INSTITUTO NACIONAL DE ESTADISTICA

Labour Force Survey

Sample Design and Evaluation of Data Quality

Technical Report

February 2025

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 and to adapt to European regulations.

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.

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.

Likewise, Regulation (EU) 2019/1700 of the European Parliament and of the Council of October 10, 2019, which establishes a common European framework for surveys directed at people and households, determines the need to carry out a survey to research on the workforce.

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.

In order to comply with the requirements of the European Union that establish a common framework for the aforementioned household surveys, as well as Implementing Regulation (EU) 2019/2240 of December 16, 2019, which sets other characteristics of the European survey in the field of the labor force, it has been necessary to introduce a series of adaptations with effect from the first quarter of 2021.

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 year-round or most of the year l 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 Statis-tics) 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 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.

The framework used for the selection of the sample of both first and second stage units is the Georeferenced Addresses Framework (GRF), which is a structured and hierarchical information system made up of all the entities that are part of the territorial model: autonomous community, province, municipality, district, section, street and other elements that make up the postal address. In it, all addresses are structured, have unique identifiers and have a high percentage of georeferencing. In addition, these addresses are associated with the registered inhabitants and their territorial entities are adjusted to the Electoral Census street map.

Likewise, the GRF is synchronized with the Spatial Data Infrastructure of the INE (SDIINE), which provides it with a visualization layer in said environment.

It is updated annually, based on the information on registrations, cancellations and modifications of both the population and the territory, which the municipalities transmit monthly to update the Register. And the maintenance of the limits of the census sections, consisting of the partition, disappearance or creation of new sections, is done with the information reported by the Provincial Delegations of the INE.

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, throughout the national territory with the exception of the autonomous cities of Ceuta and Melilla, where sampling is in one stage.

The Primary Sampling Units (PSUs) are **census sections**. The PSUs sample remains fixed in time, except when they have been exhausted, due to having visited all surveyable dwellings, or when in the process of updating the sectioning some sections correspond to leave the sample due to the readjustment of the selection probabilities associated with population variations (see section 5).

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.

In the case of Ceuta and Melilla, a simple random sampling is carried out in each one, where the selected units are the main family dwellings, collecting information from all the people who have their habitual residence in them.

4.2 STRATIFICATION OF SAMPLING UNITS

In each autonomous community, the first-stage units are stratified following a geographical criterion, which assigns the stratum according to the size, measured in terms of population, of the municipality to which the section belongs.

Primary sampling units are stratified following a double criteria:

1. **Geographical criterion (stratification)**

   Sections are grouped by strata within each province, according to the size of the municipality which they belong to, measured in terms of population.

2. **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 theoretical strata considered respond to the following sizes:

- Stratum 0: Municipalities with 500.000 inhabitants or more.
- Stratum 1: Province capital municipalities under 500.000 inhabitants.
- Stratum 2: Municipalities with more than 100,000 inhabitants, except the above.
- Stratum 3: Municipalities between 50,000 and 100,000 inhabitants, except the above.
- Stratum 4: Municipalities between 20,000 and 50,000 inhabitants, except the above.
- Stratum 5: Municipalities between 10,000 and 20,000 inhabitants, except the above.
- Stratum 6: Municipalities under 10.000 inhabitants.

In some provinces it has been necessary to unite contiguous strata, either because there are no municipalities in any of them, or because the population is too small and therefore it would not correspond to a sample in its proportional distribution between strata.

4.2.2 Substrata

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

1. **Sections from strata 6**. 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 using estimation techniques in small areas.
2. **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 the data added at this territorial level from the 2018 Precensal File 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:

– Percentage 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 variable was used:

– Total household income declared by its contributors.

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 tax variable 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

In order to comply with the precision criteria imposed by the new European Regulation, the sample sizes have been calculated through a procedure to minimize the cost of the survey, subject to precision restrictions in terms of dimensioning of the variance of the estimator. To define the optimization problem we need the decision variables, the constraints and its objective function.

The optimization problem that has been raised has as decision variables:

– n_d^* with $d = 1,\ldots, 17$ is the number of census sections in the sample of the Autonomous Community d , excluding Ceuta and Melilla .

– m^* is the number of dwellings selected from each census section, which is set as an independent constant for the Autonomous Community.

The decision to select the same number of dwellings within each census section has been taken to facilitate the organization and management of the collection tasks. As in Ceuta and Melilla, the sampling is simple random, it is not necessary to consider the sample sizes associated with the first stage, but rather the number of dwellings collected in each autonomous city.

The precision requirements imposed by European regulations give rise to the following restrictions:

– Criterion 1: Accuracy requirement for national unemployment:

$$\sqrt{\hat{V}(\hat{p}_u)} \leq L_u$$

Where $\hat{V}(\hat{p}_u)$ is the estimate of the variance for the estimated proportion of unemployment in the population aged 16-74 at the national level andL_u It is 75% of the upper limit imposed by regulation.

– Criterion 2: Accuracy requirement for employment nationwide:

$$\sqrt{\hat{V}(\hat{p}_e)} \leq L_e$$

Where $\hat{V}(\hat{p}_e)$ is the estimate of the variance for the estimated proportion of employment in the population aged 16-74 at the national level andL_e It is 75% of the upper limit imposed by regulation.

– Criterion 3: Precision requirement for unemployment in each Autonomous Community:

$$\sqrt{\hat{V}(\hat{p}_{u,d})} \leq L_{u,d} \quad d = 1, \dots, 19$$

Where$\hat{V}(\hat{p}_{u,d})$ is the estimate of the variance of the estimated proportion of unemployment in the population aged 16-74 years in the autonomous community *d* and$L_{u,d}$ is 75% of the upper limit in the community$d$ imposed by regulation

The objective function to optimize is the cost function of the survey. We have started from a cost $Q_d(n_d^*, m^*)$ for the autonomous community$d$ which collects the costs derived from conducting the EPAS interviews through the two collection methods that are mainly used: CAPI (computer-assisted personal interviewing) and CATI (computer-assisted telephone interviewing). So the expression for the cost function would be:

$$Q_d(n_d^*, m^*) = C_1 n_d^*(\alpha + \beta m^*) + C_2 m^* n_d^* f_2$$

Where:

– C_1 = Average daily cost of researching a primary unit (section) per CAPI

– The straight line $\alpha + \beta m^*$ = number of days needed to investigatem^* dwellings in one section. Parametersα Yβ are defined based on the experience acquired in the collection in previous periods of the survey.

– C_2 = Average cost of investigating a home per CATI interview.

– f_2 = Proportion of dwellings that are investigated by CATI.

Adding in all the communities we obtain the final expression of the objective function:

$$\sum_{d=1}^{19} Q_d(n_d^*, m^*)$$

The restrictions are given by the precision limits established above.

This problem has been solved using the SAS PROC OPTMODEL procedure for nonlinear programming.

The sizes resulting from the resolution of the optimization problem have been revised to guarantee minimum sample sizes by provinces and to limit the changes with respect to the old sample sizes.

The number of **sections in the sample is 5,298**, investigating **14 dwellings per section** . Of these, 288 correspond to the sample increase collected by the Galician Institute of Statistics under the agreement signed with the INE. In the case of the autonomous cities of Ceuta and Melilla, the sampling will be in a single stage and 286 main dwellings are selected in both cases. The following table shows the distribution of the number of sections between the different Autonomous Communities.

Autonomous Community	Tracts
01 Andalucía	751
02 Aragón	230
03 Asturias, Principado de	188
04 Balears, Illes	191
05 Canarias	239
06 Cantabria	159
07 Castilla y León	531
08 Castilla-La Mancha	336
09 Cataluña	490
10 Comunitat Valenciana	366
11 Extremadura	210
12 Galicia	576
13 Madrid, Comunidad de	325
14 Murcia, Región de	168
15 Navarra, Com.Foral de	174
16 País Vasco	239
17 Rioja, La	125
18 Ceuta (*)	286
19 Melilla (*)	286

(*) Number of dwellings investigated

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. A compromise allocation between uniform and proportional has been adopted for the distribution of the sample of the Autonomous Community among the provinces that comprise it.

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

The table below shows the distribution of the sample sections by provinces and strata.

	ESTRATO21							
	0	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
01 Araba/Álava		37				4	8	49
02 Albacete		27			13		24	64
03 Alicante/Alacant		24	15	29	37	11	18	134
04 Almería		20		18		17	18	73
05 Ávila		16					29	45
06 Badajoz		27		11	17	10	59	124
07 Balears, Illes		66		17	53	29	26	191
08 Barcelona	89		59	43	43	25	31	290
09 Burgos		33			13		22	68
10 Cáceres		19			15		52	86
11 Cádiz		10	28	33	17	8	9	105
12 Castellón/Castelló		20		5	19	5	17	66
13 Ciudad Real		12			22	17	26	77
14 Córdoba		31			16	9	21	77
15 Coruña, A		48	18	14	44	26	60	210
16 Cuenca		12					33	45
17 Girona		9			23	13	26	71
18 Granada		23			17	18	28	86
19 Guadalajara		16				12	22	50
20 Gipuzkoa		20		7	15	21	16	79
21 Huelva		17			17	9	18	61
22 Huesca		12				13	23	48
23 Jaén		12		5	13	13	24	67
24 León		23			23		37	83
25 Lleida		17				8	31	56
26 Rioja, La		59			9	14	43	125
27 Lugo		28	0	0	0	16	48	92
28 Madrid	170		71	41	15	10	18	325
29 Málaga	44		11	38	19		21	133
30 Murcia		51	24	18	47	28		168
31 Navarra		55			20	26	73	174
32 Orense		30	0	0	0	14	46	90
33 Asturias		40	50	23	19	28	28	188
34 Palencia		22					23	45
35 Palmas, Las		42	11	20	32	10	7	122
36 Pontevedra		16	58	0	40	40	30	184
37 Salamanca		29				7	28	64
38 Santa Cruz de Tenerife		24	17	22	30	8	16	117
39 Cantabria		48		14	23	17	57	159
40 Segovia		15					30	45
41 Sevilla	55		10	9	29	23	23	149
42 Soria		19					26	45
43 Tarragona		12	9		21	6	25	73
44 Teruel		12					33	45
45 Toledo		12		19		15	54	100
46 Valencia/València	52			19	46	19	30	166
47 Valladolid		54			10		27	91
48 Bizkaia		35	10	7	26	13	20	111
49 Zamora		16				4	25	45
50 Zaragoza	97					11	29	137
51 Ceuta (*)		(286)						26
52 Melilla (*)		(286)						26
Total (**)	**507**	**1.170**	**391**	**412**	**803**	**577**	**1.438**	**5.298**

(*) Número de Viviendas investigadas

(**) Sin contabilizar Ceuta y Melilla

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 the population data. 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 14 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:

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

in which:

m = 14

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 of the stratum h.

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.

In Ceuta and Melilla, 286 dwellings are selected in each autonomous city using simple random sampling. Therefore, each houseV_i has the same probability of being selected:

$$P(\mathrm{V_i}) = \frac{286}{V}$$

Where V is the total number of dwellings in Ceuta or Melilla, as the case may be.

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 (except as discussed 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.

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

The distribution of the number of sections per stratum and week is similar in each rotation group.

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 $U = \{u_1 .....u_N\}$, from which it is selected the 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 : 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 2021, in the EPA calibration process for population and housing estimates, auxiliary variables are used in two levels of geographical breakdown:

– Provincial
  - Population by age groups (0-14, 15-29, 30-49, 50 and +), and sex.
– Autonomous community
  - Population by five-year age and sex groups.
  - Population aged 15 years and over by nationality: Spanish or Foreign.
  - Dwelling by size according to number of members (1, 2, 3 and 4 or more)[1]

[1] Note: Until 2020, the term household is equivalent to dwelling, as all residents in the dwelling are considered members of the household. From 2021, with the entry into force of the new European regulations associated with the Framework Regulation on Social Statistics 2019/1700, in order to be part of the household, a common budget must be shared and it is possible that more than one household can be found in the same dwelling.

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 15 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 sample size, auxiliary variables are grouped to allow the calibration process:

Population by age group(0-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65 and over) and sex considering Ceuta and Melilla jointly.

Dwelling by size considering Ceuta and Melilla jointly: (1, 2, 3 and 4 or more).

Population aged 15 and over in Ceuta and Melilla, for each of them.

Population aged 15 and over, by Spanish or foreign nationality considering Ceuta and Melilla jointly.

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,** updated Georeferenced Addresses Framework is available on an annual basis, which aims to incorporate all the changes produced in the dwellings of the national territory (changes in the postal address, registrations).
- **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:

1. **Section S is broken down to form two or more complete sections.**

   In this case the following steps are performed:

   a) 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.

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

   c) The number of dwellings that must be interviewed is

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

      with m = 14

      Thus the sample remains self-weighted**.**

2. **Section S is fragmented to be annexed to one or more existing sections.**

   In this case:

   a) 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.

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

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

      where

      m = 14

      $V'_{S'j}$ = 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 = 14

$V'_{S'}$= 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

An update of the section selection probabilities is carried out every three years based on the most current population data available.

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 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 effect of the interview number on the information and a variable workload on the interviewer.

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 every three years.

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 every ten years

Every ten years, 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 simplicity of its application. 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

When, from the data of a sample survey, it is intended to estimate a population parameter, under the hypothesis that an appropriate estimator is being used, an estimate of it will be of high quality if the data on which it is based are of high quality. Conversely, if the survey data is of low quality, the estimates will also be of low quality.

Furthermore, the sample size on which the estimates are based is also an important determinant of quality. Even with high quality data, an estimate based on a very small number of observations will be unreliable. Therefore, the quality of an estimator of a population parameter is a function of the total survey error , which includes, on the one hand, an error that derives only from the fact of selecting a sample instead of carrying out a complete census, called sampling error , and on the other, an error related to data collection and processing procedures, known as non-sampling error .

Optimizing the design of a survey involves striking a balance between sampling and non-sampling errors.

2 Sampling errors

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

The Jackknife method is used to obtain the sampling errors.

This procedure is based on the creation of sub-samples, where each of these is obtained by eliminating a primary unit from the total sample. In the autonomous cities of Ceuta and Melilla, given that the sampling used is not two-stage, in each stratum fictitious primary units are formed by dividing the sample of dwellings into random groups.

From each subsample, or jackknife sample, the quarterly estimate of the characteristic whose sampling error we want to obtain is obtained. This estimate is calculated in the same way as the quarterly estimate on the full sample, that is, including the non-response and calibration adjustments.

Once all the estimates have been calculated with each of the jackknife samples, as well as the estimate with the complete sample, the estimator of the variance is given by the expression:

$$\hat{V}(\hat{Y}) = \sum_{h} \frac{n_h - 1}{n_h} \sum_{j \in h} \left(\hat{Y}_{(hj)} - \hat{Y}\right)^2$$

where:

$\hat{Y}_{(hj)}$ is the estimate based on the jackknife sample obtained by removing primary unit j from stratum h from the complete sample.

$\hat{Y}$ is the estimate based on the full sample.

n_h is the number of primary units in stratum h.

The tables publish the relative sampling error in percentage (coefficient of variation), which is given by the following expression:

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

3 Non sampling errors

Non-sampling errors are errors that occur at any stage of the survey. They come from, among others, the following sources:

1. **Specification and measurement** errors: These errors occur when what is intended to be measured or found by the survey and what is actually obtained in the interview process do not match. They can have multiple causes: concepts or definitions that are not well specified or are confusing for the informants, questions that are not written correctly, inadequate formulation by the interviewer, inadequate response by the interviewee...
2. **Framing errors**: occur when there are elements of the population that are omitted or duplicated in the sampling frame or when there are elements included in the same that should not be (elements mistakenly included).
3. **Errors due to lack of response** : three types can be distinguished here:
   a) **Lack of response from the unit**: occurs when an element of the sample does not collaborate in the survey for different reasons (refusal to collaborate, absence, inability to answer, etc.).
   b) **Lack of response to one or more questions**: occurs when the questionnaire has only been partially completed, due to the fact that there have been questions that have remained unanswered.
   c) **Incomplete answer**: occurs when in open questions the informant provides some information, but the answer is too short to allow an adequate coding.

The methods for evaluating these errors are generally expensive and difficult to implement. At the EPA, currently, the study of non-sampling errors is focused on the analysis of errors due to frame defects and the lack of response from the reporting units.

Frame errors and those due to lack of response originate situations, called incidents, which cause some units not to collaborate in the survey.

On the one hand, a quantification of the incidents is carried out according to various variables, such as the data collection method used, the number of interviews (first or subsequent), the stratum to which the unit belongs (provincial capitals and the rest municipalities). Likewise, the distribution of incidents by autonomous communities is obtained.

On the other hand, a specific study is carried out of those selected units that are surveyable but that refused to provide the requested data.

For these units that refuse to collaborate in the survey, a refusal questionnaire is completed, in which an attempt is made to collect a series of basic minimum characteristics.