

14 Other service personnel

15 Non-agricultural foremen

16 Qualified and specialised non-agricultural operators

17 Non-specialised non-agricultural operators

18 Armed Forces Professionals

Finally, the following tax variables were used:

Overall income per dwelling with recipients

Livestock and real estate capital income regarding to overall income.

Agricultural income regarding the overall income

(These last variables have not been implemented in the País Vasco)

Prior to the cluster analysis, variables have been standardised in each stratum with average 0 and standard deviation 1, except for the variables for the percentage of unemployed persons, percentage of youths and the three fiscal variables, that have been standardised with standard deviation 2. This aims to ensure the latter variables have a greater weighting than the rest, and therefore more influence in the formation of the substrata.

Furthermore, cluster analysis has not been performed for variables that represent less than 1 per cent of the total of the stratum in each province, in order to avoid tiny substrata.

The survey uses an accumulative algorithm that obtains hierarchical clusters. The survey stems from as many clusters as there are sections and, in each stage the most similar clusters, i.e. with a minimum distance, are united. In the end, all sections form a single cluster. Finally, the intermediate point of the grouping process is established, in terms of the number and size of the clusters considered most appropriate.

The distance between sections is Euclidean and defined using standardised variables as explained previously. Between clusters, the distance is measured using the Ward method, which tends to produce clusters with a similar number of sections.

This procedure has been performed implementing the CLUSTER procedure of the SAS SAS/SAT module.

4.3 SAMPLE SIZE

When the survey was implemented, the size was established via the application of a procedure of minimum variance for fixed cost. The base was a budget (Q) which was used to determine the number of sections (n) and the number of dwellings (m) that minimise the variance of the estimates. This was performed using a lineal cost function and the expression of the variation coefficient for a proportion in the sampling of clusters with subsampling.

The following cost function was observed:

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

where:

Q = Total budget

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 P established by

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

in which:

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

and δ the intraclassical correlation coefficient, which for the active population was calculated and equals 0.05.

The minimum of the expression $CV^2(\hat{P})$ as regards variables m and n is obtained calculating the minimum of expression F (δ, m, n) that is independent from $\hat{P}$.

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

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

obtain different values for F (δ, m, n).

The minimum value for F (δ, m, n) as regards m and n corresponded to m=20 and n=3,000.

Considering this result, the sample established a total of 3,000 sections, researching an average of 20 dwellings per section.

Subsequently, the sample has been extended several times with a view to fulfilling the different European Union requirements and improve the representation of the most broken down areas. As of the first quarter of 2005, a sample size of 3,588 sections and 18 dwellings per section is established, except in the provinces of Madrid, Barcelona, Sevilla, Valencia and Zaragoza, in which the number of interviews per section rises to 22. In the third quarter of 2009, an agreement was signed with the Autonomous Community of Galicia, increasing the sample in this Autonomous Community to a total of 468 sections, assigning separate strata to the municipalities of Santiago de Compostela and Ferrol. **The size of the final sample for the national total stands at 3822 sections.**

4.4 ALLOCATION

This section includes the criteria followed for the distribution of sample sections between the provinces, in the province between strata and in these between substrata.

The following aspects were considered when performing the provincial allocation:

a) National results have to be as reliable as possible. In this sense it is important to recall that, in general, the further one is from the ideal allocation by provinces and strata, the greater the loss of precision in the national estimate for a fixed size of the sample.

b) In each province there should be minimum sample size that will enable estimates of the same.

c) In each province, the number of sections must be a multiple of thirteen. This facilitates the distribution of the sample between the weeks of the quarter. In order to make the three aforementioned conditions compatible, a compromise allocation has been adopted between the uniform and the proportional method, after grouping provinces with a similar demographic importance and allocating them 3 to 12 interviewers, that is to say from 39 to 156 sample sections (with the exception of Ceuta and Melilla, which given their small population size only have one agent and therefore there are 13 sample sections in each).

In each province, the allocation between strata is proportional to the size of each one of them, although strata with the largest municipalities have been strengthened, since it is expected that most of the characteristics analysed are correlated with the social-economic and cultural levels of the inhabitants, and it is precisely in these strata where, in general, dispersion should be greater and the cost per interview is lower.

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

Table 1 shows the distribution of the sample sections by provinces and strata.

Distribution of the section samples by provinces and strata

	Strata									
Provinces	1	2	3	4	5	6	7	8	9	Total
02 Albacete	18				8		3	5	5	39
03 Alicante/Alacant	18	10		15	12	12	8	3		78
04 Almería	16			9		5	3	6		39
01 Araba/Álava	30					3	6			39
33 Asturias	30	33		9	23	19	7		9	130
05 Ávila	13						3	8	15	39
06 Badajoz	20			6	10	6	15	11	10	78
07 Balears, Illes	44				27	12	12	9		104
08 Barcelona	55		33	19	21	12	10	3	3	156
48 Bizkaia	29	7		9	15	8	4	6		78
09 Burgos	20				7		3		9	39
10 Cáceres	19				7	5	10	12	25	78
11 Cádiz	13	13	6	26	7	6	7			78
39 Cantabria	35	10			8	11	9	9	9	91
12 Castellón/Castelló	26				26	6	4	7	9	78
13 Ciudad Real	13	9			14	15	12	8	7	78
14 Córdoba	34				13	9	11	11		78
15 Coruña, A	42	14	12		24	26	26	12		156
16 Cuenca	10						8	6	15	39
20 Gipuzkoa	26			6	13	20	7	6		78
17 Girona	15				19	12	10	13	9	78
18 Granada	28			5	5	12	12	10	6	78
19 Guadalajara	20				4			6	9	39
21 Huelva	15					13	3	8		39
22 Huesca	13					10	6		10	39
23 Jaén	17	7			15	12	12	15		78
24 León	24	10				10	15		19	78
25 Lleida	15					5	3	6	10	39
27 Lugo	26					14	14	24		78
28 Madrid	92		30	15	9	4	6			156
29 Málaga	36			10	18	5	9			78
30 Murcia	36	18		6	26	12	6			104
31 Navarra	36				9	9	8	15	14	91
32 Ourense	30					12	10	14	12	78
34 Palencia	20						5	5	9	39
35 Palmas, Las	44			9	28	14	9			104
36 Pontevedra	18	52			24	36	16	10		156
26 Rioja, La	33					9	7	7	9	65
37 Salamanca	20					4	3		12	39
38 Santa Cruz de Tenerife	25	15			24	10	10	7		91
40 Segovia	16						5	3	15	39
41 Sevilla	52			11	20	18	10	6		117
42 Soria	18						8		13	39
43 Tarragona	19	12			12	9	9	9	8	78
44 Teruel	10					4	9		16	39
45 Toledo	13	13				7	11	21	13	78
46 Valencia/València	45			10	24	19	7	7	5	117
47 Valladolid	36					4	6		6	52
49 Zamora	16					4			19	39
50 Zaragoza	59					4		6	9	78
51 Ceuta	13									13
52 Melilla	13									13
Total	1,384	223	81	165	472	447	397	314	339	3,822

4.5 SELECTION OF THE SAMPLE

The sample selection has been performed to ensure that in each stratum all family dwellings have the same probability of being selected, in other words, to ensure there are **self-weighted samples in each stratum.** This type of samples provide equal design weights by stratum in the estimators. For this, first stage units (census sections) are selected with a probability proportional to the number of main family dwellings, according to data from the last Census or Continuous Register. In each section selected in the first stage, a pre-set number of family dwellings with the same probability is selected by implementing a random start systematic sample. For this survey, 18 dwellings have been selected per section (see section 4.3)

Therefore, the probability of selection of each dwelling i, belonging to section j of stratum h, where K_h sections have been allocated, would be:

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

in which:

$P\,(S_{jh})$ = Probability of selection of section j in stratum h

$P\,(V_{ijh}/S_{jh})$ = Probability of selection of dwelling i conditioned by the selection of section j.

V_{jh}= Total dwellings in section j

V_h = Total dwellings in stratum h.

This probability does not depend on i or j, in other words, it does not depend on the dwelling or the section, and therefore the section is self-weighted.

4.6 DISTRIBUTION OF THE SAMPLE IN TIME

The distribution of the sample is uniform over time, which means there is a constant number of sections per week in each province.

Furthermore, the distribution of sample sections by province, stratum and week is homogeneous, as by province, rotation shifts (see section 4.7) and week. 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 whole of the sample is divided into three independent subsamples, each one, representative of all the population.

4.7 ROTATION SHIFTS

As in the previous paragraph, each period of the survey is a quarter, which is repeated successively.

The census sections are set in the sample indefinitely (except for the exceptions in section 4.1). Nevertheless, family dwellings are renovated partially every quarter of the survey, in order to avoid tiring the families. This renovation is performed in a sixth part of the sections.

Therefore, the total sample is divided into six subsamples, that are called *Rotation shifts*. Each section is identified by a five digit code. The last digit expresses the corresponding rotation shift, which is numbered from 1 to 6.

Each quarter, the dwellings in the sections of a specific rotation shift are renewed. Thus, each dwelling remains in the sample for six consecutive quarters. After this period, it is removed from the sample and replaced by another from the same section.

In order to be able to renew the dwellings appropriately, each quarter the framework of dwellings in the sections of the rotation shift whose dwellings are interviewed for the sixth and final time is updated. Consequently, in the following period, the sample can include dwellings, both newly constructed and those that have become family dwellings, which, did not exist or were uninhabited or used for other purposes other than the main dwelling when the last Census or Register was performed.

These dwellings are included in the sample with the same probability as the original dwellings in the section.

The distribution of the number of sections per stratum and week is similar in each rotation shift. Thus, it is a case of avoiding possible measurement bias errors in the estimates, due to the different behaviour of the families taking part in terms of the time they have been in the survey.
Each rotation shift can, therefore, be considered as a representative subsample. This makes it easier to obtain estimates using structural variables by means of joining said subsamples.

4.8 ESTIMATORS

Up until year 2001, the survey used **ratio estimators**, taking the figures of the population resident in main family dwellings, deduced from the Population Now Cast compiled by the INE as the auxiliary variable. The expression of the estimator for a specific characteristic Y in a certain quarter of survey is as follows:

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

the sum h is extended to the strata of a province, an autonomous community or the national total, in which:

P_h: is the population resident in family dwellings, in stratum h, referred to half the quarter.

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

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

y_{hi}: is the value of the characteristic researched in dwelling i-th, of stratum h.

As of the first quarter of 2002, **Reweighting techniques** are applied to estimators so as to adjust the survey estimates to the information from external sources.

The reweighting technique involves the following:

A population is considered $U = \{u_1.......u_N\}$ to extract a sample

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

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 characteristic researched in sample unit K.

d_k: Raising factor for unit K obtained using the expression $\frac{P_h}{p_h}$, h being the stratum to which the unit belongs.

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

There are J auxiliary variables, whose values are known for the sample and whose totals are known for the population.

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

The purpose is to find a new estimator

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

in which the new weights w_k fulfil the following conditions:

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

- Similar to initial weights d_k
- Verify the balance equation

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

The problem aims to find values w_k that minimise the expression:

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

in which:

G = Function of distance.

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

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

The solution of the problem depends on the function of distance G 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 using Lagrange multipliers which facilitate obtaining a set of factors w_k that verify the balancing conditions and provide the same estimates as the generalised regression estimator.

In the specific case of the APS, the linear distance function is implemented in a truncated version (to avoid negative solutions when using the equation system), in order to maximise the properties of the regression estimator, with a small variance and minimum bias errors.

The survey uses the following auxiliary variables:

- Population aged 16 years and over by age groups and sex by Autonomous communities (22 groups).
- Population aged 16 years and over by autonomous community and nationality, Spanish or foreign.
- Population aged 16 years and over by province
- Population under 16 years old by age groups and sex (6 groups) by Autonomous Community
- Population under 16 years old by province

Therefore, the estimators used at present in the APS present a correct estimate of the population by age group and sex and the total number of Spaniards and foreigners aged over 16 years old by autonomous communities.

In order to obtain the practical solution for this problem, the survey uses CALMAR (CALage sur MARges) software, programmed by the INSEE (Institut National de la Statistique et des Études Économiques) in France.

5 Updates in the framework of the sample

The constant population variations either as regards the characteristics or spatial distribution require updates in the framework that necessarily have a bearing on the sample structure.

The APS framework considers four types of updates:

Updates in the sample sections framework, as a consequence of modifications (see section 5.1) caused by different incidents such as partitions, merges or variations of the limits of the selected sections. In each of these cases, it is necessary to determine the probability of selection of the new sections, as well as the number of interviews to be performed in the same.

Updates in the housing framework, restricted and exclusive to sample sections. This update, as aforementioned in section 4.7., aims to incorporate main dwellings *newly registered* in the section in the list of dwellings.

Update of the selection possibilities of the selected persons. This aims to, undertaking as few changes as possible, ensure the section sample is equal to a sample selected in the year of the update. Performed every three years.

General update concerns al sections and dwellings. The definition of strata and substrata is revised and the probability of selection of the section is updated. Performed using information from the Population Censuses (See section 5.2.2).

5.1 MODIFICATIONS IN SECTIONS OF THE SAMPLE

The following cases are considered:

5.1.1Partition of sections

Refers to a section S in which the increase of the number of main dwellings requires a division into parts S_1, S_2 ... S_K, either to form new sections or to become part of other pre-existing ones.

This considers the problem of determining the selection probabilities for new sections in order to know which will remain in the sample, as well as the number of dwellings to be interviewed in the same to ensure the sample is still self-weighted.

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 part 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.

Thus, the sample continues to be self-weighted.

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

In this case:

1) One of the fragments is selected with probability proportional to its size according to the last Census V_{sj} / V_s and the new section S'_j where the part has been incorporated will be automatically selected.

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

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

in which

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

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

5.1.2 Merging of sections

Due to migration and natural movements of the population, some sections become empty or uninhabited. They are, therefore, merged with others, to ensure that if they are selected, there will be sections to research.

The section-merging is a particular case of the partition analysed in section 5.1.1.B.

Therefore, if section S_j selected merges with another to form a new section S, the latter is automatically included in the sample and the number of dwellings to be interviewed is:

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

in which

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

V_S = Number of main dwellings, according to the last Census, in the limits of new section S.

5.1.3 Variation of limits

This is the case of a section formed with fragments from two or more sections due to a readjustment of the limits.

To calculate the selection probability, this case can be considered as a process consisting of two stages: the first involves the partition of each section and the second the appropriate merging of the sections resulting from the partition.

In all the aforementioned cases, the new sections are incorporated into the sample when due to *Rotation shifts* the families in the sections affected by said incidents are renewed.

5.2 RENEWAL OF THE SAMPLE AFTER UPDATING SELECTION PROBABILITIES

When information is available from the electoral files, Population Censuses or the Continuous Register, section probabilities are updated and the number of interviews is adjusted to 18 per section.

Changes in the sample of sections as a consequence of the update are included in the same by rotation shift, that is to say, during a period of six quarters, as occurs in the case of the renewal of dwellings. Consequently, and with a view to providing certain stability in the survey's time series, the updates of the section probabilities are performed every two or three years.

The more direct manner of updating the selection probabilities is the selection of a new sample using the most updated available framework. Yet such a radical change in a continuous survey, such as the APS, causes three types of problems:

- Loss of essential information for selection and visiting the dwellings selected in the second stage. This information, which has to be recompiled, includes tangible aspects such as the directories of dwellings or the planimetry of the area, and intangible aspects that are also very important, such as the whether or not the population in the section know the interviewer, since this fact makes it easier to access families and reduces non-response notably.
- Loss of precision in estimates for interannual quarterly variations, since this reduces the common sample between both periods considerably.
- Possible lack of continuity in the sample's time series, due to the causes mentioned in the previous section.

Therefore, it was decided to arbitrate a procedure that, without distorting the selection probabilities that actually correspond to each section, maintains the sample of sections with minimum variations.

This considers two types of updates of the selection probabilities in terms of the information available for the same.

5.2.1. Updates performed using information taken from the Continuous Register.

In this case this does not modify the definition of the strata and maintains the one preset for each municipality, even though the population may have changed and exceeded the limit of the lower or higher stratum. In order to update the information, the survey uses the procedure proposed by Kish and A. Scott(JASA 1971).

When S is a section from stratum h, whose selection probability in the previous update(t-1) was given by:

$$Ps = \frac{V_s}{V_h} = \frac{Dwellings\ in\ \sec tion\ S\ in\ (t-1)}{Dwellings\ in\ stratum\ h\ in\ (t-1)}$$

and when at the moment of updating(t), the corresponding selection probability is defined by

$$P's = \frac{V_s'}{V_h'} = \frac{Dwellings\ in \sec tionS\ in\ (t)}{Dwellings\ in\ stratum\ h\ in\ (t)}$$

P_s is compared with P'_s with one of the following cases being possible:

1) If $P'_s > P_s$ section S remains in the sample with probability P'_s, since if it was selected with a probability P_s lower than the probability corresponding at present, there is greater reason for it to have been selected implementing the current probability P'_s.

2) If $P'_s < P_s$ the section remains in the sample with probability P'_s/P_s and is removed from the sample with probability $1 - P'_s / P_s$.

This criterion will cause the removal of certain sections from the sample. These will be replaced by others sections from the same stratum, selected among **those that have increased their probability and did not belong to the sample.**

This criterion maintains the diagram that proves that the probability of a section belonging to the sample is in fact the correct probability, in other words, it is proportional to the current number of dwellings.

5.2.2.Updates performed using information taken from the Population Census.

As this information is more comprehensive, definitions of strata and substrata are revised, and each municipality is given the corresponding allocation in terms of the new population figures.

In view of the former, many strata modifications take place and the Kish-Scott procedure is too complex and does not guarantee ideal results, since it does not prove that the least number of modifications are undertaken.

Therefore, the survey uses the method proposed by J. M. Brick, R. Morganstein and CH. L. Wolter(Westat 1987), based on the Kish and Scott method mentioned in the previous section.

The following expressions are the probabilities of Section 'S' of belonging to the sample in the last update and in the new one, respectively:

$$\pi_{hs} = n_h * \frac{V_s}{V_h} \qquad \pi'_{h^*s} = n'_{h^*} * \frac{V'_s}{V'_{h^*}}$$

where n_h and n'_{h^*} are the number of sections allocated by stratum in 't-1' and in 't', and in strata h and h* respectively. Therefore:

- Ifπ'_{h*S} is greater than π_{hS} and the section is in the sample, it remains in it.
- If π'_{h*S} is greater than π_{hS} and the section is **not** in the sample, it will enter the sample with probability:

$$\left(\pi'_{h*s} - \pi_{hs}\right) / \left(1 - \pi_{hs}\right)$$

- If π'_{h*S} is less than π_{hS} and the section was in the sample, it will remain in the same with probability:

$$\pi'_{h^*s} \ / \pi_{hs}$$

- If π'_{h*S} is less than π_{hS} and the section was not in the sample, there is no possibility of entering the same.

This shows that the probability of a section s to belong to the sample is π'_{h*S}, in other words, the possibility updated in t in the new stratum.

The main characteristic of this algorithm is the fact that it is quite simple to apply in quite complicated situations. On the contrary, it presents the inconvenience that it does not provide a sample with a set size by stratum which makes it necessary to perform one last adjustment, removing remaining sections with equal probability and selecting the missing sections with probability proportional to their size.

III. Assessment of the quality of the information

1 Introduction

The errors that usually affect surveys can be divided into two large groups:

Sampling errors, caused by obtaining results on the features of a population, using the information garnered in a sample of the sample.

Non-sampling errors, which are common to all statistical researches, both if the information is garnered by sampling or after performing a Census. These errors appear in any stage of the statistical process:

- Before collecting the data: due to framework deficiencies and inadequacies in the definitions and questionnaires.

- During the collection of data: caused by defects in the task of the interviewers and incorrect statement from the informers.

- After collecting the data: errors in filtering, coding, recording, tabulating, etc. the results.

2 Sampling errors

Sampling errors are calculated quarterly for the estimates of some of the main characteristics investigated.

The *successive semisamples* method is used to obtain sampling errors.

This procedure consists in obtaining successive semisamples from the initial sample. Each semisample is used to calculate the estimate of the characteristic for which the sampling error is to be obtained. After calculating all estimates with each of the semisamples, and the estimate with the whole sample, the variance estimator is established 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, that is the number of repetitions

$\hat{Y}_i$: is the estimate obtained with the i-th repetition

The general estimate process is performed for each repetition, i.e. the reweighting technique is implemented using CALMAR software.

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

In the APS the number of repetitions is 40. The following steps were taken to form them:

a) All sections in each stratum were grouped in pairs, trying to ensure both sections in each pair were from the same APS rotation shift.

b) Randomly, the first section of each pair was given 20 repetitions and the other section the other 20.

Therefore, each repetition is composed by a number of sections equivalent to 50 per cent of the sample (semisample) and each section appears in half of the repetitions.

The tables include the *corresponding sample error as a percentage (variation coefficient*), as established in the following expression:

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

3 Non-sampling errors

The analysis of non-sampling errors presents a great number of difficulties due to the vast amount of reasons that cause them, as well as the hypotheses on which the theoretical models are based which, in general, are never fulfilled. This aspect leads to obtaining approximate results.

In the APS, the analysis of non-sampling errors is based on the mathematical model created by the United States Census Office, thanks to Hansen, Hurwitz and Bershad, which, operationally, is based on repeating the survey interviews in a subsample of the dwellings selected originally. Subsequently, data are compared with those obtained in the previous occasions, with a view to researching inconsistencies and quantifying errors using different quality indices.

Apart from the *repeated interview*, the selected units that are surveyable but refused to facilitate the requested data are studied specifically.

For units that refused to take part in the survey, a *Refusals questionnaire,* is compiled, collecting a series of basic characteristics, like sex, age and relationship of the person refusing the interview with the main person, as well as age, sex, nationality, finished studies, economic activity, branch of activity and occupation of the main person.

3.1 EVALUATION SURVEY

This survey has two objectives:

- Control the task of collecting information in all autonomous communities.
- Assess the quality of the results.

Comparing results in the evaluation survey (repeated interview, ER) with those obtained in the original interview (EO) allows the survey to assess two major types of non-sampling errors:

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

b) Content errors, that affect the features investigated in surveyable persons.

Fieldwork is carried out by specialised agents, who perform the repeated interview up to three weeks after the original, with both interviews referring to the same period of time.

The fact that over 70 per cent of the first time refusals are produced in the first interview with the families, alongside the existence of technical difficulties that hamper the realisation of the evaluation survey (ER) with CATI, have determined the fact that ERs will only investigate **sections that are in the first interview in the EO**. The collection method used in these sections, both EO and ER, is CAPI.

As a consequence of the aforementioned, there is less sample material in the evaluation survey, compared to previous years; therefore the four quarterly samples will be grouped to offer results annually, in order for them to be more representative.

Four zones have been created for the quarterly selection of the sample for the evaluation survey, with each one grouping several Autonomous Communities, so that each of these are included in only one of the zones.

Each week, the sections of the sample (in the first interview) in one of the zones are investigated, with a random allocation of weeks and zones, so that each of them is researched at least in three weeks of the quarter.

Consequently, approximately between 130 and 150 sections are investigated each quarter.

In the sections selected, the interview is repeated in half of the dwellings. The ER uses a slightly reduced questionnaire, compared to the EO, i.e. with a few questions less.

This procedure is used to investigate a number of dwellings between 1,300 and 1,500, which represents approximately 2 per cent of the APS sample.

As well as the evaluation survey, and so as to detect errors committed when updating the sample sections, each quarter a sample of 50 sections is selected (one from each province, except Ceuta and Melilla) to asses the quality of the updates.

3.2 COVERAGE ERRORS

The comparison of the results obtained in both interviews provides indicators regarding the coverage of dwellings and persons, as well as indicators of content errors.

Coverage of dwellings: provides dwellings that are surveyable in both interviews, surveyable in ER and not in EO and vice versa.

Coverage of persons: to analyse these errors, persons are classified into:

- Suitable persons, those who both agents have considered surveyable.

- Omitted persons, those whose data has been collected by the ER agent after considering them surveyable, but with no information from the EO.

- Persons included mistakenly, those who appear in the EO but not in the ER, since the agent performing the repeated interview thought they were not surveyable.

3.3 CONTENT ERRORS

Data on content errors are based on the information supplied in both interviews by suitable persons.

Two types of tables are created to facilitate data analysis: consistency tables and quality indicator tables.

Thus, for a characteristic C with modalities M_1, ..., M_k, the consistency table will be as follows:

E.O. E.R.	Total persons	M_1	M_2	. . .	M_j	. . .	M_k
Total Persons	**n**	$\mathbf{n}_{.1}$	$\mathbf{n}_{.2}$	**. . .**	$\mathbf{n}_{.j}$	**. . .**	$\mathbf{n}_{.k}$
M_1	$\mathbf{n}_{1.}$	n_{11}	n_{12}	. . .	n_{1j}	. . .	n_{1k}
M_2	$\mathbf{n}_{2.}$	n_{21}	n_{22}	. . .	n_{2j}	. . .	n_{2k}
. .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
. .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
. .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
M_i	$\mathbf{n}_{i.}$	n_{i1}	n_{i2}	. . .	n_{ij}	. . .	n_{ik}
. .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
. .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
. .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
M_k	$\mathbf{n}_{k.}$	n_{k1}	n_{k2}	. . .	n_{kj}	. . .	n_{kk}

n_{ij} represents the number of persons classified in modality M_i according to ER and in M_j according to EO.

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

Each modality M_i of the characteristic C provides the following reduced table:

E.O. E.R.	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 this with the previous table, provides the following equivalences:

$a = n_{ii}$ number of persons classified in modality M_i in both surveys.

$b = n_{i.} - n_{ii}$ Number of persons in modality M_i in ER and in another in EO.

$c = n_{.i} - n_{ii}$ Number of persons in modality M_i in EO and in another in ER.

$d = n - n_{i.} - n_{.i} + n_{ii}$ Number of persons classified in a different modality to M_i in both interviews.

$n = a + b + c + d$ Total number of persons classified in both interviews compared to the characteristic C under study.

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

a) Percentage classified identically

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

Varies from zero to one hundred. This is an indicator of response stability. The optimal value (100) expresses that all persons who according to the ER belong to modality M_i obtained the same classification in the EO.

b) Net change index

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

This element can be positive ($c > b$ o $n_{.i} > n_{i.}$) or negative ($b > c$ o $n_{i.} > n_{.i}$). Indicator of the response bias error, expressed as the percentage of the number of persons classified in M_i according to ER.

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 option, but in this case the percentage refers to the total number of persons classified in both interviews compared to the reference characteristic.

c) Gross change index

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

This element can be non-existent or positive. Indicates the response variance.

e) Gross rate of difference

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

Similar to the previous option, but refers to the total number of persons classified in both interviews compared to the characteristic under study.

To compare the general quality of the different characteristics assessed, the survey uses the **global consistency index**, which for each characteristic C is obtained using the table which includes all the modalities of the same. Defined as

$$I.C.G.(C) = \frac{\sum_{i=1}^{k} n_{ii}}{n} x100$$

An I.C.G. value = 100 indicates the lack of classification errors.