BANCO DE PORTUGAL
EUROSYSTEM

Methodological note

Portuguese Household Finance and Consumption Survey

The Portuguese Household Finance and Consumption Survey (ISFF, the Portuguese acronym for *Inquérito à Situação Financeira das Famílias*) is conducted by Banco de Portugal and Statistics Portugal and is part of the Household Finance and Consumption Survey (HFCS) project. The purpose of this project, coordinated by the European Central Bank (ECB), is to obtain harmonised microeconomic data on the financial situation of households residing in the euro area. The HFCS working group (Household Finance and Consumption Network - HFCN) has established a set of variables to be collected in all countries and agreed on the methodological principles underlying the implementation of the HFCS. These principles have been adopted in the ISFF and adapted to the Portuguese reality.

This note presents the main methodological aspects of the ISFF. Section 1 describes the sample selection method, section 2 presents data collection, while the remaining sections describe the post-collection data treatment targeted at attaining the ISFF microdatabase. Section 3 describes data editing, section 4 the imputation of missing values, section 5 the harmonisation of variables with the HFCS, section 6 the anonymisation procedures, while sections 7 and 8 describe the estimation of survey weights and replicate weights respectively.

Further details on the ISFF methodology can be found in the methodological documents published by Statistics Portugal. In addition, the comparison of the methodological aspects of the ISFF with those of other surveys integrating the HFCS can be found in the methodological documents published by the ECB.

1. Sample

The ISFF sample design is set out with the purpose of obtaining data that are representative of households residing in Portugal[1], particularly their wealth and assets. To characterise the wealth of a typical household, it would be sufficient to select a representative sample of the population based on geographical criteria, as is usually the case of household surveys conducted by Statistics Portugal. However, given that the distribution of wealth is quite skewed (i.e., most is held by a very small share of the population), a sample selected only with geographical criteria would have to be very large and thus entail high operational costs to allow for a correct characterisation of total household wealth and its distribution. To minimise this problem, in the ISFF sample wealthiest households are overrepresented. This higher share of wealthiest households in the ISFF sample, in relation to the population, is adjusted when calculating the survey results with the weights of each household, which ensure that the data are representative of households residing in Portugal.

The sample selected for the ISFF comprises private household dwellings of main residence (8,000 dwellings in the first three waves and 14,814 dwellings in the fourth wave, held under pandemic conditions). In all waves, the selected sample is made up of two subsamples: one subsample is selected with the aim of being representative of the households residing in Portugal based on geographic criteria and the other subsample is selected with the aim of overrepresenting the wealthiest households.

Similarly to other household surveys conducted by Statistics Portugal at that time, the ISFF sampling frame in the 2010 wave was the Master Sample, i.e., a sample of dwellings extracted from the 2001 Census. In the following ISFF waves, the sample frame became the National Dwellings Register, which has the benefit of covering all private dwellings in the Portuguese territory. This file is built from the Census and is regularly updated with administrative data on real estate property and data from other surveys.

[1] The target reference population of the survey are private households; it therefore excludes people living in collective households and in institutions, such as retirement homes or prisons, which account for less than 1 per cent of the total population.

In the 2010 wave, the strategy to oversample the wealthy households was to increase the sample in the metropolitan areas of Lisbon and Porto. This strategy was based on the available evidence, e.g., from the 2005-2006 Household Wealth and Indebtedness Survey (IPEF), which pointed to a higher probability of finding wealthier households in these two regions. In the following ISFF waves, the overrepresentation sample was made up of dwellings larger than certain limits in square meters, defined by region, based on data from the 2010 ISFF[2]. This change was implemented because the new sample frame covered information on property size, which - as shown by the analysis of the ISFF 2010 - is more correlated to household wealth than geographical location. In the 2020 wave, the selection of the overrepresentation sample considered not only the useful area classes of the dwellings, but also a sample increase in the geographical areas with higher incomes. The highest income geographic areas were obtained based on the total income of the tax aggregates of the Personal Income Tax Liquidation Statements.

The ISFF sample is a probabilistic sample, meaning that all households residing in Portugal have a non-zero probability of being selected. The sample is stratified by region (NUTS 2 and sub-divisions) and selected in two stages. At the first stage, areas are selected (1 sq. km GRID INSPIRE cells), with a probability proportional to the number of household main residences. At the second stage, dwellings are selected systematically within each area. This methodology is applied to the two sub-samples of ISFF dwellings.

2. Data collection

In surveys with questions focusing on complex issues related to private matters, such as income and wealth, it is particularly difficult to encourage households to participate and

[2] The geographical areas are the following: Norte region, excluding the municipality of Porto; the municipality of Porto; Centro; the 'Greater Lisbon' area (municipalities of Cascais, Lisbon, Loures, Mafra, Oeiras, Sintra, Vila Franca de Xira, Amadora, and Odivelas); Setúbal Peninsula (municipalities of Alcochete, Almada, Barreiro, Moita, Montijo, Palmela, Seixal, Sesimbra, and Setúbal); Alentejo; Algarve; Azores; and Madeira. The larger dwellings are the ones with useful areas equal to or greater than 100 m2 in the "Greater Lisbon" area; 120 m2 in the Porto, "Península de Setúbal", Algarve and Madeira areas; and 150 m2 in the North region except Porto, Centro, Alentejo and Azores.

ensure that responses are accurate. The methodological options adopted when collecting data for the ISFF seek to minimise this type of problems.

Interviews are based on a data collection software specifically designed for the ISFF. This software makes it possible, inter alia, to include coherence and plausibility checks that help detect and correct possible response errors during the interview.

In the first three waves of the ISFF, interviews were carried out in person. In the 2020 wave, during the pandemic, there were two alternative collection methods: by telephone and via the web. Before beginning the fieldwork, interviewers attend information sessions on the ISFF conducted by Banco de Portugal and Statistics Portugal staff members. The questionnaire is presented in full at these sessions, jointly with the underlying concepts, and interview simulation exercises are conducted using the collection software.

Also, to minimise the non-response rate, before the first contact of the interviewer, households that are part of the ISFF sample receive a letter from Banco de Portugal and Statistics Portugal seeking to raise their awareness to the importance of participating in the survey. This letter is sent along with a leaflet presenting the ISFF, including examples of the usefulness of the information collected in this type of survey.

Most ISFF questions, namely those regarding assets, liabilities and consumption, refer to the household as a whole. The most financially knowledgeable person is selected to answer these questions. Information on employment status, pension rights and specific types of income is collected for individual household members aged 16 and over. In these cases, questions should be answered by them where possible.

Many ISFF questions refer to money amounts, which are particularly difficult to collect given that respondents either do not recall exact figures or refuse to report them. In these questions, to minimise non-response, answers may be given from a range of amounts defined by the respondent or selected from a pre-defined table provided in the data collection software. Responses collected in ranges are subsequently imputed, inside the indicated range limits.

ISFF response rates are quite high in comparison with those obtained in most of the other countries that are part of the HFCS project. The number of households with completed interviews included in the final database amounted to 4,004, 6,207, 5,924 and 6,107on the first, second, third and fourth waves, respectively.

The ISFF collection stage has a duration of about four months and preferably takes place over the course of the second and/or third quarters of the year. The reference period for most variables is the time of the interview. For income variables, the values collected refer to the previous year, given that it was considered easier for households to answer in relation to amounts corresponding to a full calendar year.

3. Editing

After the fieldwork conclusion data are analysed in detail, so as to correct errors and, where possible, complete the data with auxiliary information collected during the interview. At this editing stage all the interviewers' written comments on each answer are read. In addition, data for each household are checked for plausibility and coherence, using additional checks other than those implemented in the collection software. This interview-by-interview analysis is also supplemented by an analysis of the outlier values, focusing mainly on monetary variables.

For some specific variables, such as income and interest rates on loans, auxiliary variables are collected during the interview that in some cases make it possible to complete the variables in the event of non-response. The ISFF variables on income refer to gross values. However, when households can only answer in net values, these are collected and used to estimate gross values using a micro-simulation model. Where possible, missing values in interest rates on loans are completed based on data on the reference rate and the spread, collected during the interview.

Data changes during the editing stage, either filling missing answers, correcting or deleting answers considered to be implausible, are recorded in the database under shadow variables (flags). The value of the flags indicates the origin of the content of the

variable to which is it associated (e.g., 1050 for deleted answers, 3050 for changed answers and 5050 for estimated answers).

4. Imputation

During the editing it is only possible to fill out a very limited number of missing answers. Missing data are mostly due to 'Don't know' or 'No answer' responses but may also result from answers that were deleted or changed during the editing stage. Data analysis based only on households with no missing answers may bias the output substantially. In fact, non-response to certain survey questions can be related to the characteristics of the relevant households. Furthermore, the existence of few observations in some questions may Additionally, the existence of few observations in some questions may hinder some studies.

Given the problems stemming from non-response, the HFCN decided that missing data for the main variables should be imputed. A stochastic multiple imputation method was agreed on as the imputation methodology, in order to take into account the uncertainty associated with the process. This methodology preserves the joint data distribution, since the imputed values are obtained from adding a random value (based on the data conditional distribution) to the value predicted by an imputation model. A non-stochastic imputation method, based for example on the replacement of missing values with the mean of the collected answers, makes the distribution of variables concentrated around this mean, underestimating variables variance. Choosing a multiple imputation method, i.e., a method that gives several values for variables with missing answers, aims at taking into account the uncertainty surrounding the imputation process. In the HFCN context it was agreed that the database should include five implicates, i.e., five versions of the variables which differ in terms of the values assigned to the imputed answers. The higher the number of implicates the more accurate the estimates obtained but the more computationally demanding becomes the use of the final database.

The imputation methodology used in the ISFF is detailed in Martins (2020). The imputation routines developed for the ISFF are largely based on a SAS software written by the ECB for the HFCS's multiple imputation (€MIR European Multiple Imputation Routines). The main part of the €MIR code consists in the FRITZ (Federal Reserve Imputation Technique Zeta) routine, written by Arthur B. Kennickell for the imputation of the Survey of Consumer Finances (SCF).

The imputation methodology used in the ISFF is based on the assumption that data are Missing at Random (MAR). This means that, conditional on other variables, the fact that one variable has missing values does not depend on what its value would be. When answers are MAR, the mechanism of missing values can be ignored if the parameters that determine missing data are independent from the imputation model parameters. To meet these assumptions, the imputation models need to include non-response related variables in the set of covariables, i.e., of explanatory variables. Covariables should also include variables that are good predictors of the set of variables to be imputed, as well as variables that - in accordance with the different economic theories - are related to the variable to be imputed. The inclusion of this latter type of covariable is important to prevent data from being biased in favour of a given economic model. In general, it is advisable to include a high number of covariables in imputation models.

The functional form of imputation models varies according to the type of variable to be imputed. Linear regression models are used for continuous variables, linear probability models for binary variables and hotdeck procedures[3] for categorical variables.

Imputation is carried out through an iterative sequential process. As such, it is important to define the order in which variables will be imputed. First, include those with a low number of missing answers that are good predictors of the other variables to be imputed. Each iteration involves two stages: (1) imputation based on the parameters estimated in the previous iteration and on observed and imputed data in that iteration; (2) at the end of the iteration, estimation of the imputation model parameters based on

[3] In these procedures the missing value is replaced by the value reported by households with similar covariates values.

observed and imputed data. These stages are repeated on every iteration until process convergence is achieved. The first iteration is slightly different from the remaining ones, since, while it lasts, the values imputed for a variable are used jointly with observed data to estimate the parameters to be used in the imputation of the following variables.

In the ISFF all variables related with assets, liabilities, income and consumption are imputed, along with the variables on which the former ones depend and those to which they are strongly related[4]. The imputation process takes into account the minimum and maximum values applicable to the variables, as well as some restrictions stemming from the relationship between variables.

In the ISFF database, the flags enable the identification of all imputed observations, as well as the reasons for missing data prior to imputation (e.g., code 4050 identifies the case of an imputed observation because the respondent answered "Don't know", and code 4053 is assigned for answers in euro collected within a range and subsequently imputed).

5. Harmonisation of variables

The ISFF variables can be divided into three groups, as regards the comparability with other countries' variables that are part of the HFCN: (1) core variables – have a harmonized definition with the HFCS variables and also exist for the remaining countries participating in this project; (2) non-core variables - also have a harmonized definition with the HFCS variables but are optional, so they exist only for countries that opted for their inclusion; (3) national variables – are only part of the ISFF database. Most of the ISFF variables, most notably those that comprise household assets, liabilities, income and consumption, are core variables. Non-core and national variables are easily

[4] For example, when there is a missing answer on the value of the main residence because the household did not answer the question on the main residence ownership, the answer on the ownership will also be imputed. Additionally, given the strong relation between the current value of the main residence and its value on the acquisition date, variables on the acquisition date and value at the time of acquisition will also be imputed.

identifiable in the database and correspond to those whose second name letter is N and O, respectively.

In some cases, national specificities justify that information on variables with a harmonized definition be collected with a greater degree of detail than would be necessary for its completion. Under these circumstances, harmonised (core and non-core) variables are built ex post, and this transformation is recorded under code 13 in the flag variable. In these cases, the more detailed information collected is also recorded under national variables. Consequently, in some cases, the ISFF database contains simultaneously harmonised HFCS variables and national variables on the same topic.

6. Anonymisation

ISFF data are anonymised to prevent households and any individual respondents from being identified based on the answers given. In fact, although households are only identified in the ISFF database as a randomly defined number, it is also necessary to ensure that households with uncommon characteristics cannot be identified by crossing their answers to the various survey questions.

The anonymisation rules applied follow the principles set out within the HFCN, in order to ensure the comparability of results across countries. In this process, discrete variables which present outliers are trimmed (top or bottom coded). In some continuous variables, which can be matched with external data sources, values are randomly rounded. For categorical variables, with infrequent categories, a greater aggregation of categories is performed. Finally, there are some variables, notably on the geographical location or on the sample selection, which although belong to the list of ISFF variables are empty in the database for anonymisation purposes. The flags for these latter variables have the code 2050. The other anonymised variables appear in the ISFF database with a name that ends in '_COD', '_R' and '_B', depending on the values having been trimmed, rounded or aggregated, respectively. All these cases can be found in the ISFF variables codebooks available on the ISFF database section of the survey page on Banco de Portugal's website.

7. Weighting

The ISFF sample is not a simple random sample, this means the probability of selection differs across population elements. This requires weights for each household to be used in the calculation of extrapolated statistics for households residing in Portugal. The weight for each household in the sample corresponds to the number of households residing in Portugal with similar characteristics.

In a household sample survey, the initial design weight for each household corresponds to the inverse of this household's probability of being selected for the sample. However, these initial weights must be adjusted as the final sample differs from the selected sample, namely due to non-response. Weights adjusted for non-response correspond to the inverse household's probability of being part of the initial sample times a non-response correction factor in the household main residence's region. The probability of response may be estimated through a more or less complex model and depends in particular on the information that can be obtained on the non-response generating process. In the ISFF only geographical information is used. For each NUTS 2 region an adjustment factor is applied corresponding to the ratio of the total number of units in the population and the number of units who have participated in the survey.

Finally, weights adjusted for non-response are corrected to take into account information on the distribution of specific population characteristics that may have some influence over the main variables of interest. This procedure is named calibration and aims at aligning some variables' distributions in the sample with their distributions in the population. In the ISFF there is a simultaneous calibration of households and individuals to ensure consistent estimates. At the household level, the auxiliary variables (margins) considered are the number of households by NUTS 2 and the number of households according to their size (1, 2, 3 and 4 or more persons), as well as the outstanding amount of loans for house purchase by NUTS 2. At the individual level, margins are population estimates by gender and five age groups (except for the first and last groups, that include all persons aged 16 and less and those aged 75 and over, respectively). In the 2020 wave, additional margins were also considered: the number

of households by occupation regime of the main residence (owners, tenants or free use), the number of households by typology of urban areas (predominantly urban areas, moderately urban areas and predominantly rural areas) and the number of individuals aged 16 or over by level of education (basic or lower, secondary and higher).

8. Replicate weights

The use of final weights (calibrated and adjusted for non-response) makes it possible to obtain, based on the final sample point estimates for specific population statistics. These estimates depend on the selected sample. In fact, another sample obtained from the same sample frame, under the same conditions and with the same size, would lead to different estimates. It is therefore important to have an accuracy measure of the estimates, i.e., to calculate their variance.

There are basically two types of method for estimating variance: analytical methods, which can be particularly difficult to implement in the case of complex sample designs, and replication methods, obtained by building replicate weights. The latter procedure was adopted for the ISFF, as agreed on within the HFCN. This choice was largely encouraged by arguments on the use of the database by researchers. In fact, estimating variance based on analytical models implies the release of information on the sample design, which would not be possible to include in the database to be made available to researchers for confidentiality purposes. By contrast, replicates contain all the necessary information to calculate the variance of estimators without the need to release any information on the sample selection process that might lead to a household being possibly identified.

The HFCN established that there should be 1,000 replicates, with sub-samples obtained through the bootstrap method and subsequently calibrated with the same margins used in final weights. Hence, the ISFF database includes 1,000 replicates, corresponding to the weights associated with the different sub-samples selected from the total sample. These replicates should be used to calculate the variances of the point estimates obtained from ISFF data, as explained here.