INSTITUTO NACIONAL DE ESTADISTICA

Labour Force Survey

Sample Design and Evaluation of Data Quality

Technical Report

Madrid, June, 2016
Department of Survey Sample Design

Index

I. Introduction

The Labour Force Survey (LFS) is a continuous survey aimed to investigate the socio-economic characteristics of the population, and has been carried out by the INE since 1964.

Since the beginning, the survey has undergone several changes, always aimed to improve the information provided.

This report is intended to reflect the methodological aspects of the current design and evaluation of data quality contained within.

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

II. Survey Design

1 Objectives

The main objective in LFS is the knowledge of economic activity in the country, with regard to 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 topics cannot satisfy the objectives of this survey due to different reasons.

The Census is inappropriate because:

1) Its long periodicity prevents us from knowing the situation in intercensal periods.

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

3) Data are obtained by self-classification, so there are, in many cases, difficulties of interpretation of the concepts used by the informant.

The Industrial and Wages Surveys only provide information on the wage-earning population.

Registered Unemployment figures published by the Public Employment Service (Servicio de Empleo Público Estatal, SEPE) and Social Security Affiliation do not provide homogeneous series, since the legal regulation that governs them is variable. Furthermore, they do not collect 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 degree of economic activity of the population, together with other characteristics closely related to that activity.

The survey has been designed to provide detailed results at national level. For Autonomous Communities and provinces, information is offered only for the main characteristics at the level of disaggregation that the coefficients of variation of the estimators allow.

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, supply labour for the production of economic goods and services or are available to do so and carry out actions to incorporate themselves into said prod*uction.

According to the above, this survey considers that economically active population consists of persons aged 16 and over who in the reference week satisfy the conditions necessary for inclusion among the employed or unemployed persons according to the definitions given 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 is aimed at the population living in family dwellings, ie those used all or most of the year as a 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 0.9 per cent of the total population according to 2011 Census data.

2.2 GEOGRAPHICAL SCOPE

The survey is conducted throughout the national territory

2.3 TIME SCOPE

The LFS is a continuous survey on a quarterly basis, extending the interviews along the thirteen weeks of the quarter.

It is necessary to make a difference between:

Reference period for the results of the survey: the quarter.

Reference period for the information collected: as a rule, this refers to the previous week (from Monday to Sunday) at the date the interview takes place. This week is called the *reference week* and all data must be referred to this period, with the exceptions listed in the document *Labour Force Survey. Survey description, definitions and instructions for completing the questionnaire*.

3 Survey Framework

In order to define the sampling frame of the survey, it is necessary to consider the administrative division in Spain, which is as follows:

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

From the mentioned before, INE with the Councils make a new subdivision of the 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 conditions mentioned above.

The sections and their number vary considerably over time, so it must be up-to-date continuously. On the one hand, some sections come to have few inhabitant and have to be merged with others, and on the other, the opposite also occurs ie, sections that grow to a point that exceed the established population limits and have to be divided.

In section 5 of this document, it is analysed in detail how these updates affect the sample and its treatment.

4 Sample Design

4.1 TYPE OF PROBABILITY SAMPLE DESIGN. SAMPLING UNITS

A two-stage random sampling design with stratification of the first stage units is used.

The Primary Sampling Units (PSUs) are **census sections**. The PSUs sample remains permanently in the survey with the following exceptions:

a) Sections in which all dwellings to survey have already been visited.

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

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

Second stage units are the main family dwellings (inhabited permanently) and fix accommodations (shacks, caves, etc.). Secondary dwellings (inhabited only part of the year), or those that are for rent or sale, are not considered units to survey as they are not part of the population scope defined previously.

Subsampling is not carried out in second stage units, information is collected from all persons who regularly live in it.

4.2 STRATIFICATION OF SAMPLING UNITS

Primary sampling units are stratified following a double criteria:

A. Geographical criterion (stratification)

Sections are grouped by strata within each province, according to the demographic relevance of the municipality which they belong to.

B. Socio-economic criterion (of substratification)

Census sections are grouped by substrata within each strata, according to the socioeconomic characteristics of the section itself.

4.2.1 Strata

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

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

Self-represented municipalities are:

The capital of the province.

Municipalities that have a number of people such that in the proportional allocation within the province, they have at least 12 sections in the sample.

Municipalities that have an important demographic situation in the province, there are no other similar municipalities to group them with, although the corresponding proportional allocation is 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.

According to this classification, in general, the theoretical strata considered are:

Stratum 1: Province capital municipalities.

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

Stratum 3: Other self-represented municipalities, important areas in comparison with 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 size distribution municipalities between different provinces, the stratification is not uniform among them. For example, in the province of Alicante, the stratum 9 disappears as there is not enough population to compose it, and therefore municipalities of less than 2.000 inhabitants are included in stratum 8. By contrast, the province of Burgos has over 350 municipalities with less than 2.000 inhabitants which are included in stratum 9, and however stratum 7 and 8 are grouped in stratum 7 since there are very few municipalities with between 2.000 and 5.000 inhabitants.

Every ten years, with the information from the Census of Population, the definition of strata in each province is reviewed and updated.

4.2.2. Substrata

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

a- **Sections from strata 7, 8 and 9**. It is considered that this group of sections, belonging to small municipalities, presents a relatively small variability with respect to the target variables and in any case this variability is well explained by the territory to which they belong. For this reason, they are assigned as substrata the area (LAU1-Local Administrative Units) of the municipality to which they belong. Consequently, by doing so, in addition to 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 within their strata by applying cluster analysis techniques. In this case, as they are larger municipalities and therefore have more or less practically guaranteed the sample representation of the area (LAU-1) which they belong to, the priority has been to use the auxiliary information available to form homogeneous groups of sections and, thereby, improve the accuracy of the estimates.

The auxiliary information used for the analysis in this second group comes from Census 2011 and the State Tax Administration Agency (AEAT). The characteristics selected are the most correlated to the variables under study in the Labour Force 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 2011 census, that is, illiterate, uneducated or education first grade

Percentage of persons with level of studies 4, 5, 6 or 7, ie, ESO, EGB, Bachillerato, FP

Percentage of persons with level of studies 8, 9, 10, 11 and 12 ie, bachelor's, master's or doctoral university

Finally, the following tax variables were used:

Total household income declared by its contributors.

Porcentage of capital income on total household income.

Porcentage of agricultural income on total household income.

(These latter variables are not available for the Basque Country)

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

The algorithm used to obtain clusters (substrata) was developed by Ward (1963). This is a multivariate algorithm for hierarchical cluster analysis based on minimization of distances between clusters. At each step, two cluster are grouped so that the sum of squares of distances between clusters is minimized over all possible partitions by grouping two clusters obtained from the previous step. Thus we move from a first stage with many conglomerates as sections in stratum until the last stage with all sections in a single cluster.

This method is available in the CLUSTER procedure of SAS / STAT SAS module.

Finally the TREE process was used, SAS also, that allows to see, in an easy way, a graphic with the process used forming clusters. This graphic, a tree graphic, facilitates the decision on the final number of clusters to be considered in each stratum.

The number of substrata within each stratum is assigned based on the internal variability of the clusters, and also considering the number of sections of each one, in order to avoid substrata too small and therefore with difficult sample representation.

4.3 SAMPLE SIZE

When the survey was implemented, the size was established by applying a method of minimum variance for a 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 minimize the variance of the estimates. This was performed using a linear cost function and the expression of the variation coefficient for the estimator of a proportion in sampling clusters with subsampling.

The following cost function was used:

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

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 needed for field work

All the variables were known except n and m.

The coefficient of variation for a proportion P is defined by the formula:

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

in which:

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

δ is the intra-cluster correlation coefficient, which was calculated for the active population and it is equals 0.05.

The minimum of the expression $CV(\hat{P})$ regarding 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

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

different values are obtained 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 in order to give compliance with the requirements of the European Union and improve the representation of the most disaggregated areas. Since the first quarter of 2005, a sample size of 3,588 sections and 18 dwellings per section was established, except in the provinces of Madrid, Barcelona, Sevilla, Valencia and Zaragoza, where 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

Since 2015, the sample size of primary units is maintained in 3822 sections, but in each primary unit 20 secondary units (main dwellings) are selected, except in the provinces of Barcelona, Madrid, Sevilla, Valencia, and Zaragoza, in which 25 dwellings are selected by section.

4.4 ALLOCATION

This section includes the criteria for the distribution of sample sections among the provinces, among strata within provinces and among substrata within strata.

The following aspects were considered when performing the provincial allocation:

a) National results should be as reliable as possible

b) In each province there should be a minimum sample size that will enable estimates of the same. (This fact, that allows quality provincial estimates, implies a loss of accuracy respect the national estimate obtained with the same final size, using proportional allocation).

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 compatible the three conditions mentioned above, a compromise allocation has been adopted between uniform and proportional, grouping provinces with a similar demographic importance and assign 3 to 12 interviewers, ie from 39 to 156 sample sections (with the exception of Ceuta and Melilla, which because of its small population size only have one interviewer and there are 13 sample sections in each).

Within each province, the sample allocation among strata is proportional to the population size of each one, nevertheless, the sample sizes of strata with the largest municipalities have been increased, 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, the dispersion should be larger and the cost per interview is lower.

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

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

Table 1

Sample distribution of census sections by provinces and strata

Provinces	Strata 1	2	3	4	5	6	7	8	9	Total
02 Albacete	18				8		4	5	4	39
03 Alicante/Alacant	17	10		16	21	6	5	3		78
04 Almería	16			9		7	3	4		39
01 Araba/Álava	30					3	6			39
33 Asturias	30	33		16	14	22	7		8	130
05 Ávila	15						6	5	13	39
06 Badajoz	20			6	10	6	12	14	10	78
07 Balears, Illes	42			9	28	12	9	4		104
08 Barcelona	50		33	24	22	12	9	3	3	156
48 Bizkaia	29	7		5	18	8	5	6		78
09 Burgos	20				7		3		9	39
10 Cáceres	19				7	8	9	10	25	78
11 Cádiz	12	13	6	26	9	6	6			78
39 Cantabria	35	9			11	9	11	9	7	91
12 Castellón/Castelló	26				29	5	7	4	7	78
13 Ciudad Real	13	9			14	16	11	7	8	78
14 Córdoba	34				17	7	10	10		78
15 Coruña, A	42	14	12		30	16	28	14		156
16 Cuenca	12						8	5	14	39
20 Gipuzkoa	26			6	12	18	10	6		78
17 Girona	15				24	13	10	8	8	78
18 Granada	26			5	10	14	9	8	6	78
19 Guadalajara	16				8			8	7	39
21 Huelva	14				8	8		9		39
22 Huesca	12					15			12	39
23 Jaén	17	8			15	12	10	16		78
24 León	24	10				10	15		19	78
25 Lleida	15					5	6	5	8	39
27 Lugo	26					14	14	24		78
28 Madrid	87		34	16	9	4	6			156
29 Málaga	33		6	19	10		5	5		78
30 Murcia	36	16		10	25	13	4			104
31 Navarra	34				8	12	11	14	12	91
32 Ourense	32					12	6	10	18	78
34 Palencia	20						10		9	39
35 Palmas, Las	44			24	22	9	5			104
36 Pontevedra	22	48			24	36	16	10		156
26 Rioja, La	33					11	7	6	8	65
37 Salamanca	20					4	5		10	39
38 Santa Cruz de Tenerife	24	13		10	22	9	7	6		91
40 Segovia	16						5	5	13	39
41 Sevilla	52			13	21	18	9	4		117
42 Soria	18						9		12	39
43 Tarragona	18	12			19	5	10	9	5	78
44 Teruel	12					6			21	39
45 Toledo	13	11			6	10	10	20	8	78
46 Valencia/València	45			15	29	10	7	7	4	117
47 Valladolid	36					4	8		4	52
49 Zamora	16					5			18	39
50 Zaragoza	58					5		8	7	78
51 Ceuta	13									13
52 Melilla	13									13

4.5 SAMPLE SELECTION

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, self-weighted samples are obtained within **each stratum**. This type of samples provides equal design weights to each sampled unit at stratum level. To do 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 Municipal Register. Within each section selected in the first stage, a fixed number of households is selected with the same probability by using a random start systematic sampling. As it was mentioned before (see section 4.3), in this survey 20 households per section are selected.

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:

$$P(V_{ijh}) = P(S_{jh}) \times P(V_{ijh} / S_{jh}) = K_h \times \frac{V_{jh}}{V_h} \times \frac{20}{V_{jh}} = K_h \times \frac{20}{V_h}$$

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 sample is self-weighted.

4.6 DISTRIBUTION OF THE SAMPLE IN TIME

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

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 scheme (see section 4.7) and week.

4.7 ROTATION SCHEME

As it was mentioned in the previous section, each period of the survey is a quarter, repeating it on.

The census sections remain in the sample indefinitely (subject to the exceptions listed in section 4.1), however family dwellings are partially renewed every quarter, in order to avoid fatigue of families. This renovation is performed in a sixth part of the sections.

For this purpose, the total sample of sections is divided into six subsamples, called *Rotations Groups*. Each section is identified by a five digit code. The last digit expresses the corresponding rotation group to which the sample section belongs, being numbered from 1 to 6.

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

Therefore, the dwellings in the sections of each rotation group collaborate the same number of quarters in the survey. This number of collaborations is associated to the rotation group and vary from one to a maximum of six times.

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

These renewed 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 group. In this way, it is possible to prevent bias in the estimates due to the different behaviour of the families depending on the time they are collaborating in the survey.

Each rotation group can be considered as a representative subsample. This facilitates obtaining estimates of structural variables by joining these subsamples.

4.8 ESTIMATORS

Until 2001, **ratio estimators** have been used, taking as auxiliary variables the figures of the resident population in family dwellings which are deduced from the Population Now Cast calculated by the INE.

The expression of this estimator, for the total of a specific characteristic Y, in a certain quarter of the survey is as follows:

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

the sum is extended to the strata of a province, an autonomous community or the national total, depending on the geographical level of the estimation.

In this formula:

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

p_h : is the number of resident persons 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.

From the first quarter of 2002, **Reweighting techniques** are applied to estimators in order to adjust the survey estimates to the information from external sources.

The reweighting technique involves the following:

Given a population $U = \{u_1 \ldots\ldots u_N\}$, from which it is selected the sample

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

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

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

where:

y_k : is the value of the characteristic researched in sample unit k.

d_k : Weight for unit k obtained using the expression $\frac{P_h}{p_h}$, h is the stratum to which the unit belongs.

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

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

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

The objective 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

- They should be close to initial weights d_k, and
- 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:

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

in which:

G = Function of distance.

X = Vector of dimension (J, 1) with the population totals of the auxiliary variables.

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

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

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

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

the problem is solved using Lagrange multipliers which allow to obtain a set of factors w_k that verify the balance conditions and provide the same estimates as the Generalized Regression Estimator.

In the particular case of the LFS, the linear distance function has been chosen in a truncated version (to avoid negative solutions of the system of equations), in order to maximise the properties of the regression estimator, with a small variance and minimum bias.

For the practical solution of this problem it has been used the software CALMAR (CALage sur Margès), programmed in SAS by the INSEE (Institut National de la Statistique et des Études Économiques) in France.

Since 2014, in the calibration process for LFS estimates, auxiliary variables are used in two levels of geographical breakdown:

Provincial

-Population aged 16 years and over by age groups and sex (16-29, 30-49, 50 and +)

Autonomous community

- Population aged 16 years and over by age groups and sex (22 groups).

- Population aged 16 years and over by nationality: Spanish or Foreign.

- Population under 16 years old by age groups and sex (6 groups).

- Household by size according to number of members (1, 2, 3, 4 and 5 or more).

In this way, the current estimates used in the LFS present a correct estimations of household totals by size, population totals by age and sex and the total of Spaniards and foreigners over 16 years by autonomous community as well as provincial totals by more aggregated age-sex groups.

In the autonomous cities of Ceuta and Melilla, given the small sample size, auxiliary variables are grouped to allow the calibration process.

5 Frame updates

Continuous population variations, either in their characteristics either in its spatial distribution, require updates in the sampling frame, that necessarily have repercussion in the sample structure.

The LFS considers four types of updates in the framework:

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

Updates in the dwelling frame, restricted and exclusive to sample sections. This update, as stated in section 4.7., aims to incorporate new main dwellings in the list of dwellings of the section.

Update the selection probabilities of the sections. This update aims to ensure that the sample of sections is equal to a sample selected in the year of the update, removing the minimum number of section from the sample. It is performed every three years.

General update on all sections and dwellings in which the definition of strata and substrata is revised and the probability of selection of the section is updated. It is done with the information from the Population Censuses (See section 5.2.2).

5.1 CHANGES IN SECTIONS OF THE SAMPLE

The following cases are considered:

5.1.1 Partition of sections

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

In this case, it is necessary to solve the problem of determining the selection probability for new sections in order to know which one will remain in the sample, as well as the number of dwellings to be interviewed to ensure the sample remain self-weighted.

Two cases are distinguished:

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

In this case the following steps are performed:

1) If V_S = Number of dwellings of section S according to the last census.

V'_S = Number of dwellings of section S after update.

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

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

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 interviewed is

$$m_j = m.\frac{V'_S}{V_S}$$

with m = 20 (or 25 in Barcelona, Madrid, Sevilla, Valencia and Zaragoza)

Thus the sample remains self-weighted.

B) Section S is fragmented 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 which has joined that part is automatically selected.

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

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

where

m = 20 (or 25 in Barcelona, Madrid, Sevilla, Valencia and Zaragoza)

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

$V_{S'j}$ = Number of dwellings within the boundaries of new section S'_j according to the last census.

5.1.2 Fusion of sections

Due to migratory and natural movements of the population, some sections become uninhabited. They are, therefore, joined with others, to ensure that if they are selected there will have units to investigate.

The fusion of sections is a particular case of the partition analysed in section 5.1.1.B.

Therefore, if a section S_j , that belongs to the sample, joins 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:

$$m_j = m.\frac{V'_{S'}}{V_{S'}}$$

where

m = 20 (or 25 in Barcelona, Madrid, Sevilla, Valencia and Zaragoza)

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

$V_{S'}$= Number of dwellings within the boundaries of new section S' according to the last census.

5.1.3 Variation of boundaries

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

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

In all the cases describe previously, the new sections are incorporated into the sample when according to the *Rotation scheme,* it corresponds to renew the families in the sections affected by such changes.

5.2 RENEWAL OF THE SAMPLE. UPDATE OF PROBABILITIES

When information is available from the electoral files, Population Censuses or the Municipal Register, it proceeds to update the probabilities of selection of the sections and to adjust to 20 (or 25 in the cases mentioned before) the number of interviews per section.

Changes in the sample of sections as a consequence of the update are included by rotation groups, that is to say, during a period of six quarters, as occurs in the case of the renovation of dwellings. In order to provide certain stability in the time series of data from the survey, the updates of the selection probabilities are performed every two or three years.

The most direct way to update the selection probabilities of sections is the selection of a new sample using the most updated sampling frame available. But such a radical change in a continuous survey as the LFS, generates three types of problems:

- Loss of essential information for selection and visit of the dwellings that are selected in the second stage. This information, which has to be compiled again, includes tangible aspects such as the directories of dwellings or the maps of the area, and other intangible but no less important, such as the fact that the interviewer is known by the population in the section and this makes easier the access to families and reduces the nonresponse.
- Loss of precision in estimates for interannual variations between quarterly indicators, since this reduces the common sample between both periods considerably.

- Possible presence of discontinuities in the time series of the survey data, due to the cause mentioned in the preceding paragraph

Therefore, it was decided to set up a procedure that, without distorting the selection probabilities that actually correspond to each section, maintains the sample of sections with minimal changes.

Two types of updates of the selection probabilities, based on the information available, are considered.

5.2.1. Updates performed using information from the Municipal Register.

In this case the definition of strata is not changed and remains the stratum already assigned to each municipality, although its population has changed and exceeded the limit of the lower or the upper stratum. The procedure used for updating is proposed by L. Kish and A. Scott (JASA 1971).

Let S to be a section belonging to the stratum h, whose probability of selection in the previous renovation (t-1) was given by:

$$P_S = \frac{V_S}{V_h} = \frac{Dwellings\ in\ section\ S\ in\ (t-1)}{Dwellings\ in\ stratum\ h\ in\ (t-1)}$$

and at the moment of the update (t), the corresponding selection probability is :

$$P_S' = \frac{V_S'}{V_h'} = \frac{Dwellings\ in\ section S\ in\ (t)}{Dwellings\ in\ stratum\ h\ in\ (t)}$$

P_S is compared with P'_S, and one of the following cases should be possible:

1) If $P_S > P'_S$ then 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 to have been selected in (t) with 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 some sections from the sample. These will be replaced by others sections from the same stratum, selected among **those that not belonging to the sample have increased its probability.**

This criterion maintains the scheme 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 from the Population Census.

As the information from the Population Census is more complete, definitions of strata and substrata are revised, and each municipality is assigned to the correspondent strata according to its new population figures.

Because of this many changes between strata can take place and the Kish-Scott procedure is too complex and does not guarantee to be optimal, in the sense that 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 (t-1) and in the new one (t), 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 section sample sizes at time 't-1' and 't', and in the strata h and h* respectively.

Then:

- If $\pi'_{h^*{}_s}$ is greater than π_{hs} and the section is in the sample, it remains in it.
- If $\pi'_{h^*{}_s}$ is greater than π_{hs} and the section is **not** in the sample, it will enter in the sample with probability:

$$\frac{(\pi'_{h^*{}_s} - \pi_{hs})}{1 - \pi_{hs}}$$

- If $\pi'_{h^*{}_s}$ is less than π_{hs} and the section is in the sample, it will remain in the sample with probability:

$$\frac{\pi'_{h^*{}_s}}{\pi_{hs}}$$

- If $\pi'_{h^*{}_s}$ is less than π_{hs} and the section was not in the sample, there is no possibility of entering the same.

Proceeding in this way, it can be proved that the probability of a section *S* to belong to the sample is $\pi'_{h^*_s}$, in other words, the probability updated in t, in the new stratum.

The main characteristic of this algorithm is the fact that it is quite simple to apply in complex situations. By contrast, it presents the inconvenience that it does not provide a sample with a fix size by stratum, which makes necessary a final adjustment removing sections with equal probability or adding sections with probability proportional to the size.

III. Evaluation of data quality

1 Introduction

Survey errors can be classified into two groups:

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

Non sampling errors, which are common to all statistical researches, whether the information is collected by sampling as Census is conducted. These errors appear in any stage of the statistical process:

- Before collecting the data: due to framework deficiencies or unsuitable definitions and questionnaires.

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

- After collecting the data: errors in editing, 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 method uses to obtain sampling errors is the Balanced **Half Sample Replications**.

This procedure consists in obtaining successive half-samples of the initial sample. Each half-sample is used to calculate an estimate of the characteristic for which the sampling error is going to be obtained. After calculating all estimates with each of the half-samples, and the estimate with the whole sample, the variance estimator 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 half-samples obtained, that is the number of replications

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

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

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

In the LFS the number of replications is 40. The following steps were performed to select them:

a) Within each stratum all sections were grouped in pairs, trying that both sections belong to the same *Rotation Group*.

b) Randomly, the first section of each pair was assigned to 20 replications and the other section assigned to the other 20.

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

The published tables include the *Relative Sampling Error* in percentage *(Coefficient of Variation*), calculated with 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 many difficulties due to the variety of reasons that can cause them, as well as the hypotheses, on which the theoretical models are based and generally are never fulfilled. This leads to obtaining approximate results.

In the LFS, the analysis of non sampling errors is based on the mathematical model created by the United States Census Office and due to Hansen, Hurwitz and Bershad which is based on repeating the survey interviews on a subsample of the dwellings originally selected. Later, data obtained on both occasions are compared in order to investigate the inconsistencies and quantifying errors by using different quality indices.

Besides the *repeated interview*, a specific study of those selected units that refused to provide information is performed.

For these units who refuse to collaborate in the survey, a *Questionnaire, of Refusals* is collected, in order to investigate some basic characteristics, like sex, age and relationship of the person who refuse the interview, as well as age, sex, nationality, finished studies, economic activity, branch of activity and occupation of the main person in the dwelling.

3.1 EVALUATION SURVEY

This survey has two objectives:

- Check the work of data collection in all the autonomous communities.
- Evaluate 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 investigate two types of non sampling errors:

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

b) Measurement errors, which affect the characteristics investigated in the persons of the sample.

The fieldwork is carried out by specialized agents, who perform the repeated interview at most three weeks after the original, with both interviews referring to the same period of time.

The fact that more than 70 percent of the refusals take place in the first interview, together with the existence of technical difficulties in conducting the evaluation survey (ER) with CATI, have determined that the ER only investigates **sections that are in first interview in the EO**. The collection method used in these sections, both EO and ER, is CAPI.

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

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

Each week, the sections in the first interview in one of the zones are investigated, with a random assignation of weeks to zones, so that each of zone 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 less questions.

With this procedure, a number of dwellings between 1.300 and 1.500, representing about 2 percent of the sample, is investigated.

In addition to the evaluation survey, and in order to detect errors in the process of updating the sections of the sample, each quarter a sample of fifty sections is

selected (one from each province, except Ceuta and Melilla) to assess the quality of the updates.

3.2 COVERAGE ERRORS

The comparison of the results obtained in both interviews provides indicators on the survey's coverage as well as indicators of measurement errors.

Coverage of dwellings: provides information about 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 erroneously included*, those who appear in the EO but not in the ER, since the interviewer performing the ER thought they were not surveyable.

3.3 MEASUREMENT ERRORS

Data on content errors are based on the information provided in the two 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 divided into the classes M_1, ..., M_k, the consistency table will be as follows:

E.R. \ E.O.	Number of persons	M_1	M_2	...	M_j	...	M_k
Number of 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 classified identically in both interviews.

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

E.R. \ E.O.	In category M_i	Not in Category M_i	Total
In category M_i	a	b	a + b
Not in category M_i	c	d	c + d
Total	a + c	b + d	n

The comparison between both tables provides the following equivalences:

$a = n_{ii}$ Number of persons classified in category M_i in both surveys.

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

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

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

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

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

a) Identically Classified Percentage

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

The value of PIC ranges is between zero and one hundred. This is an indicator of response stability. The optimal value (100) expresses that all persons who according to the ER belong to the class Mi obtained the same classification in the EO.

b) Index of Net Change

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

This element can be positive ($c > b$ o $n_{.i} > n_{i.}$) or negative ($b > c$ o $n_{i.} > n_{.i}$). It is an indicator of the response bias error, expressed as the percent difference between the original survey figure for class M_i and the reinterview figure.

c) Net Difference Rate

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

Similar to the previous option, but in this case the percentage refers to the number of persons classified in both interviews regarding to the reference characteristic.

c) Index of Gross Change

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

This element can be zero or positive. It is an indicator of the response variance.

e) Gross Difference Rate

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

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

To compare the overall quality of the different characteristics assessed, the **Index of Global Consistency** for a characteristic C is obtained from the table which includes all the categories of C.

It is defined as

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

An I.C.G. = 100 indicates no errors in the classification.