EU-SILC CONFERENCE LISBON 2014

Workshop, Lisbon, 15 October 2014
*Session 4*

# STATISTICAL MATCHING OF IT-SILC AND HBS: SOME CRITICAL ISSUES

Gabriella Donatiello, Doriana Frattarola, Antony Rizzi, Mattia Spaziani

(Italian National Institute of Statistics - ISTAT)

Abstract[1]

In recent years, there has been increasing interest in statistical matching techniques applied to consumption expenditure and income in order to provide more appropriate measures of standard of living. The purpose of this paper is to evaluate the possibility of using two different data sources to create an integrated database with detailed information on households consumption and income, using the sample of IT-SILC 2012 (income reference year 2011) and the HBS (Household Budget Survey) 2011 sample. In this paper different approaches were used and improvements from using auxiliary information in order to relax the CIA (Conditional Independence Assumption) are also highlighted. From this point of view, the aim is to discuss the advantages in having better harmonization of common variables of SILC and HBS (e.g. housing costs) or a more reliable monthly household income in HBS. Additionally new shared questions on consumption in SILC questionnaire and /or SILC module (e.g. variables on food expenditures) would improve the quality of the matching process. Evidence from these issues are finally presented.

[1] The views expressed in this paper are solely those of the authors and do not involve the responsibility of ISTAT.

## 1. Introduction

The growing demand to provide data for measuring households economic well-being at the micro level has encouraged the production of integrated statistics on household income, consumption and wealth. The need of new indicators that cover cross-cutting information on social and economic aspects is among the current priorities of the National Statistical Institutes (NSIs) as well as a major goal at European level. The redesign of the social statistics framework towards a better integrated system of European social surveys also looks at the integration techniques as a good opportunity to enhance the potential information of the existing data sources. In particular the statistical data matching is recognized as a complementary tool, among other integration techniques, for producing statistics on variables not jointly collected in a single survey, with reduction of survey costs and response burden. However, there are several methodological issues involved in the statistical matching that is a complex process and this aspect needs to be taken into account in particular for assessing the quality of the final estimates.

This paper will focus on the statistical matching as an additional tool to enhance the social and economic data on household surveys currently available. The aim of our work is to evaluate the possibility of integrating two different data sources in order to provide joint information on household income and consumption expenditures in Italy at the micro level. For this goal we used IT-SILC 2012, with income reference year 2011, and the HBS (Household Budget Survey) 2011. In addition this paper will shed light on the data requirements and those pre-conditions necessary for an effective use of the micro integration techniques.

It is well known that an *ex-post* integration of existing micro data sets has to face several challenges that could be mainly resolved at an earlier phase of data collection. In order to fully utilize the matching techniques the advantages in having a more efficient *ex-ante* data collection system as well as a better harmonization of common variables of SILC and HBS and other important social surveys are widely discussed.

Based on our exercises in matching consumption expenditures from HBS into IT-SILC, the paper presents the most significant features related to the harmonization and reconciliation issues and the role of the auxiliary information in improving the matching estimates. In effect it is not always possible to perform statistical matching under the *conditional independence assumption* (CIA), i.e. independence between income and consumption given some common information in both the data sources. The only way to bypass the CIA is to introduce some auxiliary information in the matching procedures and the advantages in using a reconstructed HBS household income is definitely underlined. Moreover an assessment of the predictive power of the housing costs components of HBS and IT-SILC is also presented. In order to facilitate the integration techniques and improve the quality of the matching estimates, the introduction in the SILC module of a small number of questions on food consumption and transport which are able to act as new shared variables is finally debated.

## 2. Some critical factors in matching income and consumption data

### 2.1 A brief introduction to statistical matching

Statistical matching (SM) procedures usually refer to a broad range of model-based techniques that generally aim to achieve a micro data file from different sources that have a set of variables in common but do not contain the same units or the same identifier. The primary object of SM is to provide joint information on variables not collected through a single survey.

In general terms the statistical matching can be considered as an inferential problem with incomplete information and it can be treated as an imputation from a donor survey to a recipient survey. Many SM

techniques are in effect based on methods developed for the imputation of missing values such as parametric (e.g. regression imputation), nonparametric (hot deck imputation) or mixed methods (e.g. methods based on predictive mean matching).

In a standard SM framework, the surveys to integrate, indicated as *A* and *B*, present a set of common variables *X*, while the variable *Y* is observed only in *A*, and the variable *Z* is observed in *B*. In order to generate a fused data set that contains all the information on X,Y,Z, it is possible to impute the missing variables (*Z* in this case) in the recipient survey (*A*). The synthetic data set can be otherwise produced by concatenating the two data sources and then filling in the missing variables.

It is worth noting that data integration at the micro level it is not always necessary if the final objective is the estimation of one or more parameters (correlation coefficient between *Y* and *Z*; regression coefficients, contingency table $Y \times Z$). On the contrary data integration is necessary when the final goal is a *fused or synthetic* data set which contains all the variables of interest (X,Y,Z) (Donatiello *et al.* 2014).

The statistical matching procedure is rather a complex process that raises important methodological concerns regarding especially the validity of results, since it relays on underlying assumptions not always verifiable. In effect most of the matching techniques assume (i) the conditional independence (CIA) of the target variables given the common variables (i.e. the measures of association between Y and Z conditional on X cannot be estimated and they are usually assumed to be 0, in other words Y and Z are independent ) and, (ii) the observations in the samples are independent and identically distributed (i.i.d.) (i.e. the sample is a simple random sample).

The conditional independence is particularly important for assessing the quality of matching estimates, however it rarely holds in practice. When this condition holds, the matching estimates reflect the true joint distribution of variables collected in different sources and give the same results as a linkage procedure. In the case where the conditional independence does not hold, this assumption can be relaxed if some auxiliary information on the relationship between *Y* and *Z* is available (e.g. estimates of a correlation coefficient, third data source observing jointly the target variables such as a small sub-set of units with complete information on the joint distributions). In the event that auxiliary information is not available, the model will have identification problems and the fused datasets may lead to incorrect inferences. When the joint distributions of target variables is not available, some proxy variables with very high predictive power can be used. These variables are able to mediate the relationship between Y and Z and make reasonable the conditional independence assumption.

In addition the conditional independence can be overcome by approaching SM in terms of analysis of uncertainty that assess the sensitivity of estimated results to different assumptions through the estimation of specific contingency tables (D'Orazio *et al.* 2006).

Moreover in case of data matching of complex sample surveys involving two or more stages of selection of the sample units, as usually applied in social surveys, the i.i.d. assumption is also difficult to be maintained. The statistical matching procedure turns out to be more challenging as it has to include the treatment and harmonization of survey weights. In this case, SM methods that take into account the sampling design and the unit weighs such as (i) Renssen's approach based on calibrations of the weights (Renssen 1998), and, (ii) Rubin's *file concatenation* (Rubin 1986) would be applied (D'Orazio *et al.* 2010 and 2012). When the variables of interest (*X*, *Y* and *Z*) are categorical, as often in households surveys, the Renssen's approach seems more appropriate and promising.

The main objective in our exercises in matching HBS with IT-SILC is to enhance IT-SILC data on income and social exclusion with consumption data derived from HBS survey. We used HBS as a donor data set and we imputed consumption expenditures classes in SILC in order to obtain a synthetic micro data set. More specifically both a non-parametric method micro (random hot deck) and the exploration of SM uncertainty have been performed. Also an exercise based on SM method for complex sample surveys (Renssen's approach) has been applied, but further examinations are needed in order to obtain valid results.

## 2.2 Statistical matching: some pre-conditions

Statistical matching can be used as additional tool in order to cover crossing needs that are particularly difficult to collect such as the joint distribution of income, consumption and wealth. It should be noted that income and consumption are very complex concepts that generally need exhaustive list of questions to be collected in household surveys. Two individual surveys are frequently used with different mode of data collection as the information on household consumption are mostly based on dairies. As a consequence statistics on the joint distribution of income and consumption are very challenging to obtain with a single survey. The integration methodologies could represent a good chance for exploiting the available surveys if the data requirements that are able to facilitate the integration process are really met.
It is known that the statistical matching procedures strongly depend on the quality and coherence of data sources and of the common variables (Eurostat 2013). It is worth noting that a greater and effective use of matching techniques is actually limited by the current extend of harmonization of EU-SILC, HBS and other important social surveys. However a new approach to statistical matching based on the ex-ante identification and incorporation, at the design stage, of some pre-conditions of micro integration that need to be fulfilled is rapidly spreading.
It should be noted that the inconsistencies in data sources that need harmonization and reconciliation can only partly be dealt with an ex–post integration technique. The incoherencies between surveys can arise at different levels of the statistical process and basically depend by differences in data collection (e.g. dissimilar definitions, different variables measuring comparable concepts, etc) and in survey methods (e.g. sampling design, weighting, calibration, treatment of missing values). The current process of modernization of social surveys at European level is going towards a better integration and coordination of surveys also in order to facilitate the matching process. In this contest an ex-ante harmonization of common variables, statistical units and concepts in SILC and HBS could effectively enhance the application of matching techniques and could simplify the estimation of parameters or indicators on the joint distribution of variables of interest.

### *2.2.1. The harmonization of IT-SILC and the Italian HBS*

An essential point in the success of ex-post matching procedures is the existence of a set of common variables in different data sources that are homogeneous in their statistical content. It can be said that EU-SILC and HBS show a large number of common variables, mostly related to demographics, household composition, dwelling, labour, income, whose quality and coherence are in general quite good. In our matching exercises the step of selection and harmonizing of the common variables has nonetheless resulted in an intense phase of reconciliation of classifications and definition of units, with a re-coding of several variables in order to have the same degree of detail. We believe that this time-consuming step can be more easily performed with the new editions of the Italian HBS. It is worth noting that in recent years ISTAT has undertaken a deep process of harmonization of national social surveys starting from the changing of the mode of data collection from Papi to Capi in 2011. In particular HBS, after a long testing phase, will switch to a new consumption expenditures survey in 2014, with a first data release in 2015. Alongside with some important methodological improvements, aiming at fostering data comparability at European level, the new Italian HBS has been designed to harmonize as much as possible the main common variables with IT-SILC. These reconciliation and harmonization efforts clearly go beyond the core social variables (Eurostat 2011) and affect all the variables measuring comparable concepts. Particularly attention has been paid to demographic variables, household composition, family relationship with the reference person, level of education, ILO labour status. Furthermore the information on dwelling facilities have been extended in order to get closer to those provided by IT-SILC, so as to allow also the estimation of the imputed rent by a regression method as applied in IT-SILC. In effect the Italian HBS does not estimate the imputed rent

but provides a measure of the subjective rent. It is likely that the inclusion of imputed rent in HBS housing costs will make more comparable the corresponding IT-SILC variable. We are confident that all these adjustments will make easier the integration procedures. In effect these changes will be definitely in the direction of greater coherence and harmonization of social surveys by the fulfilment of those pre-conditions essential for data matching and micro integration.
It is well known that the matching exercises that rely on a restricted number of common variables do not usually provide good quality estimations of the target joint distributions mainly due to the underlining assumptions in SM. As pointed out before, in order to improve matching results the use of auxiliary information and/or proxy variables for one of the two target concepts are essential.
As regard consumption and income, HBS could be an appropriate source of information but the quality and coherence of income data (few questions on net monthly income at household level) are not comparable with EU-SILC, where income is more extensively and better collected. Nonetheless the existence of income information in both surveys has been essential in our matching exercises in order to overcome the CIA and reduce the uncertainty associated with our target joint distributions. As a matter of fact, we used HBS income information in the estimation process after a reconciliation of income statistics that has implied a reconstruction of HBS income variable in a new variable. This issue will be treated more extensively in a following paragraph but it is worth noting that we used the HBS available information on income and savings. The presence of few valuable questions about the use of the household income (e.g. consumption and savings) has allowed us to reconstruct HBS income classes and compare them with those of IT-SILC. From this point of view, the inclusion of one or two questions on savings in HBS can be useful for data integration purposes, as well as for improving the quality of information on household monthly income.
At present, few consumption variables mainly related to housing costs (utilities, rents, mortgage interests, regular maintenance and repairs) are collected in SILC questionnaire. The housing costs may actually represent shared variables with high predictive power for matching purposes, even though they have not been selected as a matching variables in our current exercises. In a following paragraph a deep analysis of the current housing costs components of HBS and IT-SILC is also presented. We believe that the information on housing costs have a great potential and explanatory power so as to be exploited more in our next exercises.

## 3. Statistical matching of IT-SILC and HBS

It is well known that an *ex-post* integration procedure of two different sources mainly consists of several steps which can be summarized as follows: (i) preliminary analysis of the data sets; (ii) reconciliation of the data sets through the harmonization of definition and classification; (iii) selection of the matching variables; (iv) selection of the matching methods more suitable with the final objective; (v) quality assessment of the results. In the next paragraphs some important aspects of our exercises in matching consumption classes from HBS (donor) into IT-SILC (recipient) are presented. The main objective is to highlight significant features related to the harmonization step and the role of the auxiliary information in improving the matching estimates.

### *3.1Preliminary analysis of the data sets*

In Italy, both the HBS and IT-SILC cover the same population of private households and are equally based on a two-stage simple random sampling design. The primary sampling units (PSU) are the municipalities and the second stage units (SSU) are the households. Inside each administrative region, the PSU are stratified according to their demographic size and, in order to guarantee self-weighting design in each region, the total of residents in each stratum is approximately constant. The evaluation of frequency distributions (weighted and non-weighted) of the variables in both datasets proved that

keeping the respective weights of the two surveys is rather suitable (Table 3.1), although additional analyses for dealing with the treatment and harmonization of survey weights are also considered. It should be noted that the discrepancies shown in the table are mainly due to the differences in reference period for population size (end of year for IT-SILC and quarterly population for HBS).

**Table 3.1 Comparison between HBS and IT-SILC by sample size and population size**

| | Household level | | Individual level | |
|---|---|---|---|---|
| | **HBS** | **IT-SILC** | **HBS** | **IT-SILC** |
| **Sample size** | 23,158 | 19,578 | 57,613 | 47,365 |
| **Population size** | 25,165,002 | 25,429,176 | 60,286,784 | 60,797,109 |

*3.2 Harmonization of the datasets*

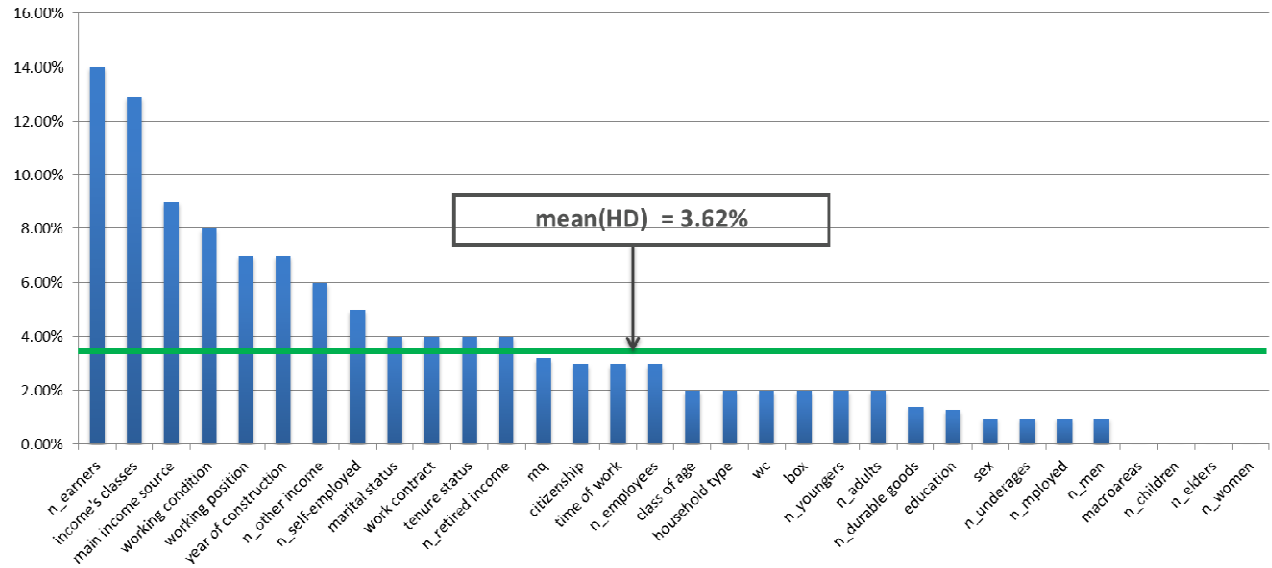
In order to apply a statistical matching procedure it is necessary to choose a set of common variables that have to be comparable. At the outset the common variables need to be harmonized across the two datasets by comparing the definitions in the two surveys and afterwards by harmonizing the definitions and classifications in such a way as to make them homogeneous (D'Orazio et al 2006). The table in Annex 1 contains the final codification of these derived variables. As we have imputed the HBS household consumption classes, the analyses are done at household level and in some cases the variables are aggregated from the individual level. Each individual variable refers to the reference person and in both survey it is the holder of the registry form. It should be noted that the selected matching variables must satisfy at least two criteria. First of all, there must be homogeneity in the distribution across the two surveys. Second, the variable must be good predictors of both income and consumption. The selected common variables are shown in table 3.2. The analysis is clearly different when dealing with categorical and continuous variables. Only the categorical common variables are used in our current matching exercises, while the housing costs are explored as potential auxiliary information with high predictive power to be used in the next exercises.
With respect to categorical variables, as a measure of coherence, the weighted frequencies and the Hellinger Distance (HD) have been used for analyzing the similarity/dissimilarity of the variables distributions across the two data sets. Annex 2 include all the relative frequencies and Figure 3.1 shows the HD of the common variables. The monthly household income and number of earners finally present the highest values of HD. The large discrepancy of income variable and number of earners is quite foreseeable since in HBS the latter variables have not the same quality and level of detail as in SILC. Marginal distributions which have HD distance below 5% (the chosen arbitrary threshold) are considered coherent. Values of HD greater than 5% and lower than 10% refer to the following variables: main income source, main activity, professional status of reference person and year of dwelling construction.

## Table 3.2 Selected common variables HBS -IT-SILC

| Categorical common variables | |
|---|---|
| **Household reference person** | Sex, Marital status, Age, Educational level attained, Citizenship, Main activity, Professional status, Type of contract, Classification of economic activities (NACE), Number of hours usually worked per week in main job, Main income source |
| **Household structure** | Number of children (0-8), Underage people (9-17), Younger people (18-39), Adults (40-64), Elderly people (65- ), Number of women and men in the household |
| **Income** | Number of employed people, Individuals with employee income, Individuals with self-employed income, Individuals with retired income, Number of income earners, Monthly household income (in classes) |
| **Housing condition** | Type of housing, Year of construction, Macroareas, Square meters, Tenure status, Imputed rent |
| **Presence/absence of housing amenities** | Kitchen, Bathroom, Hot water supply, Garage |
| **Number of durable goods** | Refrigerator, Dishwasher, Washing Machine, Car, Phone, Tv, Vcr, Personal computer |
| **Household type** | Single person households, Households with or without dependent children |

| Continuous common variables | |
|---|---|
| **Housing-related expenses** | Water, Electricity, Modified HH070, Mortgage repayment, Rent, Subjective Rent, Total Housing Expenses |

## Figure 3.1 - Hellinger distance of the common variables




### *3.2.1 An assessment of housing costs*

As regards the analysis of continuous variables, a comparison between measures of location and dispersion is carried out as well as parametric and non-parametric test. It's worth mentioning that a substantial amount of consumption expenditures related to the target variable HH070 (Total Housing Costs) is collected in SILC. The costs of utilities (water, electricity, gas and heating) and in general all kind of expenses connected with the household right to live in the accommodation are also included.

For owners and tenants this variable include mortgage interest payments and rent payments, respectively. HBS collects most of the components included in HH070 except the expenses for municipal solid waste, the sewer services and the mortgage interest payments[2]. In order to compare the housing costs, a modified variable of HH070 is calculated in SILC excluding the costs not covered in HBS. After a comprehensive analysis of each component collected in the two surveys, a new variable is created in each survey, by adding to the modified variable HH070 the rent payments for tenants and the subjective rent for non-tenants as usually covered in HBS housing costs. The table 3.3 shows descriptive statistics for the common continuous variables. Looking at means it would appear that the only mortgage repayment and the modified HH070 have large discrepancy between surveys. It is worth noting that standard deviation presents considerable difference notably for water expenses, modified HH070 and mortgage repayment. Further investigations with specific tests are needed in order to evaluate the real similarity of the distributions.

Both parametric independent sample t-test and non-parametric Kolmogorov-Smirnov test are performed by comparing if the means between two groups, in this case the two surveys, are the same. The assumption underlying a t-test is that each of the two populations being compared should follow a normal distribution. Another attention that should be addressed before using the t-test is whether the population variance can be considered to be equal. This assumption is necessary in order to use pooled variance in the calculation of the t statistics. The Levene's F Test is the most commonly used statistic to test the assumption of homogeneity of variance. In table 3.4 we reject the null hypothesis (no difference) for the assumption of homogeneity of variance and we conclude that there is a significant difference between the two group's variances. Looking at the corresponding t-test we accept the null hypothesis for the rent payment and total housing expenses and conclude that there are not significant differences between HBS and SILC.

**Table 3.3 Descriptive statistics for the common variables on housing costs**

| | Survey | Mean | Std. deviation | Std. error mean |
|---|---|---|---|---|
| **Water** | HBS | 273.5 | 236.5 | 1.90 |
| | SILC | 241.4 | 165.7 | 1.38 |
| **Electricity** | HBS | 528.1 | 382.3 | 2.51 |
| | SILC | 551.7 | 345.6 | 2.49 |
| **Mortgage repayment** | HBS | 497.3 | 253.1 | 4.93 |
| | SILC | 636.9 | 356.3 | 6.81 |
| **Modified HH070** | HBS | 221.7 | 187.8 | 1.23 |
| | SILC | 199.4 | 109.2 | 0.78 |
| **Subjective rent** | HBS | 565.1 | 278.0 | 2.01 |
| | SILC | 582.6 | 302.1 | 2.36 |
| **Rent** | HBS | 357.9 | 194.2 | 3.06 |
| | SILC | 393.2 | 222.0 | 3.94 |
| **Total Housing Expenses** | HBS | 750.8 | 371.9 | 2.44 |
| | SILC | 751.3 | 359.4 | 2.57 |

[2] HBS collect variables on mortgage repayment.

**Table 3.4 Parametric and non parametric tests for housing costs**

| | | Levene' Test for Equality of Variances | | T-test for Equality of Means | | | Kolmogorov-Smirnov Test | |
|---|---|---|---|---|---|---|---|---|
| | | F | Sig. | T | df | Sig. (2-tailed) | KSA | D |
| **Water** | Equal variances assumed | 404.991 | .000 | 2.306 | 42734 | .021 | 8.718 | 0.085 |
| | **Equal variances not assumed** | | | **2.358** | **42321.370** | **.018** | | |
| **Electricity** | Equal variances assumed | 22.095 | .000 | -4.023 | 42734 | .000 | 8.418 | 0.082 |
| | **Equal variances not assumed** | | | **-4.053** | **42464.147** | **.000** | | |
| **Mortgage repayment** | Equal variances assumed | 53.961 | .000 | -16.502 | 5374 | .000 | 4.951 | 0.135 |
| | **Equal variances not assumed** | | | **-16.607** | **4950.272** | **.000** | | |
| **Modified HH070** | Equal variances assumed | 836.881 | .000 | 14.602 | 42734 | .000 | 6.8444 | 0.06645 |
| | **Equal variances not assumed** | | | **15.220** | **38165.592** | **.000** | | |
| **Subjective rent** | Equal variances assumed | 33.557 | .000 | -6.476 | 42734 | .000 | 6.35198 | 0.062 |
| | **Equal variances not assumed** | | | **-6.446** | **40709.349** | **.000** | | |
| **Rent** | Equal variances assumed | 11.043 | .001 | -1.041 | 42734 | .298 | 1.172 | 0.011 |
| | **Equal variances not assumed** | | | **-1.034** | **40348.577** | **.301** | | |
| **Total Housing Expenses** | Equal variances assumed | 83.214 | .000 | -.228 | 42734 | .819 | 4.47542 | 0.043 |
| | **Equal variances not assumed** | | | **-.229** | **41986.678** | **.819** | | |

Kolmogorov-Smirnov test is the non-parametric test equivalent to the independent sample t-test with unequal variances and quantifies a distance between the empirical distribution function of two samples. The null distribution of this statistic is calculated under the null hypothesis namely that the samples are drawn from the same distribution. In this case, H0 is rejected for almost all the variables, except rent payments and total housing expenses. As a conclusion we can say that there are significant differences between the ways each survey collects individual housing's cost, but the analysis on the reconstructed variable that comprehends all the comparable costs confirms a good degree of comparability among HBS and SILC.

*3.3 The matching procedure*

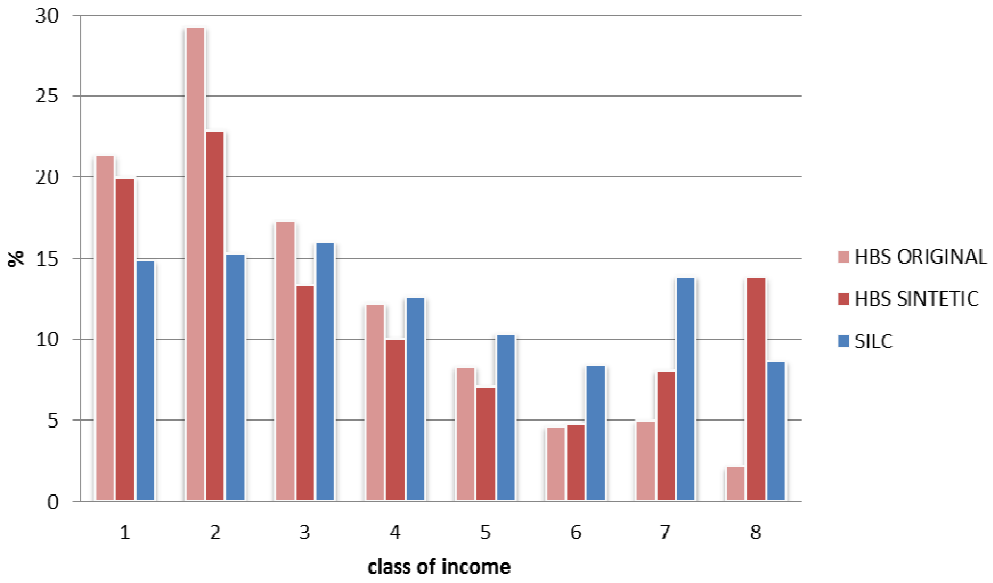
A first step in our matching procedure consisted in the application of random hot deck under CIA, using the R package StatMatch (D'Orazio, 2013). Then the exploration of SM uncertainty was also applied (Donatiello *et al.* 2014). The random hot deck is performed by specifying the donation classes and an actual observed value for classes of consumption is imputed in to IT-SILC.
It is known that the CIA cannot be verified from the matched datasets and it is clearly an unsatisfactory model for expenditures and income whatever the conditioning variables are. This conclusion is confirmed by the uncertainty analysis carried out by calculating the Fréchet bounds for the contingency table between the variables of interest given the two common variables being considered. The only way to bypass the CIA is to introduce some auxiliary information in the matching step.

*3.3.1 The role of the auxiliary information in matching*

Some previous works on SM techniques applied to social surveys (Coli et al 2005) have highlighted the importance of using the household income variable as auxiliary information in order to overcome the

CIA and improve the final estimation. Nonetheless the income variable is clearly collected in different ways between the two surveys (Donatiello *et al.* 2014). In HBS income is observed at household level from a multi-response variable that is included in an ad hoc section about income and savings. After a question about the average household income (in classes), an additional information about the use of income is collected. In particular the household can choose between two options and indicate if the whole income is spent in household consumption or, instead, if there is a saving. In the latter case, the household has to declare the amount of the saving. This additional information from the income section has been used to estimate a new income variable in order to reduce the large income discrepancy in the two surveys, as shown in Figure 3.2.
In actual fact we have considered a variable, denoted as *diff*, constructed as the difference between the class of declared income and the class of consumption. If the household has indicated to save a part of income and *diff*<0 the new income variable is constructed as the sum of consumption and saving declared. The reason of this condition comes from the greater reliability of consumption information collected in HBS with respect to the income information. Finally the new income variable has decreased the household income's underestimation in HBS and the Hellinger distance also decreases from 17,2% to 12,9%.

**Figure 3.2 Comparison of HBS and IT-SILC income classes**




Moreover the reconstructed HBS income variable has been used among the selected matching variables in order to perform the SM procedures. In this case the auxiliary information can be represented by the approximation of the actual income/expenditure relationship. In particular, the auxiliary information concerning the reconstructed income was used in the matching procedure in further restricting the subset of potential donors. In this work we use the synthetic class of income as covariate of consumption.
As a result it is possible to consider the difference between the imputed consumption classes in SILC using the original HBS income or using the new HBS income variable as auxiliary information. The Figure 3.3 shows a valuable improvement of the estimates in consumption highest classes, with a decrease from 5.1% to 1.5% of the Hellinger distance.
Furthermore the unlikely assignments between classes of consumption and income are rather limited. In other words there is a significant decrease of those frequencies corresponding to classes of consumption that differ more than three from the respective class of income. It is worth nothing that

looking more closely at the impact of the imputations on other variables not selected as matching variables, it should be noted that the household typology presents a similar distribution to the original in HBS (Table 3.5). We believe that this is a very promising starting point for the distributional analysis on the propensity to consume by main socio-demographic variables such as household type.

**Figure 3.3 Comparison of class of consumption imputed with synthetic variable and original variable**

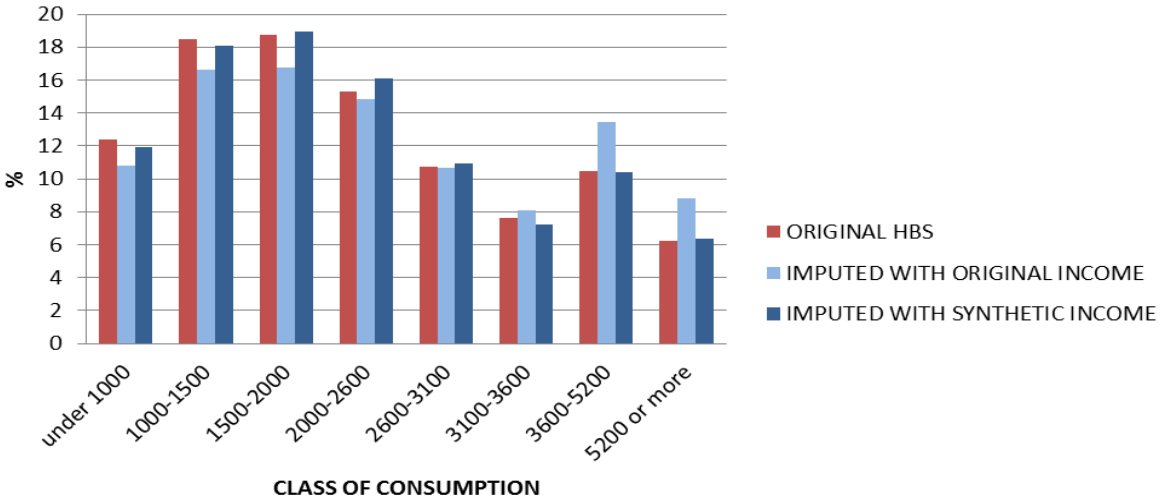



**Table 3.5 Comparison of HBS and Imputed consumption classes by household typology**

| | | Consumption | under 1000 | 1000-1500 | 1500-2000 | 2000-2600 | 2600-3100 | 3100-3600 | 3600-5200 | 5200 or | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Household typology | Single member under 35 | Hbs | 5.4 | 5 | 4.7 | 2.7 | 2.2 | 2.2 | 1.8 | 1.7 | **3.5** |
| | | Imputed | 6.3 | 7.1 | 5.3 | 5.1 | 3.3 | 3.8 | 3 | 2.7 | **5.1** |
| | Single member 35-64 | Hbs | 18.3 | 18.7 | 16.6 | 14.9 | 11.3 | 8.9 | 8 | 9 | **14.3** |
| | | Imputed | 17.5 | 15.1 | 13.4 | 10.3 | 8.9 | 7 | 9.8 | 7.2 | **12.2** |
| | Single member 65 and over | Hbs | 45.3 | 25.8 | 15 | 8.8 | 5.4 | 6.7 | 4.1 | 3.8 | **16** |
| | | Imputed | 40.6 | 22.7 | 12.6 | 7.9 | 6.4 | 4.6 | 3.8 | 2.8 | **15.2** |
| | Couple with r.p. (a) under 35 | Hbs | 0.7 | 1.6 | 1.9 | 1.9 | 1.2 | 2.3 | 2.2 | 0.9 | **1.6** |
| | | Imputed | 0.4 | 1.4 | 2.1 | 2.8 | 1.9 | 3.4 | 3.2 | 1.3 | **2** |
| | Couple with r.p. 35-64 | Hbs | 2.3 | 4.9 | 7.4 | 8.5 | 9.4 | 9.7 | 7.6 | 5.6 | **6.8** |
| | | Imputed | 3.1 | 4.2 | 6.3 | 6.3 | 6.6 | 7.5 | 8.7 | 9.6 | **6** |
| | Couple with r.p. 65 and over | Hbs | 9.7 | 11.1 | 11.1 | 10.5 | 9.6 | 8.8 | 7 | 7.2 | **9.8** |
| | | Imputed | 8.2 | 10.5 | 9.7 | 7.4 | 7.8 | 6.7 | 6.5 | 4.6 | **8.3** |
| | Couple with 1 child | Hbs | 4.8 | 10.7 | 14.1 | 16.8 | 21.5 | 20 | 24.8 | 22.5 | **15.8** |
| | | Imputed | 9.2 | 11.7 | 16.7 | 22.5 | 23.1 | 21.1 | 22.1 | 29.4 | **17.8** |
| | Couple with 2 children | Hbs | 2 | 7.1 | 11.6 | 17.5 | 21.2 | 22.5 | 25.5 | 26.5 | **15** |
| | | Imputed | 5.3 | 9.6 | 15.5 | 18 | 22 | 24.5 | 24.5 | 22.7 | **15.8** |
| | Couple with 3 or more children | Hbs | 0.9 | 1.4 | 3 | 3.4 | 4.1 | 4.9 | 7 | 6.8 | **3.5** |
| | | Imputed | 1 | 3.3 | 4.3 | 3.8 | 2.7 | 5.9 | 4.1 | 4.7 | **3.5** |
| | Single parent | Hbs | 6.7 | 9 | 10 | 10.2 | 7.8 | 8.9 | 7.7 | 8.5 | **8.8** |
| | | Imputed | 5.1 | 9.5 | 8.1 | 8.6 | 9 | 9.6 | 6.1 | 5.6 | **7.9** |
| | Other typology | Hbs | 3.8 | 4.8 | 4.6 | 4.8 | 6.2 | 5.2 | 4.5 | 7.5 | **5** |
| | | Imputed | 3.1 | 4.9 | 6 | 7.2 | 8.3 | 5.8 | 8.4 | 9.4 | **6.2** |

## 4. An ex-ante approach to data matching

In order to facilitate the integration techniques and improve the quality of the matching estimates an *ex-ante* collection of information on wealth/consumption in SILC can be a great opportunity for having new shared variables with high predictive power. For instance the introduction of a small number of questions on food consumption and transport in SILC, together with the variables on housing costs, could add valuable information for estimating a total consumption variable usable as auxiliary information in the matching procedures.
This section focuses on the identification of those consumption components that are good predictors for total consumption in HBS. The aim is to improve the quality of the matching estimates similarly to the use of the income information in HBS.

### *4.1 The structure of total consumption*

The main goal is to analyse the structure of total consumption and compare the shares that different items have across the classes of income. Afterward the explanatory power of each amount is investigated through the use of a statistical model. In Table 4.1 the main consumption components at aggregated level are shown.

**Table 4.1 Food and Non-Food components**

| | **Consumption components** |
|---|---|
| **Non-Food** | Tobacco, Clothing and footwear, Personal care and health, Transport, Education, Recreational and cultural services, Furnishings and household equipment, Restaurants and hotels, Household maintenance and secondary residence, Total housing expenses, Miscellaneous goods and services |
| **Food** | Quantitative food consumption |

Figure 4.1 shows that three very large components (Total housing costs, food, transport) in effect represent 63% of total consumption. Looking at distribution for different volume of income in Figure 4.2, a well-known trend of the food costs can be noted. As expected, the share that people reserve to food consumption decreases as income increases: from 24% for the first class of income to the 14% for the last and richest class. A similar trend observable in the total housing expenses is largely due to decreasing amount of rent payment for higher classes of income.
The method for selecting explanatory variables of total consumption is stepwise regression, in which a sequential procedure evaluates candidate predictors for possible inclusion and tests variables, already in the model, for possible removal. For instance, if the significance of a given estimated coefficient is above some specified threshold, it is eliminated from the model. The dependent variable is the logarithmic transformation of monthly total consumption expenditures. Socio-demographic characteristics of household and of the reference person (including the synthetic class of income) next to all main consumption components are considered as covariates. As shown in table 4.2, total housing costs, food and transport are the most important predictors. If we include the class of income, the model reaches a very high R-square of 0.71. Two out of four most relevant variables are common to the two surveys. As already underlined in other studies, these results suggest that predicting total consumption from a limited set of variables could explain a large fraction of the total variability.

**Figure 4.1 Percentage of Food and Non-Food components on total consumption**

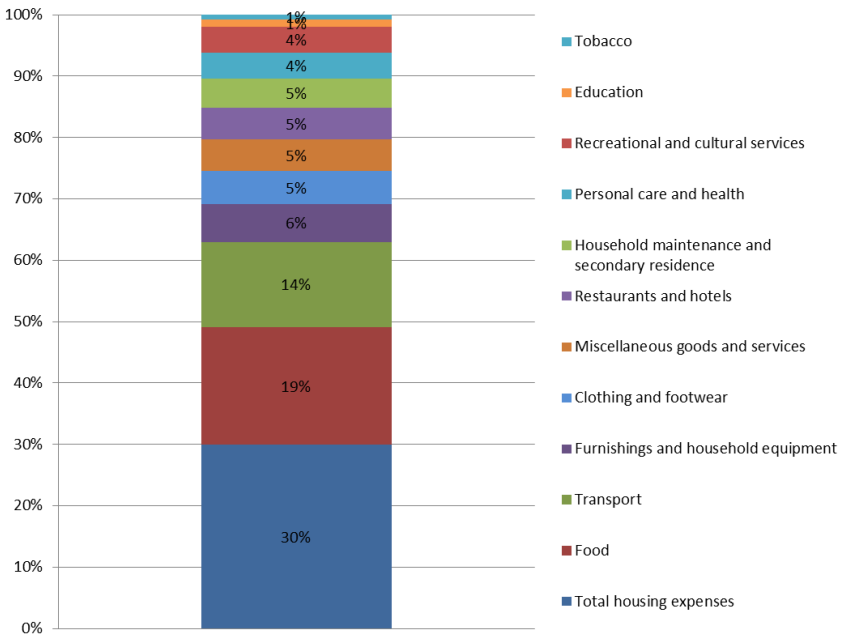



**Figure 4.2 Percentage of Food and Non-Food components on total consumption by income classes**

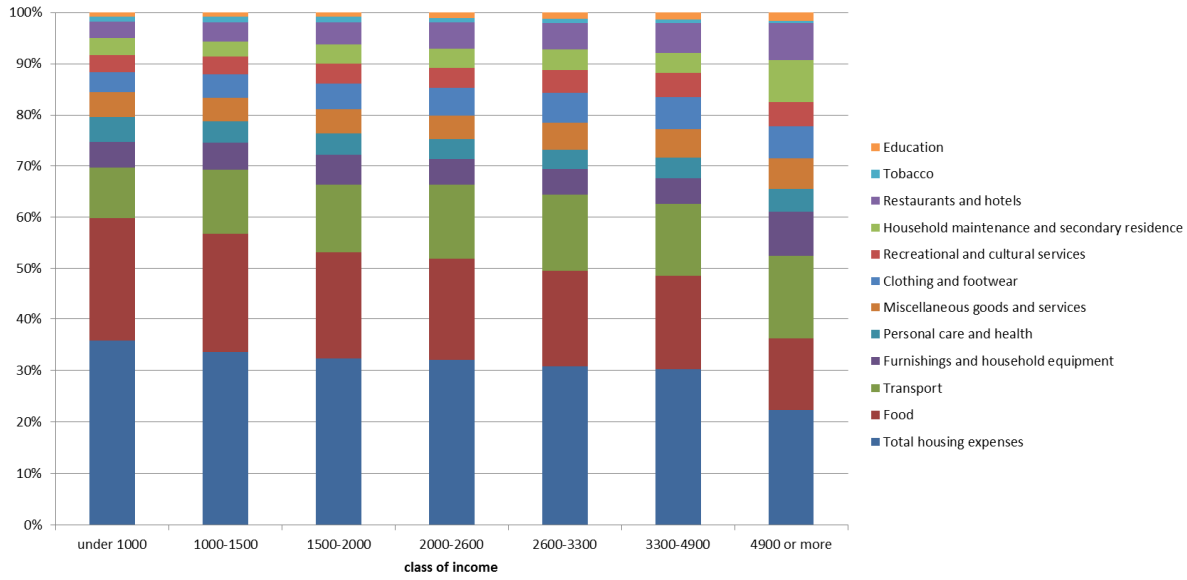


**Table 4.2 Summary of stepwise selection of variables of interest**

| Stepwise selection | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable entered | Presence in survey | | Partial R-square | Model R-square | C(p) | F value | Pr > F |
| | | HBS | SILC | | | | | |
| 1 | Total housing expenses | X | X | 0.3596 | 0.3596 | 104864 | 13002.5 | <.0001 |
| 2 | Food | X | | 0.2031 | 0.5627 | 64265.2 | 10754.2 | <.0001 |
| 3 | Transport | X | | 0.0992 | 0.6619 | 44430.8 | 6796.19 | <.0001 |
| 4 | Class of income | X | X | 0.0523 | 0.7143 | 33972.6 | 4239.88 | <.0001 |
| 5 | Household maintenance and secondary residence | X | | 0.0385 | 0.7527 | 26283.7 | 3602.26 | <.0001 |
| 6 | Restaurants and hotels | X | | 0.029 | 0.7817 | 20490 | 3075.07 | <.0001 |
| 7 | Clothing and footwear | X | | 0.0228 | 0.8045 | 15937.1 | 2698.28 | <.0001 |
| 8 | Furnishings and household equipment | X | | 0.0186 | 0.8231 | 12216.7 | 2437.1 | <.0001 |
| 9 | Number of durable goods | X | X | 0.0184 | 0.8415 | 8544.51 | 2684.48 | <.0001 |
| 10 | Personal care and health | X | | 0.0179 | 0.8594 | 4961.28 | 2953.57 | <.0001 |

*4.2 A predictive model*

This section presents some simulation on HBS data, using different methods of classification. The goal of classification is to build a rule for composing information available on the explanatory variables in HBS, with the aim of allocating observations to the estimated classes. The models that will be used for this purpose are the multinomial logistic regression[3], classification trees[4] and random forest[5]. The dependent variable is monthly household consumption expenditure divided into seven classes using the same monetary thresholds of income classes. In order to verify which set of covariates has better performance in predicting the consumption classes, we select different groups of variables, as shown in the Table 4.3. We tested each set individually and each combination with the common variables.

[3] It is an extension of the general logistic regression that uses independent variables to predict which groups observations belong to. The procedure fits *n-1* separate binary logistic models and measures the probability of being in each categories compared to a reference category.

[4] This method allows to classify the observation in the dataset repeatedly splitting the observation with respect to certain characteristics of explanatory variables. The division produces a tree hierarchy, where the subsets of observations are called nodes. In each nodes test on an attribute is performed. The branches emanating from a decision node design the set of decision alternatives that are available and create different leaves. Each leaf represents a class label and a decision taken after computing all attributes. Each observation is thus classified following a path along the tree that leads from root, the starting point, to a leaf. All possible paths are represented by the branches of the tree, which provide a set of rules of classification, expressed as a function of the dependent variables, for building homogeneous groups with respect to the response variable. We choose CHAID (Chi-squared Automatic Interaction Detection) as growing method: at each step, the algorithm selects from all possible splits, the predictor that has the strongest interaction with the dependent variable.

[5] Ensemble technique, developed by Breiman (2001), that constructs a combination of decision tree from a multitude of different trees. In ensemble terms, each decision tree is a weak learner, while all classifiers taken together are a strong learner. A group of tree can come together to compose a better classification and improve performance. In standard trees, each node is the best split among all characters. Instead, in a random forest the split picked is the best among a random subset of predictors. As a result of this randomness, this strategy performs very well compared to other classifiers and is robust against overfitting.

**Table 4.3 Selected covariates for the model**

| Set of covariates | |
|---|---|
| **SET 1**<br>**Common variables** | Total housing expenses<br>Class of income<br>Macroareas<br>Number of durable goods<br>Education |
| **SET 2**<br>**Most predictive** | Food<br>Transport |
| **SET 3**<br>**Housing related** | Furnishings and household equipment<br>Household maintenance and secondary residence |
| **SET 4**<br>**Food out and clothing** | Restaurants and hotels<br>Clothing and footwear |

4.*3 Comparison between sets of covariates*

Once estimated the classification models, it is necessary to evaluate the results obtained, in order to determine which is the best set of variables to classify the units. Comparing the overall classification error between models and covariates, we can note that every models identify the same set (the union of 1 and 2) and the best model is finally the multinomial logistic model (Table 4.4). The combination of common variables and most predictive ones classifies correctly the 56,3% of total households in HBS survey. This set is also the least demanding in terms of information to be collected. If we build a simplified module in SILC focused on the collection of food and transport costs and given the common variables available, we could have enough information to achieve a trustable prediction of classes of consumption. The percentage of correct predictions, as Table 4.5 shows, is very high for the lowest classes, 79.9% and 69.4% for first and second class, respectively. In addition, we note that despite the low percentages for the highest classes, prediction in classes non-contiguous to the diagonal is very limited.

**Table 4.4 Overall classification error by set of covariates and classification model**

| | Set of covariates | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **1** | **2** | **3** | **4** | **1+2** | **1+3** | **1+4** | **1+2+3+4** |
| **Multinomial** | 39.6 | 39.5 | 29.5 | 31.1 | 56.3 | 45.9 | 45.3 | 75.7 |
| **Classification tree** | 38.9 | 38.5 | 29.2 | 31.0 | 49.6 | 42.0 | 41.4 | 50.2 |
| **Random forest** | 36.6 | 35.3 | 29.3 | 29.0 | 53.5 | 44.2 | 42.7 | 70.0 |

**Table 4.5 Multinomial regression - confusion matrix between observed and predicted class of consumption**

| | PREDICTED | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **OBSERVED** | **under 1000** | **1000-1500** | **1500-2000** | **2000-2600** | **2600-3300** | **3300-4900** | **4900 or more** | **Correct prediction** |
| **under 1000** | 2295 | 576 | 1 | 0 | 0 | 0 | 0 | **79,9%** |
| **1000-1500** | 409 | 2994 | 891 | 15 | 8 | 0 | 0 | **69,4%** |
| **1500-2000** | 62 | 849 | 2669 | 648 | 156 | 15 | 0 | **60,7%** |
| **2000-2600** | 23 | 193 | 998 | 1225 | 919 | 131 | 0 | **35,1%** |
| **2600-3300** | 3 | 78 | 365 | 663 | 1486 | 802 | 10 | **43,6%** |
| **3300-4900** | 8 | 31 | 145 | 198 | 721 | 1556 | 284 | **52,9%** |
| **4900 or more** | 3 | 18 | 43 | 53 | 173 | 618 | 823 | **47,5%** |
| | | | | | **Overall correct prediction** | | | **56,3%** |

## 5. An assessment of consumption questions in a SILC module

The development of a simplified module on consumption in SILC was lunched by the Task Force on the revision of the EU-SILC legal basis meeting on 4-5 March 2014. Between the topics identified from the working group on Income and Living Conditions statistics for fixed every 6-years modules, there is one about over-indebtedness, wealth and consumption. The base for the development of this topic could be found in ad hoc module 2008 'Over-indebtedness and financial exclusion' but is not sufficient: it covers only over-indebtedness and needs to be revised; so there will be a meeting on this topic that is planned for the 17 February 2015. This work and the previous case study on income and wealth (SILC-HBS)[6] can be useful to the new module's development.

As yet underlined in the ESTAT'S work, items that have the largest share in explaining consumption's variance are food, housing and transport expenditures. As described in the previous paragraphs this variables are still best predictors for total variability of consumption (63%). The result confirms that total consumption can be predicted using a limited set of variables, that explain a large fraction of the total variability.

In SILC, a large set of variables about housing costs is available; in HBS most of the components included in European target variable HH070 is collected, so it is possible, as explained previously, to construct an harmonized variable in the two surveys, very similar in the distribution. This common variable is used in order to have a good prediction of classes of consumption; so it is reasonable to think that it is not necessary to collect other information about housing costs in the new module.

There were several exercises in literature that explore the feasibility of imputing consumption values using a limited number of questions (Browning, Crossley and Weber, 2003). In effect the selection of few questions on consumption to introduce in an income survey (and having a large number of questions on other delicate subjects as happen in EU-SILC), it is not an easy task. Several cautions need to be considered in order to identify a short list of variables to be included in SILC, as well as measurement differences between methods of data collection. As suggested by Attanasio et al (2006) recall and diaries data are not perfect substitutes and often the difference is correlated with certain household characteristics. As experienced in COEP, Canadian Out of Employment Panel survey, it is possible to collect this information asking only one question about the total expenditure or asking a non-exhaustive list of sub-items.

[6] See "Data matching – Final report ESTAT study" presented in 7th meeting of the Task-Force on the revision of the EU-SILC Legal Basis, 5-6 December 2013

The first method seems not to be the most feasible for the new module, because it is difficult to choose what include or not in the cues (the list of expenditure that the respondent have to consider in order to answer) of a general question and choose the referent time period[7].
Components of total household expenditures to introduce in the module can be evidently those that explain a good part of total variability of total consumption. Browning, Crossley e Weber (2003) indicate "food at home" and "food outside home", as two predictors that explain a good part of non-durable expenditures. This issue was explored using Canadian data (FAMEX, Survey of Family Expenditures, 1996) and Italian data (SFB, Survey of Family Budgets): the results suggest that imputing the total from the sub-items a great part of total consumption variability can be explained.
According to the structure of total consumption (par. 4.1), asking for "food" in the double form of "food at home" and "food outside home", an increase in the explained total variability of consumption is achieved (from 63% to 66%). As stated before, food, housing and transport expenditures are three good predictors of classes of consumption.
The experience of INSEE is the first about asking consumption in a general purpose survey. As Browning, Crossley and Weber have suggested, twelve questions were added in the monthly consumer confident survey (the French COMME 2008): three about expenditure on food and utilities and eight questions to collect information on household regular expenses on clothing, public transport and other expenditure sub-components (binary variables). The good results using data obtained from this pilot survey and from HBS 2008 have suggested to introduce a similar set of questions in the questionnaire of the Household Wealth Survey runs between 2009 and 2010. The regression model showed a good fit, and the imputation of non-durable consumption showed a good match with HBS 2010 data.
The new module in EU-SILC can benefit from this experience. Collecting information on food (at home and outside) and transport expenditure[8] seems to be the path to follow.
INSEE did not ask the amount of transport expenditure but only two binary variables were collected[9] (about having regular expenses regarding public transport and other transport with motorized vehicle or motorcycle); then an overall question about the expenses for usual monthly consumption that include transport expenditure is asked.
Analyzing HBS's set of questions about transport expenditure, the first thing to notice is that one third of the total amount of transport expenditure is performed by expenses for gasoline or other fuel for cars and motorcycles: this can be a good reason to collect this information separately from the total

[7] For an extended dissertation on this subject see Browning M. et al. "Asking Consumption Questions in General Purpose Surveys" (2006) and Savic M. "Questions about Household Consumption in Surveys"(2007).
[8] In HBS, transport expenditure is collected through a list of items that is expenditure for gasoline, diesel oil, tickets and subscriptions for using public transport, taxis, tolls and parking costs, expenditure for buying a new car, automobile insurance and car's maintenance. In SHIW, total transport expenditure is collected through one general question about the monthly amount of expenditure of "transport (fuel for cars and motorcycles; bus, tram, metro tickets and subscriptions, taxis, parking, motorway tools, not counting cost of trips and vacations)".
[9] In particular COMME's questions are:
Q4-Q11: Over the last 12 months has any member of your household had regular expenses regarding:
- Clothing: (Yes) (No)
- Public transport (train, bus, plane, subway and taxi): (Yes) (No)
- Other transport with motorized vehicle or motorcycle (gas expenses, insurance, etc. but not the vehicle acquisition expenses themselves): (Yes) (No)
- Cultural and recreational goods or services (books, movies, music, concert, museum and art exhibitions, etc.): (Yes) (No)
- Other form of recreational goods or services: (Yes) (No)
- Health (expenses not covered by public or employer insurance scheme): (Yes) (No)
- Children education or childcare: (Yes) (No)
- Personal services (housekeeping, garden keeping, other): (Yes) (No)

Q12: How much do you spend, on an average month, for your usual consumption only (food, clothes, heating, transports, leisure, various services,…), excluding rents, repayments, large expenditure on durables (e.g. buying a car, a refrigerator, a washing-machine, furniture…)?: (Amount)

amount in the new module. A similar proposal can be done regarding to car or motorcycles' insurance that represents one fifth of total expenditure in HBS.
There is another aspect to take into account when observing the composition of transport expenditure in HBS, that is the expense for buying a new car which represents 15% of total expenditure. The collecting of this component of expenditure has an evident problem of periodicity, as a consequence it could be risky to collect the monthly amount.
It is worth noting that for reducing the response burden and preserve a reasonable length of the questionnaire, the use of computer assisted technique in collecting data can be an important advantage: in effect a list of sub-components of consumption can become the first step, in order to ask later the expenditure amount for only components that the respondent will indicate.

## 6. Concluding remarks

A general consensus on the need for distributional measures of economic well-being as a joint function of income, consumption and wealth is wide spreading, even though there is not yet a common framework for their joint collection and analysis. At present the production of integrated statistics on income, consumption and wealth poses several difficulties for National Statistical Institutes and a better exploitation of existing data sources turns out to be an up-to-date challenge for NSIs. The use of administrative archives for statistical purposes is a well-established practice and the combining of survey and administrative sources is considered a primary tool for obtaining relevant data on income or wealth. In this contest, the micro integration techniques may be regarded as a valid alternative for producing statistics on variables not jointly collected in a single survey if some data requirements that are able to facilitate the integration process are really fulfilled.
The current process of modernization of social surveys at European level is going towards a better integration and coordination of surveys also in order to facilitate the matching process. At national level deep efforts of an ex-ante harmonization of common variables, statistical units and concepts in IT-SILC and the Italian HBS have been undertaken. This reconciliation process clearly goes beyond the core social variables and we are confident that will effectively enhance the application of matching techniques and will simplify the estimation of parameters or indicators on the joint distribution of consumption and income.
Our matching exercises have shown the importance of the auxiliary information in improving matching estimates. The presence of few valuable questions about the use of the household income in HBS (e.g. consumption and savings) has allowed us to reconstruct new income classes and compare them with those of IT-SILC. We believe that the inclusion of one or two questions on savings in HBS can be useful for data integration purposes, as well as for improving the quality of information on household monthly income. Similarly another source of auxiliary information could come from the introduction of a small number of questions on food consumption and transport in SILC, combined with with the variable on total housing costs currently present in the survey. As explained in the previous paragraphs, these variables have a great potential and explanatory power so it can be used for estimating a total household expenditures variable in SILC that may act as auxiliary information in the matching procedures.
The main results presented in this paper are finally a part of a work that is in progress. The next step in our matching procedures will be an exercise for extending SM in order to match not just overall consumption expenditures but also the main consumption components. Since IT-SILC presents a long tradition (more than ten years) in using Fiscal Agency archives for the construction of the net and gross income target variables, we are planning to use the register information on income also for HBS respondents. The fiscal income could be eventually used as auxiliary information with respect to the income collected in HBS and the estimated HBS income variable in order to overcome the CIA and improve the accuracy of the final integrated data set.

**References**


Ahemed N., Brzozowki M., Crossley T. F. (2006)"Measurement errors in recall food consumption data", *The Institute for Fiscal Studies*, Wp06/21, October 2006.

Andridge R.R., Little R.J.A. (2009) "The Use of Sample Weights in Hot Deck Imputation". *Journal of Official Statistics*, No. 25, pp. 21-36.

Attanasio O. P., Battistin E. and Leicester A. (2006) "From Micro to Macro, from Poor to Rich: Consumption and Income in the UK and the US", *National Institute Economic Review,* October 2011, vol 218 n. 1: R44-R57.

Breiman, L. (2001). Random forests. *Machine Learning*, October 2001, Volume 45, Issue 1, pp 5-32.

Brewer, M., & O'Dea, C. (2012), *Measuring living standards with income and consumption: evidence from the UK.* Retrieved June 23, 2012, from Institute for Social & Economic Research (ISER) - University of Essex: https://www.iser.essex.ac.uk/publications/working-papers/iser/2012-05.pdf.

Breiman, L. and A. Cutler (2008). *Random forests – Classification manual*; Website accessed in 1/2008; http://www.math.usu.edu/~adele/forest.

Breiman, L., A. Cutler, A. Liaw, and M. Wiener (2006). *Breiman and Cutler's Random Forests for Classification and Regression*. R package version 4.6 – 10.

Browning M. Crossley T.F. Weber G "*Asking Consumption Questions in General Purpose Surveys*" (2003), The Economic Journal, 113(491):F540-F567, 2003. ISSN 1468-0297 http://dx.doi.org/10.1046/j.0013-0133.2003.00168.x.

Coli, A., F. Tartamella, G. Sacco, I. Faiella, M. Scanu, M. D'Orazio, Di Zio, M., Siciliani, I., Colombini, S. and Masi, A. (2005), "*La costruzione di un archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi delle famiglie italiane e l'indagine Banca d'Italia sui bilanci delle famiglie italiane*", Technical Report, Working Group ISTAT- Bank of Italy, Rome.

Conti, P.L., Marella, D., Scanu, M. (2012), "Uncertainty analysis in statistical matching". *Journal of Official Statistics*, No. 28, pp. 69-88.

Donatiello, G., D'Orazio, M.,Frattarola,D., Rizzi, A., Scanu, M., Spaziani, M. (2014) "Statistical Matching of Income and Consumption Expenditures". *International Journal of Economic Sciences*, Vol. III(3), pp. 50 – 65.7

D'Orazio, M. (2013), *StatMatch: Statistical Matching (aka data fusion). R package version 1.2.1*. http://CRAN.R-project.org/package=StatMatch.

D'Orazio, M., Di Zio, M., Scanu, M. (2006), *Statistical Matching: Theory and Practice*. John Wiley & Sons, Chichester, ISBN: 0-470-02353-8.

D'Orazio, M., Di Zio, M., Scanu, M. (2010), "*Old and new approaches in statistical matching when samples are drawn with complex survey designs".* (Sessione specializzata "Matching techniques, censuses and administrative data") Atti della 45ma riunione scientifica della Società Italiana di Statistica, Padova 16-18 giugno 2010.

D'Orazio, M., Di Zio, M., Scanu, M. (2012), "*Statistical Matching of Data from Complex Sample Surveys*". Proceedings of the European Conference on Quality in Official Statistics - Q2012, 29 May - 1 June 2012, Athens, Greece.

Eurostat (2014), Meeting of the task-force on the revision of EU-SILC legal basis: 9th - 17-18 September 2014 - Item n° 3 EU-SILC 2017 ad hoc module: testing (a/part of) new rolling module(s).

Eurostat (2013a), Meeting of the task-force on the revision of EU-SILC legal basis: 7th - 5-6 December 2013 - Item n° 4.1 Sub-topic 2: Different modes of data collection - 4.1 Data matching – Final report ESTAT study, Case study on income and wealth (SILC-HFCS).

Eurostat (2013b), Meeting of the task-force on the revision of EU-SILC legal basis: 8th - 4-5 March 2014 - Item n° 4 Sub-topic 1: SILC contents - 4.1 First discussion on the every 6-year modules.

Eurostat (2011), Implementing core variables in EU social surveys - Methodological guidelines, Luxembourg 2011.

Eurostat (2013c), *Statistical matching: a model based approach for data integration.* Methodologies and Working Paper. Luxembourg: Publications Office of the European Union, 2013.

Meyer, B., & Sullivan, J. (2011), "Further Results on Measuring the Well-Being of the Poor Using Income and Consumption". *Canadian Journal of Economics*, Vol 44, No 1, pp. 52-87.

OECD (2013), *OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth*. Paris OECD Publishing.

Renssen, R. H. (1998), "Use of Statistical Matching Techniques in Calibration Estimation". *Survey Methodology*, No 24, pp. 171-183.

Rubin, D.B. (1986), "Statistical matching with adjusted weights and multiple imputations". *Journal of Business and Economic Statistics*, No 4, pp. 87-94.

Savic M. (2007) "Questions about Household Consumption in Surveys", *PANOECONOMICUS*, 2007, 3, str. 347-357.

Singh A.C., Mantel H., Kinack M., Rowe G. (1993) "Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption". *Survey Methodology*, No. 19, pp. 59-79.

# ANNEXES

## ANNEX 1

| VARIABLE[10] | EU-SILC | | HBS | | HARMONIZZED VARIABLE |
|---|---|---|---|---|---|
| **MARITAL STATUS** | 1 | Never married | 1 | Never married | 1 Never married |
| | 2 | Married and living together | 2 | Married and living together | 2 Married |
| | 3 | Married but living apart | 3 | Married but living apart | 3 Separated/Divorced |
| | 4 | Separated | 4 | Separated | 4 Widowed |
| | 5 | Divorced | 5 | Divorced | |
| | 6 | Widowed | 6 | Widowed | |
| **EDUCATION** | 1 | No educational attainment, illiterate | 1 | Doctorate | 0 Under 15 years old (not applicable) |
| | 2 | No educational attainment | 2 | Master's degree | 1 No educational attainment |
| | 3 | Pre-primary education | 3 | Bachelor's degree | 2 Pre-primary education |
| | 4 | Primary education | 4 | Upper secondary education | 3 Primary education |
| | 5 | Lower secondary education | 5 | Lower secondary education | 4 Lower secondary education |
| | 6 | Upper secondary education | 6 | Primary education | 5 Upper secondary education |
| | 7 | Post-secondary non tertiary education (for example conservatory) | 7 | Pre -primary education | 6 Degree or doctorate |
| | 8 | Bachelor's degree | 8 | No educational attainment | |
| | 9 | Master's degree | | | |
| | 10 | Doctorate | | | |
| **MAIN ACTIVITY STATUS** | 1 | Employed | 1 | Employed | 1 Employed |
| | 2 | Unemployed | 2 | Unemployed | 2 Unemployed |
| | 3 | Fulfilling domestic tasks | 3 | Looking for first job | 3 Fulfilling domestic tasks |
| | 4 | Student | 4 | Fulfilling domestic tasks | 4 Student |
| | 5 | Permanently disabled or/and unfit to work | 5 | Student | 5 Permanently disabled or/and unfit to work |
| | 6 | Retired | 6 | Permanently disabled or/and unfit to work | 6 Retired |
| | 7 | Other condition | 7 | Retired | 7 Other condition |
| | | | 8 | In compulsory military community or service | |
| | | | 9 | Other condition | |
| **CITIZENSHIP** | | Country code | | Country code | 1 Italian |
| | | | | | 2 European (in EU-15) |
| | | | | | 3 European (not in EU-15) |
| | | | | | 4 Rest of the world |

[10] All the individual variables in Annex 1 and Annex 2 are related to the "Reference person".

| VARIABLE | EU-SILC | | HBS | | HARMONIZZED VARIABLE | |
|---|---|---|---|---|---|---|
| **NUMBER OF HOURS WORKED** | 1 | Working full time | 1 | Working full time | 1 | Working full time |
| | 2 | Working part time | 2 | Working part time | 2 | Working part time |
| **PROFESSIONAL STATUS** | | Employee | | Employee | | |
| | 1 | Employee excluding cooperative member | 1 | Manager | 1 | Manager/Executive |
| | 2 | Co.Co.Co. (Temporary subcontractor workers) | 2 | Executive | 2 | Employee |
| | 3 | Occasional self-employed | 3 | Employee | 3 | Workman, Apprentice |
| | 4 | Entrepreneur | 4 | Employee special categories | 4 | Entrepreneur, Cooperative member |
| | 5 | Self-employed professional persons | 5 | Workman | 5 | Self-employed, Co.Co.Co. |
| | 6 | Self-employed | 6 | Other kind of employee | 6 | Self-employed professional persons |
| | 7 | Cooperative member | 7 | Apprentice | | |
| | 8 | Family Worker | 8 | Homeworker | | |
| | | | 9 | Military, policeman | | |
| | | Employee | | Self-employed | | |
| | 1 | Manager | 10 | Entrepreneur | | |
| | 2 | Executive | 11 | Self-employed | | |
| | 3 | Employee | 12 | Professional persons | | |
| | 4 | Workman | 13 | Cooperative member | | |
| | 5 | Apprentice | 14 | Family Worker | | |
| | 6 | Homeworker | 15 | Co.Co.Co (Temporary subcontractor workers) | | |
| | | | 16 | Occasional self-employed workers | | |
| **TYPE OF CONTRACT** | 1 | Temporary job/work contract of limited duration | 1 | Permanent job/work contract of unlimited duration | 1 | Permanent job/work contract of unlimited duration |
| | 2 | Permanent job/work contract of unlimited duration | 2 | Temporary job/work contract of limited duration | 2 | Temporary job/work contract of limited duration |
| **DWELLING TYPE** | 1 | Detached house | 1 | One family dwelling | 1 | One family dwelling |
| | 2 | Semidetached house | 2 | Dwelling shared by more than 1 household | 2 | Dwelling shared by more than 1 household |
| | 3 | Apartment or flat in a building with less than 10 dwellings | 3 | Fixed habitations like a hut or a cave | 3 | Fixed habitations like a hut or a cave |
| | 4 | Apartment or flat in a building with 10 or more dwellings | | | | |
| | 5 | Some other kind of accommodation | | | | |

| VARIABLE | EU-SILC | HBS | HARMONIZZED VARIABLE |
|---|---|---|---|
| **TENURE STATUS** | 1 Tenant or subtenant paying rent<br>2 Owner<br>3 Usufruct<br>4 Accommodation provided rent free | 1 Tenant or subtenant paying rent<br>2 Owner<br>3 Usufruct<br>4 Accommodation provided rent free | 1 Tenant or subtenant paying rent<br>2 Owner<br>3 Accommodation provided rent free |
| **HOUSEHOLD TYPE**[11] | | | 1 Single adult under 35<br>2 Single adult 35-64<br>3 Single adult 65+<br>4 Couple without children (reference person under 35)<br>5 Couple without children (reference person 35-64)<br>6 Couple without children (reference person 65+)<br>7 Couple with 1 child<br>8 Couple with 2 children<br>9 Couple with 3 children<br>10 Single parent<br>11 Other typology (with aggregate members) |

[11] Harmonized family structure is assumed equal to HBS structure to include aggregate members in the residual category.

ANNEX 2

| | **HBS** **Relative frequency** | **EU-SILC** **Relative frequency** | **DIFF** | **HD (%)** |
|---|---|---|---|---|
| CITIZENSHIP | | | | **3.41** |
| Italian | 95.3 | 92.7 | 2.6 | |
| European (in EU-15) | 1.5 | 2.1 | 0.6 | |
| European (not in EU-15) | 1.3 | 2.1 | 0.8 | |
| Rest of the world | 1.9 | 3.1 | 1.2 | |
| MAIN ACTIVITY STATUS | | | | **8.22** |
| Employed | 60.8 | 57.4 | 3.4 | |
| Unemployed | 3.8 | 5.4 | 1.6 | |
| Fulfilling domestic tasks | 3.8 | 6.7 | 2.9 | |
| Student | 0.2 | 0.3 | 0.1 | |
| Permanently disabled or/and unfit to work | 0.6 | 1.5 | 0.9 | |
| Retired | 29.4 | 26.8 | 2.6 | |
| Other condition | 1.6 | 2 | 0.4 | |
| CLASS OF AGE | | | | **2.24** |
| 16-24 | **0.4** | **0.6** | 0.2 | |
| 25-44 | 28.6 | 30.6 | 2 | |
| 45-64 | 45.2 | 42.5 | 2.7 | |
| 65 or more | 25.8 | 26.3 | 0.5 | |
| NUMBER OF HOURS WORKED | | | | **2.54** |
| Not Applicable | 39.2 | 41.6 | 2.4 | |
| Working full time | 57.1 | 53.6 | 3.5 | |
| Working part time | 3.7 | 4.8 | 1.1 | |
| EDUCATION | | | | **1.3** |
| No educational attainment/pre-primary education | 21 | 22.6 | 1.6 | |
| Primary education | 32.5 | 31.3 | 1.2 | |
| Lower/upper secondary education | 34.3 | 34.8 | 0.5 | |
| Degree or doctorate | 12.2 | 11.2 | 1 | |
| PROFESSIONAL STATUS | | | | **6.64** |
| Not Applicable | 39.2 | 41.9 | 2.7 | |
| Manager/Executive | 8.1 | 4.7 | 3.4 | |
| Employee | 18.6 | 15 | 3.6 | |
| Workman | 21.1 | 22.3 | 1.2 | |
| Entrepreneur/Cooperative member | 2.7 | 3.4 | 0.7 | |
| Self-employed/Co.Co.Co | 7.4 | 9.7 | 2.3 | |
| Self-employed professional persons | 2.9 | 3.1 | 0.2 | |
| | | | | |
| | | | | |

| | HBS Relative frequency | EU-SILC Relative frequency | DIFF | HD (%) |
|---|---|---|---|---|
| SEX | | | | **0.53** |
| Male | 76.4 | 75.5 | 0.9 | |
| Female | 23.6 | 24.5 | 0.9 | |
| MARITAL STATUS | | | | **4.04** |
| Never married | 10.4 | 11.1 | 0.7 | |
| Married | 73.2 | 69.7 | 3.5 | |
| Separated/Divorced | 5.9 | 7.9 | 2 | |
| Widowed | 10.6 | 11.2 | 0.6 | |
| TYPE OF CONTRACT | | | | **3.62** |
| Not Applicable | 52.2 | 58.1 | 5.9 | |
| Permanent job/work contract of unlimited duration | 43.1 | 38 | 5.1 | |
| Temporary job/work contract of limited duration | 4.7 | 3.9 | 0.8 | |
| NUMBER OF CHILDREN | | | | **0.5** |
| 0 | 85.3 | 85 | 0.3 | |
| 1 | 10.4 | 10.6 | 0.2 | |
| 2 | 3.9 | 3.9 | 0 | |
| 3 or more | 0.4 | 0.5 | 0.1 | |
| NUMBER OF UNDERAGE PEOPLE | | | | **1.11** |
| 0 | 83.4 | 84.4 | 1 | |
| 1 | 11.9 | 11.3 | 0.6 | |
| 2 | 4.2 | 4 | 0.2 | |
| 3 or more | 0.5 | 0.4 | 0.1 | |
| NUMBER OF YOUNGER PEOPLE | | | | **1.86** |
| 0 | 56.5 | 55 | 1.5 | |
| 1 | 26.9 | 26.9 | 0 | |
| 2 | 15.3 | 17 | 1.7 | |
| 3 or more | 1.3 | 1 | 0.3 | |
| NUMBER OF ADULTS | | | | **1.8** |
| 0 | 41.1 | 43.6 | 2.5 | |
| 1 | 29.7 | 28.8 | 0.9 | |
| 2 | 28.8 | 27.3 | 1.5 | |
| 3 or more | 0.4 | 0.3 | 0.1 | |
| NUMBER OF ELDERLY PEOPLE | | | | **0.43** |
| 0 | 63.7 | 63.3 | 0.4 | |
| 1 | 24.3 | 24.5 | 0.2 | |
| 2 | 11.8 | 12 | 0.2 | |
| 3 or more | 0.1 | 0.2 | 0.1 | |

| | HBS Relative frequency | EU-SILC Relative frequency | DIFF | HD (%) |
|---|---|---|---|---|
| NUMBER OF EMPLOYED PEOPLE | | | | **0.95** |
| 0 | 36.7 | 37.8 | 1.1 | |
| 1 | 38.1 | 37.9 | 0.2 | |
| 2 | 22.3 | 21.6 | 0.7 | |
| 3 or more | 2.9 | 2.7 | 0.2 | |
| NUMBER OF EMPLOYEES | | | | **3.22** |
| 0 | 45.5 | 49.7 | 4.2 | |
| 1 | 35.9 | 34.3 | 1.6 | |
| 2 | 16.7 | 14.4 | 2.3 | |
| 3 or more | 1.9 | 1.7 | 0.2 | |
| NUMBER OF SELF-EMPLOYED | | | | **4.7** |
| 0 | 85.7 | 80.8 | 4.9 | |
| 1 | 12.1 | 16.4 | 4.3 | |
| 2 | 1.9 | 2.6 | 0.7 | |
| 3 or more | 0.2 | 0.2 | 0 | |
| NUMBER OF RETIRED INCOME | | | | **4.4** |
| 0 | 61.1 | 66.3 | 5.2 | |
| 1 | 28.2 | 25.9 | 2.3 | |
| 2 | 10.4 | 7.6 | 2.8 | |
| 3 or more | 0.3 | 0.1 | 0.2 | |
| NUMBER OF OTHER TYPE OF INCOME | | | | **5.77** |
| 0 | 45.2 | 37.8 | 7.4 | |
| 1 | 26.8 | 32.9 | 6.1 | |
| 2 | 16.6 | 17 | 0.4 | |
| 3 or more | 11.4 | 12.3 | 0.9 | |
| NUMBER OF MEN IN THE HOUSEHOLD | | | | **0.69** |
| 0 | 21.4 | 21.7 | 0.3 | |
| 1 | 49.6 | 49 | 0.6 | |
| 2 | 21.4 | 22 | 0.6 | |
| 3 or more | 7.6 | 7.3 | 0.3 | |
| NUMBER OF WOMEN IN THE HOUSEHOLD | | | | **0.29** |
| 0 | 13.7 | 13.7 | 0 | |
| 1 | 57.5 | 57.7 | 0.2 | |
| 2 | 21.9 | 21.6 | 0.3 | |
| 3 or more | 6.9 | 7 | 0.1 | |
| | | | | |
| | | | | |

| | HBS Relative frequency | EU-SILC Relative frequency | DIFF | HD (%) |
|---|---|---|---|---|
| NUMBER OF INCOME EARNERS | | | | **14.27** |
| 0 | 1.6 | 1 | 0.6 | |
| 1 | 54.8 | 40.7 | 14.1 | |
| 2 | 36.8 | 39.9 | 3.1 | |
| 3 or more | 6.8 | 18.5 | 11.7 | |
| INCOME CLASSES | | | | **12.9** |
| under 1000 | 20.5 | 16.6 | 3.9 | |
| 1000-1500 | 23.1 | 16.8 | 6.3 | |
| 1500-2000 | 13.1 | 16.3 | 3.2 | |
| 2000-2600 | 9.8 | 12.7 | 2.9 | |
| 2600-3300 | 9.7 | 14.3 | 4.6 | |
| 3300-4900 | 9.1 | 15.2 | 6.1 | |
| 4900 or more | 14.7 | 8.1 | 6.6 | |
| HOUSE SIZE IN SQUARE METERS | | | | **4** |
| until 50 | 7.2 | 8.8 | 1.6 | |
| 51-75 | 23.3 | 23.6 | 0.3 | |
| 76-100 | 39.1 | 41.2 | 2.1 | |
| 101-125 | 14.4 | 13.1 | 1.3 | |
| 126-150 | 8.2 | 7.6 | 0.6 | |
| 151-175 | 2.3 | 1.7 | 0.6 | |
| 176-200 | 1.7 | 1.2 | 0.5 | |
| over 200 | 3.8 | 2.8 | 1 | |
| TENURE STATUS | | | | **3.54** |
| Tenant or Subtenant paying rent | 18 | 18.4 | 0.4 | |
| Owner | 72.3 | 68.9 | 3.4 | |
| Accommodation provided rent free | 9.7 | 12.7 | 3 | |
| YEAR OF BUILDING CONSTRUCTION | | | | **6.6** |
| after 2000 | 10 | 8.5 | 1.5 | |
| 1995-1999 | 4.6 | 4.6 | 0 | |
| 1990-1994 | 5.8 | 5.8 | 0 | |
| 80's | 14.4 | 13.6 | 0.8 | |
| 70's | 19.3 | 17.9 | 1.4 | |
| 60's | 18.7 | 18.2 | 0.5 | |
| 50's | 11.1 | 12.5 | 1.4 | |
| 1900-1949 | 13.1 | 12.3 | 0.8 | |
| before 1900 | 3 | 6.7 | 3.7 | |
| | | | | |
| | | | | |

| | HBS Relative frequency | EU-SILC Relative frequency | DIFF | HD (%) |
|---|---|---|---|---|
| MACROAREAS | | | | **0.05** |
| North-West | 28.5 | 28.5 | 0 | |
| North-East | 19.8 | 19.8 | 0 | |
| Centre | 19.8 | 19.8 | 0 | |
| South | 21.1 | 21.2 | 0.1 | |
| Islands | 10.7 | 10.7 | 0 | |
| HOUSEHOLD TYPE | | | | **2.14** |
| Single adult under 35 | 3.1 | 4 | 0.9 | |
| Single adult 35-64 | 12.8 | 12 | 0.8 | |
| Single adult 65+ | 14.9 | 15 | 0.1 | |
| Couple without children (reference person under 35) | 1.5 | 1.7 | 0.2 | |
| Couple without children (reference person 35-64) | 7.2 | 7.1 | 0.1 | |
| Couple without children (reference person 65+) | 11.1 | 10.5 | 0.6 | |
| Couple with 1 child | 16.4 | 16.7 | 0.3 | |
| Couple with 2 children | 15.2 | 15.2 | 0 | |
| Couple with 3 children | 3.8 | 3.4 | 0.4 | |
| Single parent | 8.5 | 8.4 | 0.1 | |
| Other typology (with aggregate members) | 5.6 | 5.8 | 0.2 | |
| NUMBER OF DURABLE GOODS | | | | **1.4** |
| 4 | 10.9 | 11.8 | 0.9 | |
| 5 | 11.7 | 11.3 | 0.4 | |
| 6 | 15.7 | 15.4 | 0.3 | |
| 7 | 19.8 | 18.8 | 1 | |
| 8 | 23.6 | 24.1 | 0.5 | |
| 9 | 18.3 | 18.7 | 0.4 | |
| WC | | | | **1.86** |
| Absence | 1.1 | 0.6 | 0.5 | |
| Presence | 98.9 | 99.4 | 0.5 | |
| BATHROOM | | | | **0** |
| Absence | 0.7 | 0.7 | 0 | |
| Presence | 99.3 | 99.3 | 0 | |
| HOT WATER SUPPLY | | | | **0.85** |
| Absence | 0.8 | 1.1 | 0.3 | |
| Presence | 99.2 | 98.9 | 0.3 | |
| | | | | |
| | | | | |
| | | | | |

| | HBS Relative frequency | EU-SILC Relative frequency | DIFF | HD (%) |
|---|---|---|---|---|
| GARAGE | | | | **1.96** |
| Absence | 42.5 | 39.8 | 2.7 | |
| Presence | 57.5 | 60.2 | 2.7 | |
| MAIN INCOME SOURCE | | | | **9.17** |
| No income | 2.8 | 2.1 | 0.7 | |
| Employee income | 45.4 | 44.7 | 0.7 | |
| Self-employment income | 15.2 | 15.9 | 0.7 | |
| Old-age, survivors and disability benefits | 30.8 | 31.5 | 0.7 | |
| Unemployment benefits | 1.7 | 3.3 | 1.6 | |
| Interest, dividend, profit from capital investments | 0.3 | 1.7 | 1.4 | |
| Inter-households cash transfer received | 3.8 | 0.7 | 3.1 | |