Measuring Income, Consumption and Wealth jointly at the micro-level

Pierre Lamarche
Eurostat

June 20, 2017

PRELIMINARY VERSION - PLEASE DO NOT QUOTE

1 Introduction

Income, consumption and wealth are essential variables for the description of households' economic behaviours. In particular the way these dimensions interplay tells a lot regarding asset accumulation, consumption behavior and so on. This paper presents the result obtained with one particular method making it possible to estimate the joint distribution of income, consumption and wealth, namely statistical matching, relying on existing data. We also address the uncertainty issue which is essential when dealing with statistically matched data.

2 Strategy

The implementation of modules in a survey such as EU-SILC appears as a desirable solution. Further integration may be fostered as well, yet the negative effect of such integration may be accounted for.

Experiments of such modules have been started, with the inclusion of a module on Over-indebtedness, Consumption and Wealth in the 2017 wave of EU-SILC for volunteering countries. The module encompasses a short list of questions on the different aspects of household expenditure and assets. It complements the already existing variables on indebtedness, extending the field of analysis to non-mortgage loans. The additional questions on over-indebtedness are aimed at complementing already existing questions on arrears on mortgage repayments, rents and utility bills, asking for amounts for such arrears. It embeds a final question on assistance received to alleviate the impact of over-indebtedness. It also provides a rough estimation of assets, as the respondents are requested to provide participation and value for the real estate properties and for financial assets (deposits, bonds, shares publicly traded and mutual funds). Finally, regarding consumption, the items tat are collected through short questions are the following ones:

- Food at home
- Food outside home
- Public transport
- Private transport

In addition, a question on regular savings is also asked to the respondents. The experiments on these modules will be conducted on some EU countries (7 for the Over-indebtedness sub-module, 12 for the Consumption and Wealth sub-module, see Table 1).

On the short run, we use the available data in order to estimate the joint distribution of income, consumption and wealth. We first perform statistical matching between EU-SILC and HBS data; then for the countries belonging to the euro area we also experiment a matching between the SILC-HBS fused data and the HFCS data, in order to achieve, for the countries where this is possible, a full estimation of the joint distribution of

Sub-modules	Participating countries
Over-indebtedness	Hungary, Latvia (sub-sampling), Luxembourg, Malta, Portugal (sub-sampling), Switzerland, Norway
Consumption & Wealth	Belgium, Czech Republic, Finland (sub-sample), Iceland, Italy, Latvia (sub-sample), Lithuania, Netherlands, Austria (partly), Portugal (sub-sample), Sweden, United Kingdom

Table 1: List of countries participating in the Over-indebtedness, Consumption and Wealth module for EU-SILC 2017

income, consumption and wealth. In the rest of the paper, we denote X as the variables that are common to both HBS and EU-SILC, Y the variable(s) of interest for EU-SILC (and observed only in this survey – say disposable income and poverty – and Z the variable(s) of interest for HBS – say consumption.

2.1 Matching between EU-SILC and HBS

2.1.1 Comparability issue

As pointed out by [1], the comparability across surveys is critical for the quality of the statistical matching. Therefore, after having determining the various potential matching variables existing in both survey (and possibly having performed some harmonization works), we assess their comparability thanks to the usual indicators (Hellinger distance, Q-Q plots, etc., following [2]). Once we have selected the variables that are the most comparable across SILC and HBS, we may perform the matching.

The first and simple criterion we retain to gauge the comparability of the variables is the Hellinger distance, defined as follows for two probability measures P and Q:

$$H(P,Q) = \sqrt{\frac{1}{2}\int \left(\sqrt{\frac{dP}{d\lambda}} - \sqrt{\frac{dQ}{d\lambda}}\right)^2 d\lambda} \tag{1}$$

Applied to categorical variables V_1 and V_2 with K categories, the expression of the Hellinger distance becomes:

$$H(V_1,V_2) = \sqrt{\frac{1}{2}\sum_{k=1}^{K} \left(\sqrt{\mathbb{P}(V_1 = k)} - \sqrt{\mathbb{P}(V_2 = k)}\right)^2} \tag{2}$$

One of the drawbacks of the Hellinger distance, as stated in [2], is that it does not account for sampling error. $\mathbb{P}(V_1 = k)$ and $\mathbb{P}(V_2 = k)$ are estimated thanks to the Horvitz-Thomson estimator; yet no consideration is given to the variance associated to the estimation of such quantities. A rule-of-thumb consists of taking 0.05 as a threshold for the Hellinger distance, above which the variable should not be considered as an acceptable matching variable.

Other criteria may also be used, using inference techniques accounting also for the variance due to sampling (provided that the estimation of such variance is available for both surveys). For instance, when it comes to categorical variables, a Chi-2 test may also fit for testing the equality of the distributions.

Now regarding the specific issue of continuous variables (essentially income), classical tests for testing the equality of distribution may be implemented, such as Kolmogorov-Smirnov test. Nevertheless, it is important to properly assess the uncertainty related to sampling; if such an estimation is available for EU-SILC, this is not the case for HBS in general. One way to deal with discrepancies in the measurement of income consists of assuming the preservation of the ranking: even if the concepts and modes of collection vary across surveys, we assume that the way households are ranked remains the same.

Variables	SILC variable	HBS variable
Population density level	DB100	HA09
Household size	o.c.	o.c.
Household type	o.c.	o.c.
Age of the reference person	PB140	MB03
Level of education of the reference person	PE040	MC01
Activity status of the reference person	PL031	ME01
Occupation status of the reference person	PL050	ME0988
Tenure status	HH020/HH021	derived from EUR_HE0411
Rents paid by tenants	HH060	EUR_HE0411
Main source of income	o.c.	HI11
Income	HY010/HY020	EUR_HH095

Table 2: Potential matching variables between EU-SILC and HBS

Note: o.c. stands for "Own computation". This means that the information was derived from several variables in the data.

2.1.2 Method 1: random hot-deck

A first very conservative approach consists of performing the matching through random hot-deck: the respective samples of SILC and HBS are stratified according the n matching (categorical) variables $X_1, ..., X_n$ and every household belonging to a given stratum s in SILC is allocated the data coming from a randomly selected household in the same stratum s in HBS. Hence the matched households share the exact same characteristics $(X_1, ..., X_n)$.

One of the drawbacks of this method is that it becomes swiftly difficult to define a stratification that will account for possible combinations of characteristics, even though the number of matching variables is fairly limited. It becomes then necessary to proceed with relevant regrouping of categories; however the stratification of the samples may be performed thanks to a very systematic approach, accounting for a reasonable minimal size of the strata and at the same time for the most relevant variables in terms of matching. The algorithm to determine the optimal stratification may be described as follows:

- at step 1, selection of at most the n_1 most relevant variables with respect to explaining consumption variations. This entails a backward selection model resting on a simple OLS estimation.
- the stratification according to variables $X_1^{(1)}, ..., X_{n_1}^{(1)}$ is computed on both samples and the relative size of each cell k is evaluated. The stratification (hence the matching variables $X_1^{(1)}, ..., X_{n_1}^{(1)}$) will be retained if and only if there is enough donors in HBS compared to the receivers in SILC for each stratum; the threshold accounts for the relative sample size and is defined as follows:

$$\frac{s_{k,r}}{s_{k,d}} \geq c\,\frac{s_r}{s_d} \tag{3}$$

  where c is a constant; it is set with a rule-of-thumb to 3 in this case.
- if the threshold is not exceeded for at least 90% of the sample, the stratification is retained and the hot-deck is performed. Otherwise, the process is reiterated with the selection of $n_2 = n_1 - 1$ variables.

The backward selection of the variables consists more precisely of the elimination step by step of the less relevant variables in order to select the most optimal subset of variables according to a given criterion (which may be adjusted R^2, Akaike information criterion (also denoted AIC), Bayesian information criterion (or BIC), Mallow's C_p, etc.). In this specific case, we define for each iteration k a maximal size for the subset of variables n_k which decreases by 1 at the end of each loop. This means that the selected model should not necessarily embed n_k variables; in practice, this is always the case with our data.

Such a systematic selection of variables makes then it possible to perform matching for all EU countries, leaving the possibility that the list of matching variables varies across countries. From this viewpoint, the harmonization concerns the process of selection, but it takes into account potential specificities of the countries, provided that the matching variables reflect such specificities.

2.1.3 Method 2: a mixed approach

The first method has the advantage of simplicity; but it cannot comprehend all potential variables that could bring additional insights on consumption behaviors. A second approach consists of using a semi-parametric approach for the matching, following description given by [1]. We adopt in this paper a method combining a regression step involving basic OLS and a matching step using rank hot-deck.

The first step consists of using all the potential matching variables in order to estimate $\mathbb{E}(c|X_1, ..., X_n)$ through OLS in HBS data. Few regards are given to the specification of the model, as we are in this case not interested in the estimation of the level, and the estimated equation is the following one:

$$\hat{c} = \hat{\beta}'X + \hat{u} \tag{4}$$

Then the parameters $\hat{\beta}$ and the residuals $\hat{u}$ are used to estimate $\mathbb{E}(c|X_1, ..., X_n)$ in SILC data; such estimation is denoted $\tilde{c}$. Both variables $\hat{c}$ and $\tilde{c}$ are used in order to perform rank hot-deck on the data. The underlying assumption is that even if the level of expenditures is not very well estimated thanks to the regression, the ranking of households is preserved. This is in particular the core idea in papers experiencing matching between HBS and SILC such as in [3]. In order to test this assumption, it is possible to compute Spearman's coefficient of correlation ρ or Kendall's τ.

The rank hot-deck method is a determistic method that consists of resting on the empirical cumulative distribution functions for a variable in two different datasets in order to match the individuals according to the closest rank. In our case, considering the empirical cumulative distribution functions $\hat{F_{\hat{c}}}$ and $\hat{F_{\tilde{c}}}$, for the household i belonging to the SILC sample s_{SILC}, the household j in the HBS sample s_{HBS} so that:

$$j = \underset{k \in s_{HBS}}{Argmin}(|\hat{F_{\tilde{c}}}(\tilde{c_i}) - \hat{F_{\hat{c}}}(\hat{c_k})|) \tag{5}$$

One drawback of this method is that the matching only focuses on the rank of the households, regardless their characteristics (that still play a role for the computation of the rank, but depending on the results of the model, households with very different characteristics may be closely ranked). Therefore, in order to avoid too unrealistic matching, the samples are stratified according to the type of household.

2.1.4 Consistency and reweighting

The estimations made out of the fused dataset obtained after the matching follow the classical framework of the Horvitz-Thomson estimator. The weights used for the estimation are the ones provided in the EU-SILC User DataBase (variable `DB090`). These weights are already calibrated according to margins defined at the national level[1]; nevertheless, as pointed out by [1] and [4], it may be useful to perform a second step of calibration in order to account for results coming from the HBS data. Indeed, results from the fused dataset will probably not be consistent with the ones given by HBS, as the observations nor the weights are not the same; this may result in confusing the users with different results according to sources that are supposed to be close.

[1] It is possible to report to the national quality reports in order to check which margins have been used for the calibration step.

However, there is a trade-off between bias and variance, since imposing a high number of margins during the calibration may introduce much volatility in the final weights (additionally, the weights have also already been calibrated a first time and the re-calibration may have noxious effects on the variance). A list of margins is defined based on already published indicators (available on Eurostat's website, in Eurobase), both coming from EU-SILC and HBS data:

- mean consumption (at the household level, table hbs_exp_t111 in Eurobase)
- mean consumption by COICOP category, corrected for the Purchase Power Standard (PPS) (at the household level, table hbs_exp_t121 in Eurobase)
- mean PPS consumption by income quintile (at the household level, table hbs_exp_t133 in Eurobase)
- equivalized income deciles (at the individual level, table ilc_di01 in Eurobase)
- monetary poverty rate (at the individual level, table ilc_li02 in Eurobase)

2.1.5 Estimation of uncertainty

Following [5] among others, we replicate the process of matching a high number of times (1,000 in this case) in order to obtain an estimation of the uncertainty related to the imputation itself; this allows also to obtain estimates of parameters that do not rely on a particular matching between the two datasets. As commonly is the case for multiply imputed data, the estimation process may turn out to be computationally demanding, especially when it comes to the calculation of non-linear estimators such as quantiles or ratios. Hence, for M implicates of the same matching process, the parameter of interest θ is estimated thanks to the M estimations of θ obtained on the different implicates, as follows:

$$\hat{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \tag{6}$$

In other words, the estimation of the median is given by the average of the M estimations of the median (which may be quite long to be computed). The variance associated to the estimation is given by the following formula:

$$\hat{V}(\hat{\theta}) = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\theta}_m - \hat{\theta}\right)^2 \tag{7}$$

This between-imputation variance should be combined with the between-imputation variance reflecting the error due to sampling, following Rubin's formula [5]; nevertheless, the within-imputation variance is not considered here. The confidence interval for $\hat{\theta}$ may then be obtained thanks to the usual formula derived from the central-limit theorem, provided that $\hat{\theta}$ follows a normal law. In the case of estimators for quantile, this conclusion requires further conditions on the density function of the estimator. May these conditions not verified, it is always possible to estimate the confidence interval through non-parametric techniques. For instance, facing the M sorted implicates $(\hat{\theta}_{(1)}, ..., \hat{\theta}_{(M)})$, one can take the 2.5$^{\text{th}}$ and the 97.5$^{\text{th}}$ implicates as the lower and upper bounds of the confidence interval for a level at 95%.

Moreover, this estimation of variance does not reflect uncertainty in the broad sense of the term. Indeed, the between-imputation variance corresponds to the volatility of the estimates obtained through the matching; as the selection of matching variables rests on a procedure similar to ANOVA, the more homogeneous the strata are, the less volatile the estimates will be. But the entire procedure relies on the Conditional Independence Assumption, whose reliability represents as such a source of uncertainty.

There are ways to relax the CIA and estimate the uncertainty. A first very rough approach consists of computing Fréchet bounds for categorical variables, as in [2]. The Fréchet bounds are obtained thanks to very general algebraic formulas, applied to categorical variables Y and Z, and integrated over the distribution of the matching variable X:

$$Pmin(Y = j, Z = k) = \sum_{i=1}^{I} \mathbb{P}(X = i).min(\mathbb{P}(Y = j|X = i), \mathbb{P}(Z = k|X = i)) \quad (8)$$

$$Pmax(Y = j, Z = k) = \sum_{i=1}^{I} \mathbb{P}(X = i).max(0, \mathbb{P}(Y = j|X = i) + \mathbb{P}(Z = k|X = i) - 1) \quad (9)$$

These bounds are easily computed (using for instance in R the command `Frechet.bounds.cat` associated to the package `StatMatch` [6]) on categorical data.

When computing the Fréchet bounds, we do not account for an additional piece of information that we have at our disposal, the marginal distributions of Y and Z. One solution may consist of using the Iterative Proportional Fitting (IPF) algorithm in order to generate multiply-imputed data relaxing the CIA and reproducing the expected marginal distributions. This solution is suggested in different articles, such as [7], [8].

Regarding continuous variables, relaxing the CIA proves to be more difficult. It is possible to compute multiply-imputed data that do not rely on this assumption, thereby providing an idea of the range of plausible values, as described in [9]. However, the Bayesian algorithm described by Rässler requires alternative assumptions; in particular, it assumes that the variables Y and Z follow a normal law, which may be challenged in our case.

An alternative could consist of relying on the preceding methodology, *i.e.* using the Frechet bounds to generate plausible distributions (for instance plausible joint distributions in terms of income and consumption quantiles, thereby making categorical a continuous variable). Once this is done, we face multiply-imputed granular data, on which it is possible to estimate for instance the range of values taken by say the median.

As it is granular, it is also important the granularity to be thin enough so as to accurately estimate distributional estimators such as the median. We can test how much information we lose through this transformation of continuous variables into categorical variables; for instance, it is possible to compare between- and within-variance (V_B and V_W respectively) to the total variance V, as it is commonly done for such analyses. We can also rest on Malahanobis decomposition [10][11] in a very similar way and decompose the Gini coefficient as follows:

$$G = G_B + \sum_{k=1}^{K} a_k \, G_k + R \tag{10}$$

using denotations in [12]. The residual term R may be discarded as there is no overlap between the groups when dividing the population by quantiles of the variable on which the Gini index is computed. Then the higher the ratio $\frac{G_B}{G}$, the more the decomposition preserves the heterogeneity.

2.2 Matching between SILC-HBS and HFCS

Once the matching between EU-SILC and HBS is performed, the question of wealth remains open, as very few pieces of information on assets (quite a bit more on indebtedness) are available in both surveys. In order to measure information on (net) wealth, we use the Eurosystem Household Finance and Consumption Survey (HFCS), and more specifically the first wave of this survey [13] [14].

The first point to be borne in mind is that the HFCS does unfortunately not cover all EU countries, as it mainly involves countries belonging to the euro area. Therefore the joint distribution of wealth and income/consumption can only be estimated for those countries that have conducted the HFCS in 2010 (*i.e.* a bit more than the half of the EU countries, 15).

Several options are available for performing the estimation of the joint distribution of wealth and income/consumption. One possibility consists of using the HFCS sample for the estimation, as there exists some "hook" variables (or proxies) that make it easier to match information coming from different surveys. Indeed, regarding income, the HFCS already collects comprehensive information on gross income which can be compared with gross income as measured in EU-SILC[2]. Hence, disposable income coming from EU-SILC could be brought into the HFCS data thanks for instance to a rank hot-deck procedure on gross income. Regarding consumption, the HFCS has followed recommendations by [15] and encompasses short questions on consumption that make it possible to impute total consumption in the data [16][17].

We explore another possibility which consists of using our fused SILC-HBS data and matching it with the HFCS data when available. We then keep on relying on the same sample which has some advantages in term of consistency for the different estimations we want to make out of this exercise. We then adopt a very mere approach, using

[2] for a comparison of the figures obtained for the first wave, see the HFCS methodological report [14].

rank hot-deck on gross income and stratifying the samples on common variables that account for consumption and wealth. In this case, the rank hot-deck makes it possible to consider a measurement-error model: we implicitly assume that the levels of gross income measured for the different surveys may differ, but the ranking between households remains the same. We stratify the samples of SILC-HBS and HFCS according to the tenure status (home-owner with mortgage, home-owner without any mortgage, tenant), the quartile of food consumption and the household type and we apply rank hot-deck in each stratum.

3 Data

For the matching exercise, we use data coming from the three European surveys dealing with income, consumption and wealth, namely EU-SILC, HBS and HFCS. We face a first constraint regarding the synchronization of the different surveys: if EU-SILC is conducted on a yearly basis, in many countries, HBS is conducted every 5 years, without any coordination across European countries. This is the reason why the results that can be drawn out of this exercise will be valid "around" 2010; for each country, we take the SILC wave corresponding to the reference period for HBS, following information in Table 3.

Once SILC data are selected in accordance with the HBS reference period, we investigate the comparability issue between the potential matching variables following the framework defined in section 2. The candidates for the matching are listed in Table 2; they result in some cases from ex-post harmonization, as indicated in the table.

In particular, the definition of a reference person in the household is crucial in order to incorporate the dimension of age, education or labour status in the analysis. From this viewpoint, the reference person has to be determined in a consistent way in the two surveys. There is no *a priori* harmonization to this regard; EU-SILC does even not consider any reference period, since poverty is classically analyzed at the individual level. We chose the definition adopted by the Canberra group [18] in its handbook on household income statistics, following a list of criteria to be applied on the household members until one unique person is selected:

- one of the partners in a registered or *de facto* marriage, with dependent children
- one of the partners in a registered or *de facto* marriage, without dependent children
- a lone parent with dependent children
- the person with the highest income
- the eldest person

The advantage of this definition is that there is no need for designing the questionnaire in order to collect additional data, provided that the links between the different household members are already described. This is not the case for alternative definitions, such as

Countries	Reference year for HBS	Reference year for HFCS
AT	2010	2010
BE	2010	2010
BG	2010	-
CY	2009	2010
CZ	2010	-
DE	2008	2010
DK	2009	-
EE	2010	-
EL	2010	2009
ES	2010	2008
FI	2012	2009
FR	2010	2010
HR	2010	-
HU	2010	-
IE	2010	-
IT	2010	2010
LT	2008	-
LU	2010	2010
LV	2010	-
MT	2008	2010
NL	2010	2009
PL	2010	-
PT	2010	2010
RO	2010	-
SE	2009	-
SI	2010	2010
SK	2010	2010
UK	2010	-

Table 3: Reference period for HBS 2010 and HFCS first wave

Sources: Eurostat, European Central Bank, [13].

the respondent or the financially knowledgeable person (FKP). The Canberra definition is also used in other surveys such as the HFCS [13]. Its concrete implementation in SILC and HBS data is subject to assumptions, as the matrix giving the relationships between all household members is not always fully available. For more details, please report to Annex A.

3.1 Comparison of the variables

3.1.1 Demographics

Following the framework described in section 2.1.1, we compute the Hellinger distance for the different categorical variables that could be used as matching variables. As shown in table 4, the Hellinger distance may be very high in some cases; as already explained previously, the distance has to be less than 0.05 to consider the variables equally distributed. Based on the results obtained, variables such as level of education of the reference person, its occupation status and the household's main source of income cannot qualify to be matching, as they seem very differently distributed depending on the survey that is considered. This result is confirmed by other types of metrics, and annex B makes it possible to compare the distributions.

Country	Density level	Household size	Household type	Age of RP	Level of education	Activity status	Occupation status	Tenure status	Main source of income
AT	0.008	0.005	0.029	0.041	0.996	0.096	0.894	0.012	0.506
BE	0.025	0.005	0.053	0.165	0.334	0.149	0.218	0.028	0.439
BG	0.189	0.089	0.106	0.129	0.092	0.170	0.436	0.062	0.557
CY	0.014	0.024	0.042	0.032	0.093	0.090	0.314	0.065	0.433
CZ	0.007	0.070	0.153	0.058	0.083	0.029	0.954	0.276	0.374
DE	0.039	0.001	0.028	0.047	0.042	0.122	0.926	0.015	0.409
DK	0.202	0.046	0.082	0.039	0.901	0.111	0.315	0.042	0.522
EE	0.011	0.020	0.039	0.048	0.048	0.098	0.454	0.011	0.437
EL	0.025	0.001	0.022	0.045	0.093	0.088	0.856	0.006	0.511
ES	0.013	0.038	0.041	0.031	0.246	0.104	0.153	0.148	0.480
FI	0.226	0.003	0.009	0.040	0.942	0.177	0.625	0.023	0.453
FR	0.141	0.008	0.016	0.048	0.503	0.090	0.359	0.024	0.458
HR	0.012	0.034	0.037	0.047	0.028	0.075	0.370	0.022	0.415
HU	0.002	0.001	0.013	0.038	0.325	0.099	0.074	0.014	0.532
IE	0.026	0.005	0.016	0.096	0.264	0.106	0.763	0.022	0.256
IT	0.116	0.003	0.018	0.035	0.147	0.170	0.855	0.010	1.000
LT	0.081	0.067	0.078	0.062	0.086	0.077	0.387	0.004	0.439
LU	0.020	0.020	0.065	0.083	0.185	0.065	0.413	0.082	1.000
LV	0.092	0.042	0.058	0.052	0.063	0.069	0.414	0.163	0.515
MT	0.038	0.024	0.060	0.051	0.918	0.090	0.302	0.104	0.302
PL	0.004	0.002	0.054	0.071	0.131	0.219	0.308	0.009	0.511
PT	0.007	0.026	0.035	0.045	0.331	0.107	0.070	0.011	0.435
RO	1.000	0.053	0.063	0.072	0.377	0.079	0.325	0.025	0.506
SE	0.029	0.016	0.142	0.166	0.957	0.909	0.793	0.132	0.456
SI	0.043	0.008	0.032	0.057	0.971	0.134	0.358	0.544	0.519
SK	0.073	0.038	0.083	0.060	0.097	0.092	0.938	0.019	0.508
UK	0.126	0.019	0.020	0.039	0.879	0.297	0.291	0.001	0.368

Table 4: Hellinger distance between HBS and EU-SILC for the potential matching variables

Country	Mean	Q10	Q25	Median	Q75	Q90
AT	−34.4	−65.0	−43.1	−34.5	−29.2	−27.5
BE	−5.2	−0.6	−3.3	−3.1	−4.0	−4.0
BG	−40.4	14.1	32.5	−20.5	−42.9	−59.7
CY	−7.8	−19.0	−23.3	−11.9	−3.8	−4.8
CZ	−65.0	−65.5	−67.1	−68.5	−68.2	−62.1
DE	13.6	14.0	17.6	16.1	14.9	12.5
DK	−9.9	−14.0	−3.7	−2.8	−6.5	−9.0
EE	10.8	15.7	20.0	16.7	22.9	−2.8
EL	0.2	0.0	0.0	0.0	−4.0	−2.0
ES	−53.1	−100.0	−100.0	−86.5	−37.0	−22.4
FI	−2.8	1.7	0.0	−1.9	−2.4	−3.6
FR	−15.3	−63.1	−27.3	−12.9	−9.1	−5.6
HR	−33.8	−48.3	−63.2	−64.1	−25.1	−20.0
HU	−4.9	0.0	−7.3	−4.8	−5.0	−9.1
IE	13.6	3.1	21.8	25.0	9.6	10.0
IT	−3.1	17.9	0.0	−5.7	0.0	−4.7
LT	94.9	50.0	100.0	166.7	66.7	75.0
LU	2.5	1.2	12.0	8.4	3.4	1.2
LV	−53.5	−95.0	−93.8	−82.2	−65.2	−32.0
MT	−32.0	−59.5	−45.9	−33.7	−31.5	−63.3
PL	−26.5	−63.3	−59.2	−37.5	−16.7	−20.0
PT	10.9	20.0	22.1	21.2	20.0	4.7
RO	3.3	−12.5	−20.0	16.7	0.0	6.2
SE	−8.1	−14.5	−12.8	−8.1	−7.7	−7.7
SI	−83.0	−96.0	−96.5	−96.3	−95.0	−72.7
SK	13.4	−11.7	−16.5	−10.6	14.6	34.5
UK	11.1	5.1	11.6	12.9	10.0	9.6

Table 5: Gap between HBS and EU-SILC for rents paid by tenants

3.1.2 Rents paid by tenants

There are few quantitative variables that can be used for matching the two surveys. On the one hand, quantitative variables have the advantage to show much more variability than categorical variables (and therefore much more explanatory power); on the other hand, they suffer in much cases from measurement errors, thereby jeopardizing the models built on them. Common quantitative variables in EU-SILC and HBS are very few and turn out to be quite unevenly distributed, which confirms the measurement error issue. Rents paid by tenants are one of these very few variables. They prove to suffer from sometimes discrepancies in terms of distribution, as shown in Table 5 and in Annex B.7. Therefore, the variable is introduced in the model only under the form of quantiles. More precisely, it is combined with the tenure status so as to obtain a categorical variable combining the tenure status and, for the tenants, the quartiles of rents.

Income components	Missing	Not at all	Yes, partially	Yes, totally
Employee income	4	1	1	15
Income from self-employment	4	1	2	14
Property income	4	1	4	12
Social security pensions/schemes	4	1	2	14
Pensions and other insurance benefits	4	1	2	14
Social assistance benefits	4	1	2	14
Current transfers received from non-profit institutions	4	6	2	9
Current transfers received from other households	5	3	4	9
Direct taxes	5	5	2	9
Current inter-household transfers paid	5	4	5	7
Employee/employer social insurance contributions	5	9	1	6
Current transfers to non-profit institutions	5	10	2	4
Other	13	4	2	2

Table 6: Number of countries collecting various income components, according the EU-Survey

3.1.3 Income

When it comes to describe consumption, income turns out to be a natural candidate for explaining consumption behaviours. This is mainly the reason why income is available as a variable in HBS. However, its definition greatly varies across countries, thereby preventing the user to properly perform cross-country analyses. Moreover, since HBS is already a highly demanding survey, income is sometimes collected as a side variable, relying on only a very limited number of questions.

A survey as conducted at spring 2016 among HBS representatives in order to assess the comparability of the income variable across countries. Table 6 sums up the countries that collect various components of income in the different countries, whether the countries collect income as a whole or component by component. This gives hence an idea of the concept of income underlying the variable existing in the HBS data; it turns out that, if some components are almost there (employee income, income coming from self-employment), some others are not collected by all countries – and may affect quite significantly the level of income, as for instance taxes. From this viewpoint, it is also interesting to conduct a concrete comparison of the different variables that we have at our disposal in EU-SILC (gross income, disposable income) and compare their distribution with the one obtained for the income variable in HBS data.

Such an analysis is performed for disposable income (variable `HY020` in EU-SILC) in table 7 and more extensively in Annex B.8. If for some countries, the data show a great consistency between EU-SILC and HBS data (as is the case for for instance BE, CY, DE, FI, FR, or HU), it is far from being the case for all the countries. We therefore need to use the income variables in percentiles and not as levels, thereby making again implicitly the "rank assumption" we already mention before. Nevertheless, for some countries we observe quite well disposable income in HBS; for those countries, one could derive and analyse the link between income and consumption directly from the HBS data. However,

Country	Mean	Q10	Q25	Median	Q75	Q90
AT	8.1	−13.4	−6.0	2.6	6.1	14.7
BE	−4.2	−6.9	−9.5	−4.3	−1.4	−3.5
BG	58.1	4.6	27.9	46.8	59.8	68.2
CY	0.4	2.0	−5.6	−2.1	−1.2	−0.5
CZ	8.3	−0.7	1.0	5.5	7.2	12.5
DE	−1.9	−8.2	−1.0	1.3	−0.7	−4.8
DK	−23.4	−22.1	−18.9	−22.6	−21.9	−21.1
EE	17.4	3.0	10.2	11.3	18.9	30.4
EL	−16.6	−22.0	−16.2	−14.9	−15.0	−17.8
ES	21.1	−4.4	−0.3	13.5	23.9	27.6
FI	−0.3	−2.0	0.4	0.1	−0.8	−1.0
FR	7.6	14.2	8.8	6.1	4.7	5.9
HR	−1.8	−9.5	−7.0	−3.6	0.0	1.9
HU	2.0	−1.1	0.6	1.4	2.2	3.9
IE	−12.5	4.9	−6.8	−8.9	−12.7	−15.8
LT	12.5	−3.2	−10.1	1.5	10.7	18.9
LU	9.4	−6.1	3.8	8.6	13.5	12.5
LV	19.0	4.3	6.6	13.4	25.7	30.4
MT	3.2	0.6	0.1	1.8	4.1	3.4
PL	14.2	8.3	9.4	14.6	17.9	16.4
PT	−4.9	1.7	2.5	0.9	−4.6	−8.9
RO	−6.2	−13.5	−12.5	−7.5	−5.7	−2.7
SE	−8.5	−9.3	−13.2	−6.7	−6.5	−5.7
SI	13.5	17.5	17.8	10.9	13.1	11.2
SK	8.3	−1.4	−7.0	3.6	9.9	13.9
UK	−5.7	5.5	0.8	−5.5	−7.2	−6.6

Table 7: Gap (in %) between HBS and EU-SILC for income

we take advantage of such a situation as a "natural experience", applying the matching procedure as if we could not observe disposable income as such and comparing the results we got from the procedure with the actual data. This test is not aimed at validating the ranking assumption, but it nevertheless gives an idea on the reliability of the matching exercise in the case the ranking assumption turns out to be valid.

Finally, income turns out to be a variable whose value is quite sensitive to its definition and to its source. Whether it comes from registers or from a questionnaire which may be designed in many ways, income will not have the same distribution. However, the rank assumption may be valid; in order to test this assumption, we use the different definitions in EU-SILC (gross income, variable `HY010`; disposable income, variable `HY020`) and compute Spearman's correlation coefficient. The results of such computations are provided in Table 8. Besides Spearman's ρ, we also display the usual coefficient of correlation (called Pearson's ρ[3]). The computation of these two coefficients takes into account the survey weights; a third one – Kendall's τ – relies on the unweighted sample of SILC. Provided that Spearman's ρ proves to be quite high, the "rank assumption" seems to be acceptable in our case.

[3] These weighted indicators have been computed thanks to the R package `wCorr` [19].

Country	Spearman's ρ	Pearson's ρ	Kendall's τ
AT	0.980	0.967	0.888
BE	0.972	0.978	0.865
BG	0.996	0.996	0.953
CY	0.991	0.991	0.936
CZ	0.988	0.991	0.919
DE	0.967	0.960	0.848
DK	0.991	0.956	0.918
EE	0.994	0.996	0.940
EL	0.985	0.976	0.915
ES	0.990	0.986	0.928
FI	0.993	0.988	0.930
FR	0.990	0.988	0.919
HR	0.990	0.981	0.931
HU	0.981	0.978	0.897
IE	0.987	0.969	0.927
IT	0.985	0.984	0.905
LT	0.994	0.991	0.936
LU	0.973	0.953	0.872
LV	0.986	0.983	0.915
MT	0.996	0.993	0.953
PL	0.991	0.993	0.925
PT	0.987	0.980	0.916
RO	0.987	0.985	0.911
SE	0.991	0.981	0.923
SI	0.984	0.975	0.896
SK	0.980	0.979	0.913
TR	0.986	0.989	0.911
UK	0.984	0.979	0.902

Table 8: Correlation between gross and disposable income

3.2 Selection of matching variables

The selection of the matching variables for the classical hot-deck is performed as described in section 2.1.2; the concrete selection of the variables is performed in R thanks to the package `leaps` [20] and the function `regsubsets`. The outcome of the selection procedure is detailed in Table 9; then the hot-deck procedure is applied on the stratified samples, thereby making it possible to allocate households in HBS sample to household in SILC sample sharing the same characteristics.

Countries	Matching variables
AT	Age of RP, type of household, tenure status, main source of income, income quintiles
BE	Activity status of RP, age of RP, type of household, tenure status, main source of income, income quintiles
BG	Population density level, type of household, tenure status, income quintiles
CY	Activity status of RP, age of RP, population density level, type of household, tenure status, main source of income, income quintiles
CZ	Type of household, income quintiles
DE	Activity status of RP, age of RP, type of household, tenure status, income quintiles
DK	Age of RP, type of household, tenure status, income quintiles
EE	Activity status of RP, age of RP, population density level, type of household, tenure status, income quintiles
EL	Activity status of RP, age of RP, population density level, type of household, tenure status, income quintiles
ES	Age of RP, type of household, tenure status, income quintiles
FI	Activity status of RP, age of RP, type of household, tenure status, income quintiles
FR	Activity status of RP, age of RP, type of household, tenure status, income quintiles
HR	Activity status of RP, age of RP, type of household, tenure status, main source of income, income quintiles
HU	Activity status of RP, age of RP, population density level, type of household, tenure status, income quintiles
IE	Population density level, type of household, tenure status, income quintiles
IT	Activity status of RP, age of RP, population density level, type of household, tenure status
LT	Population density level, type of household, tenure status, income quintiles
LU	Activity status of RP, age of RP, type of household, tenure status, income quintiles
LV	Activity status of RP, age of RP, population density level, type of household, tenure status, income quintiles
MT	Activity status of RP, age of RP, type of household, income quintiles
PL	Activity status of RP, age of RP, population density level, tenure status, income quintiles
PT	Age of RP, population density level, type of household, tenure status, income quintiles
RO	Activity status of RP, age of RP, type of household, tenure status, income quintiles
SE	Age of RP, type of household, income quintiles
SI	Age of RP, population density level, type of household, income quintiles
SK	Type of household, income quintiles
UK	Age of RP, type of household, tenure status, income quintiles

Table 9: Matching variables for the different countries

Note: The subset of variables has been selected thanks to a backward regression procedure.

For the second matching procedure, the selection of the matching variables is simpler; given the fact that we do not use those variables to stratify the sample, we face less demanding constraints in terms of parsimony. It is then possible to estimate a model including all variables that are comparable enough across the two surveys. We therefore estimate a model for each country with the following covariates: age of the reference person, activity status of the reference person, population density level, type of household, tenure status, main source of income and income quintiles. Table 10 sums up the

Country	R^2	Pearson's ρ	Spearman's ρ
AT	0.839	0.617	0.663
BE	0.856	0.657	0.723
BG	0.914	0.818	0.843
CY	0.876	0.766	0.818
CZ	0.940	0.837	0.863
DE	0.901	0.771	0.834
DK	0.899	0.726	0.778
EE	0.804	0.689	0.705
EL	0.878	0.793	0.809
ES	0.854	0.683	0.726
FI	0.871	0.743	0.796
FR	0.843	0.678	0.721
HR	0.898	0.773	0.824
HU	0.910	0.744	0.778
IE	0.892	0.757	0.778
IT	0.742	0.436	0.512
LT	0.828	0.660	0.724
LU	0.848	0.609	0.647
LV	0.830	0.681	0.731
MT	0.813	0.592	0.657
PL	0.864	0.715	0.768
PT	0.795	0.694	0.711
RO	0.928	0.826	0.854
SE	0.888	0.688	0.731
SI	0.880	0.714	0.771
SK	0.914	0.771	0.822
UK	0.843	0.687	0.737

Table 10: Correlation between observed and predicted consumption in HBS

results of the different regressions and tests for the rank correlation between observed and predicted consumption. This makes it possible to gauge how the model reproduces the ranking, assumption on which the method relies. Statistics show that the ranking is reproduced in a satisfactory way; however, according to the R^2 and Pearson's ρ, the model provides a good prediction. From this viewpoint, the first method constitutes a good approach, as it also implicitly relies on such a model. This remark sketches the conclusions drawn in the next section regarding the ability of the different approaches to reproduce expected patterns.

Appendices

A Defining a reference person in SILC and HBS

Determining a unique reference person in survey data starts with the description of the household composition and the computation of the matrix describing the link between the different household members. From this viewpoint, we face strong limitations as in the harmonized data the full relationship matrix is not available. We therefore need to rest the work on some basic assumptions in order to compute in a consistent way the reference person in both HBS and EU-SILC, following the definition by the Canberra group [21]. The several steps we need to achieve are the following ones:

- computation of the number of members, children (if any) and dependent children
- computation of personal income
- determination of parents (when children are present in the household)
- determination of couples (if any)
- application of the Canberra rules depending of the household structure

A.1 Computation of the number of members, children and dependent children

A.1.1 In EU-SILC

In EU-SILC, the number of members is derived from table R and the children are defined the members whose age at the moment of the interview was less than 16. Dependent children are those members less than 15 or that are between 16 and 25, whose father or mother is in the household and who is out of the labour market.

```
proc sql ;
select max(rb030) into: max_n from out.silc_r_&year ;
quit ;

%macro determine_order(number,k=k) ;

%global &k ;
%let &k = 0 ;
%let ratio = %sysevalf(&number/(10**&&&k)) ;
%do %while (&ratio > 1) ;
  %let &k = %eval(&&&k + 1) ;
  %let ratio = %sysevalf(&number/(10**&&k)) ;
%end ;
%mend ;

%determine_order(number=&max_n,k=order) ;

proc sql ;
```

```
create table count as
select distinct rb020 as hb020, substr(put(rb030,z&order..),1,%eval(&order
    -2)) as hb030_c, count(*) as npers,
  sum(rb080 + 16 ge rb010) as number_children from in.silc_r_&year group by
     hb030_c ;
quit ;

proc sort data=in.silc_r_&year out=r ;
by rb010 rb020 rb30 ;
proc sort data=in.silc_p_&year out=p ;
by pb010 pb020 pb030 ;
run ;

data p ;
set p ;
rb010 = pb010 ;
rb020 = pb020 ;
rb030 = pb030 ;
run ;

data r ;
merge r p ;
by rb010 rb020 rb030 ;
run ;

proc sql ;
create table count2 as
distinct rb020 as hb020, substr(put(rb030,z&order..),1,%eval(&order-2)) as
    hb030_c,
  sum((rb080 + 15 ge rb010) or ( (rb080 + 15 < rb010 < rb080 + 25) and (
    rb220 is not missing or rb230 is not missing) and (pl031 not in
    (1,2,3,4,7)))) as number_dep_children
  from r group by hb030_c ;
quit ;
```

A.1.2 In HBS

On HBS data, we compte the needed variables on the `HM` table, using counterpart variables in order to obtain similar results.

```
proc sql ;
create table count as
select distinct ma04 as ha04, count(*) as npers, sum(mb03_n<16) as number_
    children,
sum((mb03_n<16) or ((15 < mb03_n < 25) and (mb05 = "3") and me01 not in
    ("1","3"))) as number_dep_children from hm group by ha04 ;
quit ;
```

A.2 Computation of personal income

A.2.1 In EU-SILC

Computing the personal income for each member who is 16+ consists of summing up all the income components that are collected at the individual level. This computation is performed on the table P.

```
data p ;
set out.silc_p_&year ;
income = sum(py010g, py021g, py050g, py080g, py090g, py100g, py110g, py120g
    , py130g, py140g) ;
if missing(income) then income = 0 ;
run ;
```

A.2.2 In HBS

On HBS data, the approach is simpler, as we use the variable EUR_MF099 which directly provides personal income.

A.3 Determination of parents

A.3.1 In EU-SILC

In EU-SILC, we use the variables RB220 (which provides the id of the father if member of the household) and RB230 (which gives the id of the mother if member of the household).

```
/* determine the parents in the hh */

data father ;
set out.silc_r_&year ;
where not missing(rb220) ;
keep rb020 rb220 ;
run ;

data father ;
set father ;
rename rb220 = pb030
     rb020 = pb020 ;
run ;

data mother ;
set out.silc_r_&year ;
where not missing(rb230) ;
keep rb020 rb230 ;
run ;

data mother ;
set mother ;
rename rb230 = pb030
```

```
     rb020 = pb020 ;
run ;

proc sort data=father nodupkey ;
by pb020 pb030 ;
proc sort data=mother nodupkey ;
by pb020 pb030 ;
proc sort data=p ;
by pb020 pb030 ;
run ;

data p ;
merge p (in=a) father (in=b) mother (in=c) ;
by pb020 pb030 ;
if a ;
parent = (b or c) ;
run ;
```

A.3.2 In HBS

In HBS, the information is far less rich than it is actually in EU-SILC. Indeed, for this computation we can only rest on the variable `MB05` which provides the link between the member and the reference person (as it is defined in the survey, which may not correspond to the Canberra definition). Then the entire relationship matrix can be approximated only through this *ad hoc* person of reference and his links with the other members of the household.

A.4 Determination of individuals in couple

A.4.1 In EU-SILC

In EU-SILC, we consider only people listed in table `P` – *i.e.* all individuals aged 16 or more – and we use the variable `PB180` in order to determine whether the person is in couple with another member of the household or not.

```
/* determine the couples in the hh */

data couple ;
set out.silc_p_&year ;
where not missing(pb180) ;
keep pb020 pb030 ;
run ;

proc sort data=couple ;
by pb020 pb030 ;
run ;

data p ;
merge p (in=a) couple (in=b) ;
```

```
by pb020 pb030 ;
if a ;
couple = b ;
run ;
```

A.4.2 In HBS

In HBS, we rest on a single variable `MB04` that indicates whether the individual is married or in a registered partnership. We implicitly assume that individuals that are in such a case live in the same household.

```
data hm ;
set out.hbs_hhm_&country._2010 ;
couple = (mb04="2" or mb042="1") ;
run ;
```

A.5 Application of the Canberra rules depending of the household structure

We now can determine the reference person, using the selection rules defined by the Canberra group, as we know which members is in a couple, which have children, who from the selected couple earn the more and who is the most aged.

A.5.1 In EU-SILC

```
data p ;
set p ;
attrib type length=$1. ;
/* types of situation */
/* type a - couple with dependent children */
if couple = 1 and parent = 1 then type = "a" ;
/* type b - couple without dependent children */
else if couple = 1 and parent = 0 then type = "b" ;
/* type c - lone parent */
else if couple = 0 and parent = 1 then type = "c" ;
/* type d - neither parent not couple */
else type = "d" ;
run ;

proc sql ;
create table p as
select *, substr(put(pb030,z&order..),1,%eval(&order-2)) as hb030_c, sum(
    type="a") as count_a,
        sum(type="b") as count_b, sum(type="c") as count_c, sum(type="d")
    as count_d
        from p group by hb030_c ;
```

```
quit ;

/* select persons eligible to be a reference person, according to Canberra
    definition */

data refp ;
retain hb030 ;
set p ;
if count_a > 0 and type ne "a" then delete ;
if count_a = 0 and count_b > 0 and type ne "b" then delete ;
if count_a = 0 and count_b = 0 and count_c > 0 and type ne "c" then delete
    ;
hb030 = hb030_c+0 ;
drop hb030_c ;
run ;

proc sql ;
create table refp as
select *, max(income) as max_income, max(pb010-pb140) as max_age, min(pb130
    ) as min_month_birth from refp group by pb020, hb030 ;
quit ;

data refp ;
set refp ;
where income = max_income ;
run ;

proc sort data=refp ;
by pb020 hb030 ;
run ;

data refp ;
set refp ;
by pb020 hb030 ;
duplicate = (first.hb030 ne last.hb030) ;
run ;

data refp ;
set refp ;
where duplicate = 0 or (duplicate = 1 and pb010 - pb140 = max_age) ;
run ;

proc sort data=refp ;
by pb020 hb030 ;
run ;

data refp ;
set refp ;
by pb020 hb030 ;
duplicate = (first.hb030 ne last.hb030) ;
run ;

data refp ;
```

```
set refp ;
where duplicate = 0 or (duplicate = 1 and pb130 = min_month_birth) ;
run ;
```

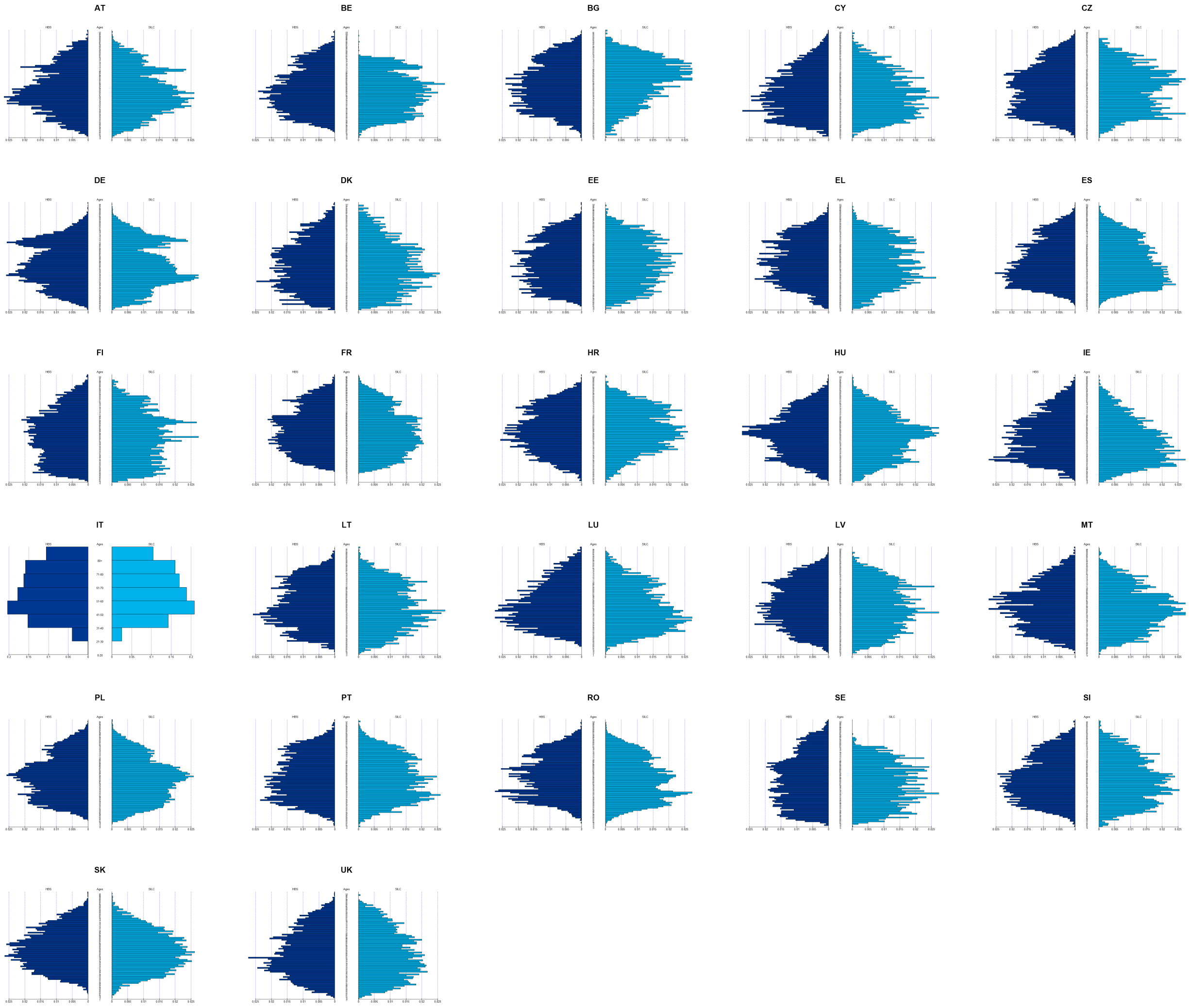
A.5.2 In HBS

```
proc sql ;
create table hm as
select *, (sum(mb05 = "3")>0) as exist_child_r, (sum(mb03_n<16 and mb05 ne
    "3")>0) as exist_child_o, (sum(mb05 = "4")>0) as exist_parent_r
from hm group by ma04 ;
quit ;

data hm ;
set hm ;
attrib type length=$1. ;
/* types of situation */
/* type a - couple with children */
if (couple=1 and ((mb05 in ("1","2") and exist_child_r) or (mb05 not in
    ("1","2","4") and exist_child_o) or (mb05="4"))) then type="a" ;
/* type b - couple without children */
else if couple=1 then type="b" ;
/* type c - lone parent */
else if (couple=0 and ((mb05="1" and exist_child_r) or (mb05="4"))) then
    type="c" ;
/* type d - neither parent nor couple */
else type="d" ;
run ;

proc sql ;
create table hm as
select *, sum(type="a") as count_a, sum(type="b") as count_b, sum(type="c")
     as count_c, sum(type="d") as count_d
from hm group by ma04 ;
quit ;

/* select persons eligible to be a reference person, according to Canberra
    definition */

data refp ;
set hm ;
ha04 = ma04 ;
if count_a > 0 and type ne "a" then delete ;
if count_a = 0 and count_b > 0 and type ne "b" then delete ;
if count_a = 0 and count_b = 0 and count_c > 0 and type ne "c" then delete
    ;
run ;

proc sql ;
create table refp as
select *, max(eur_mf099) as max_income, max(mb03_n) as max_age from refp
    group by ha04 ;
```

```
quit ;

data refp ;
set refp ;
where eur_mf099 = max_income ;
run ;

proc sort data=refp ;
by ha04 ;
run ;

data refp ;
set refp ;
by ha04 ;
duplicate = (first.ha04 ne last.ha04) ;
run ;

data refp ;
set refp ;
where duplicate = 0 or (duplicate = 1 and mb03_n = max_age) ;
run ;

proc sort data=refp nodupkey ;
by ha04 ;
run ;

data refp ;
set refp ;
age_rp = mb03_n ;
level_edu_rp = mc01 ;
status_activity_rp = me01 ;
occupation_status = me0988 ;
keep ha04 age_rp level_edu_rp status_activity_rp occupation_status ;
run ;
```

B Comparison of common variables between EU-SILC and HBS

B.1 Age of the reference person



AT BE BG CY CZ

DE DK EE EL ES

FI FR HR HU IE

IT LT LU LV MT

PL PT RO SE SI

SK UK

Figure 1: Comparison (pyramid-wise) of the structure of ages between HBS and EU-SILC

B.2 Population density level

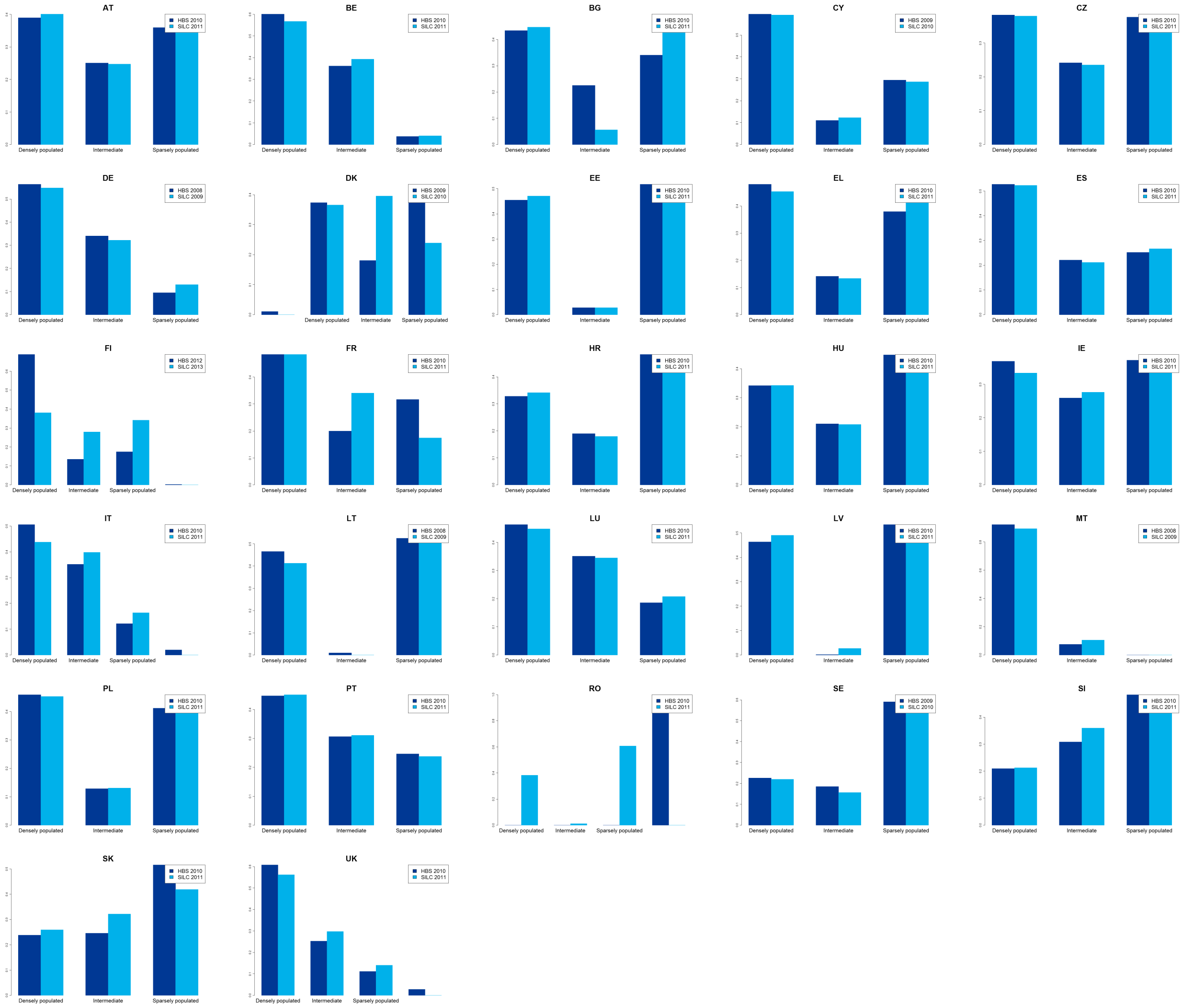


Figure 2: Comparison of the density level between HBS and EU-SILC

B.3 Household size

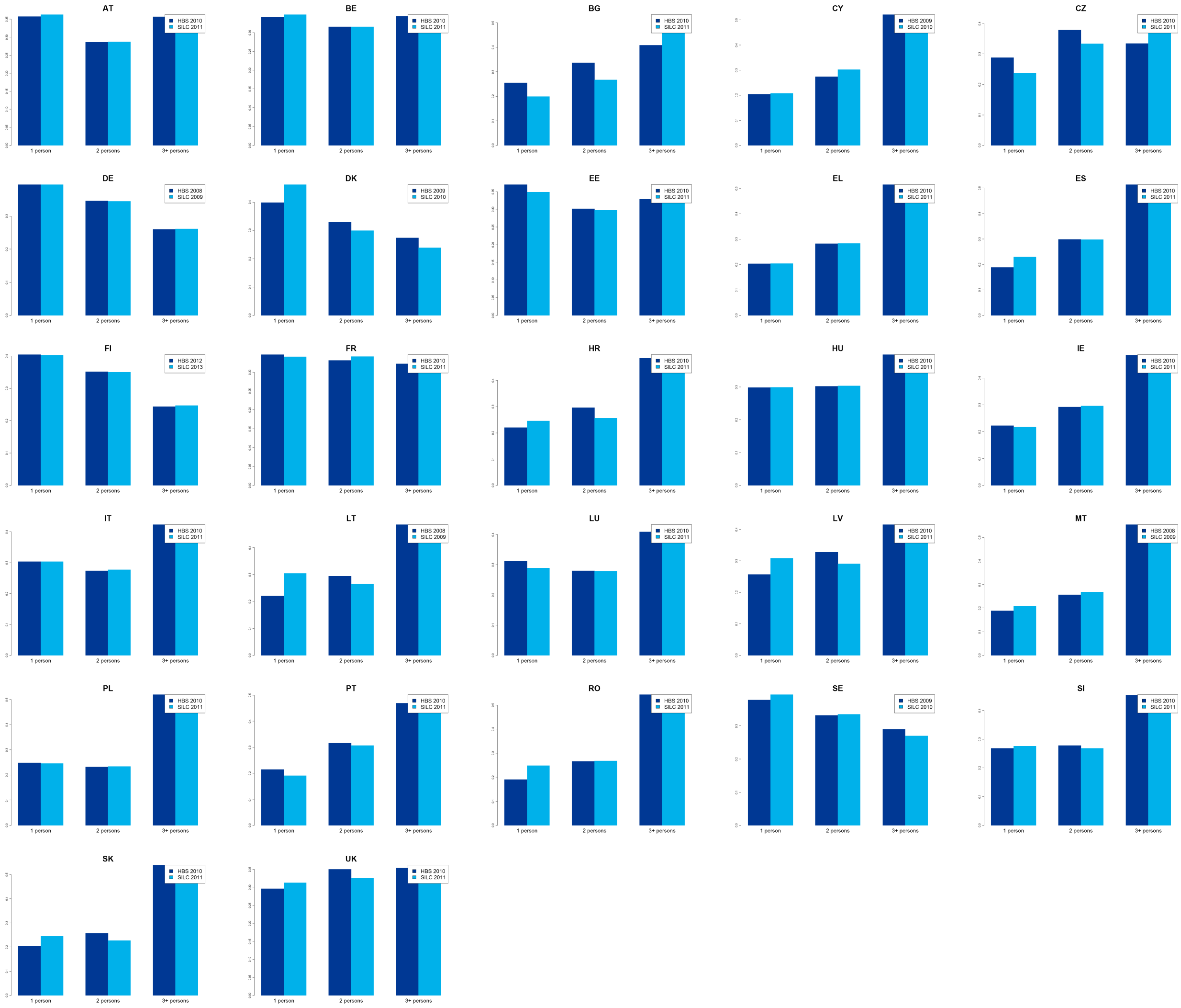


Figure 3: Comparison of the household size between HBS and EU-SILC

B.4 Household type

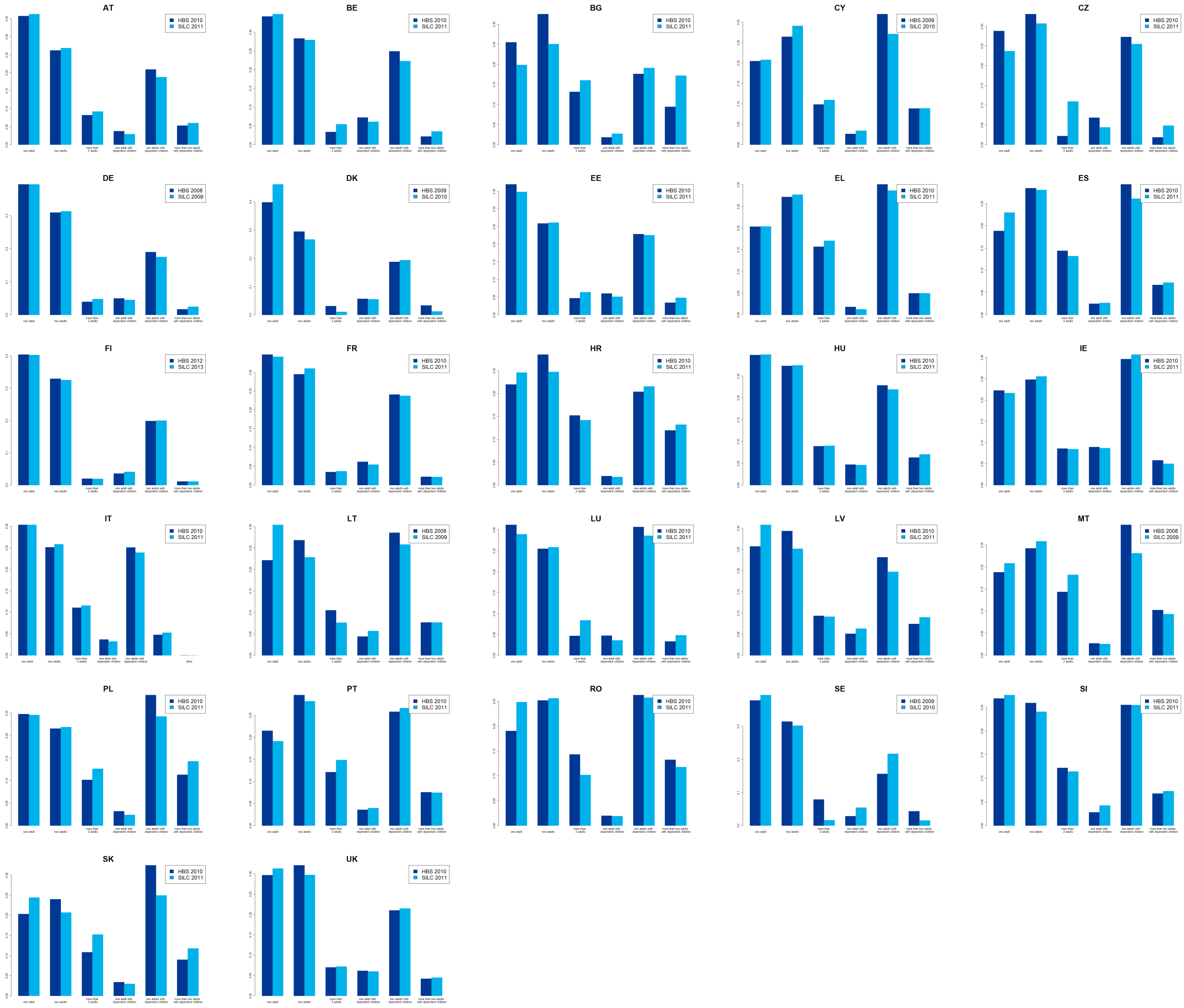


Figure 4: Comparison of the household type between HBS and EU-SILC

B.5 Activity status of the reference person

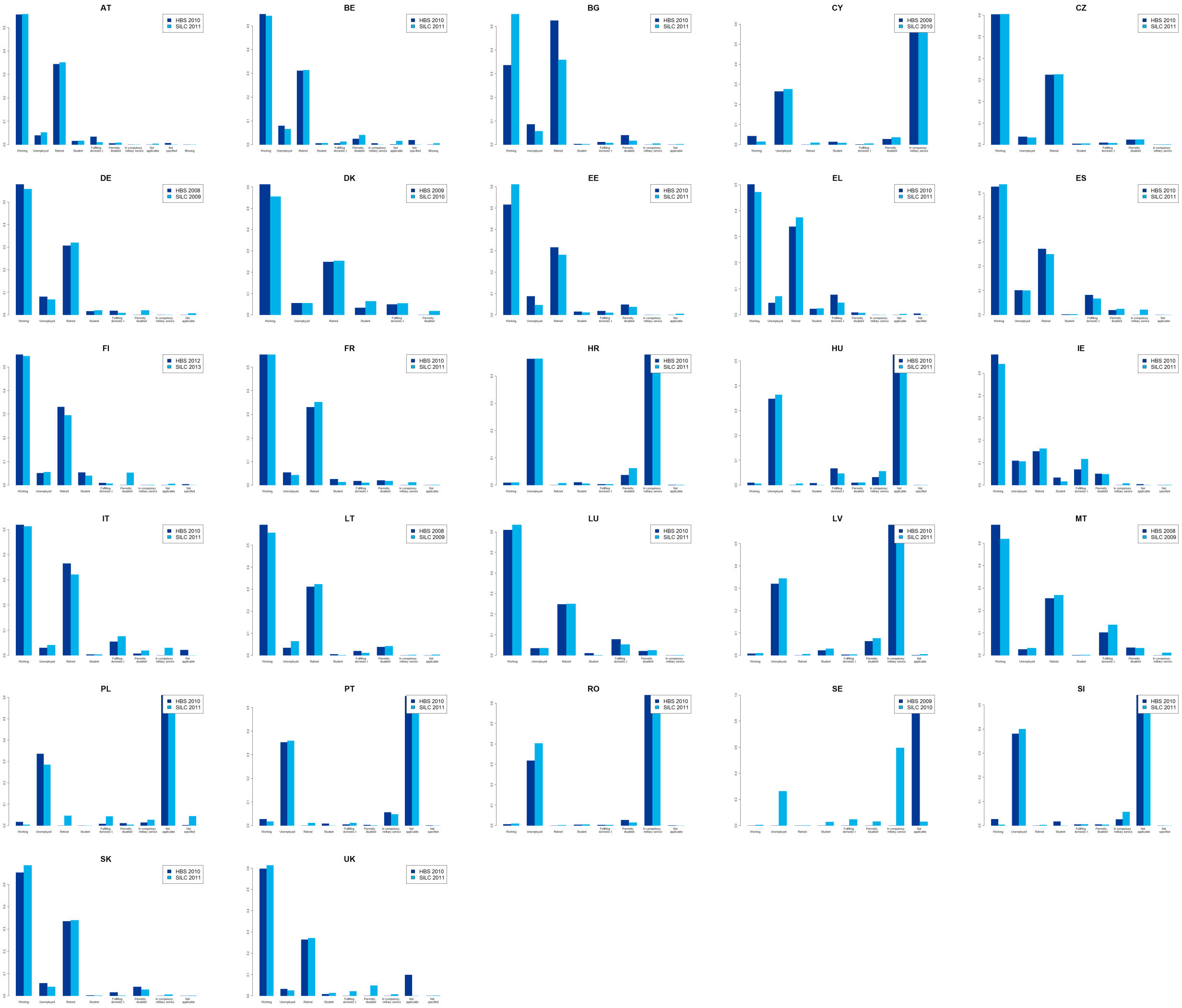


Figure 5: Comparison of the activity status of the reference person between HBS and EU-SILC

B.6 Level of education of the reference person

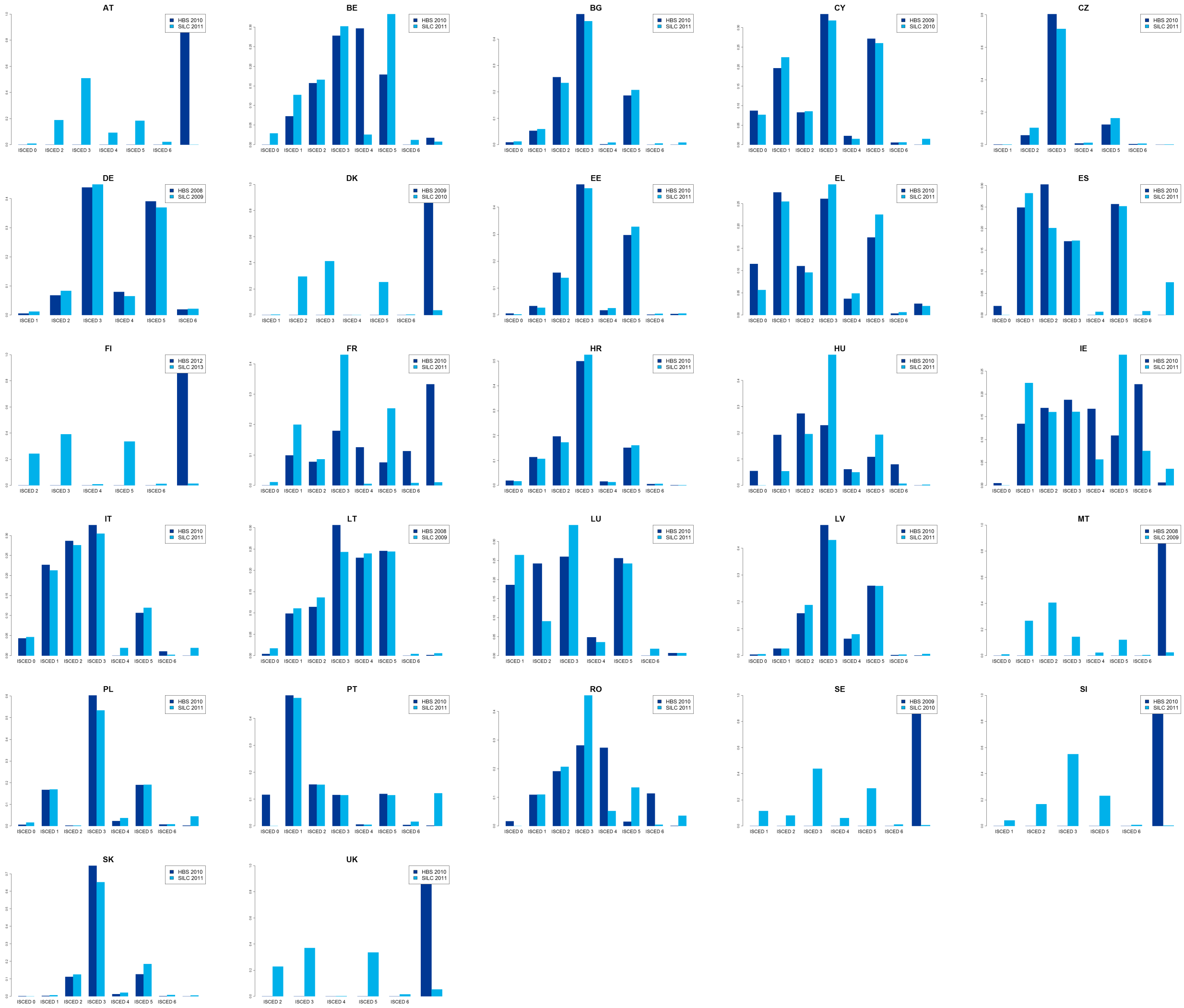


Figure 6: Comparison of the level of education of the reference person between HBS and EU-SILC

B.7 Rents paid by tenants

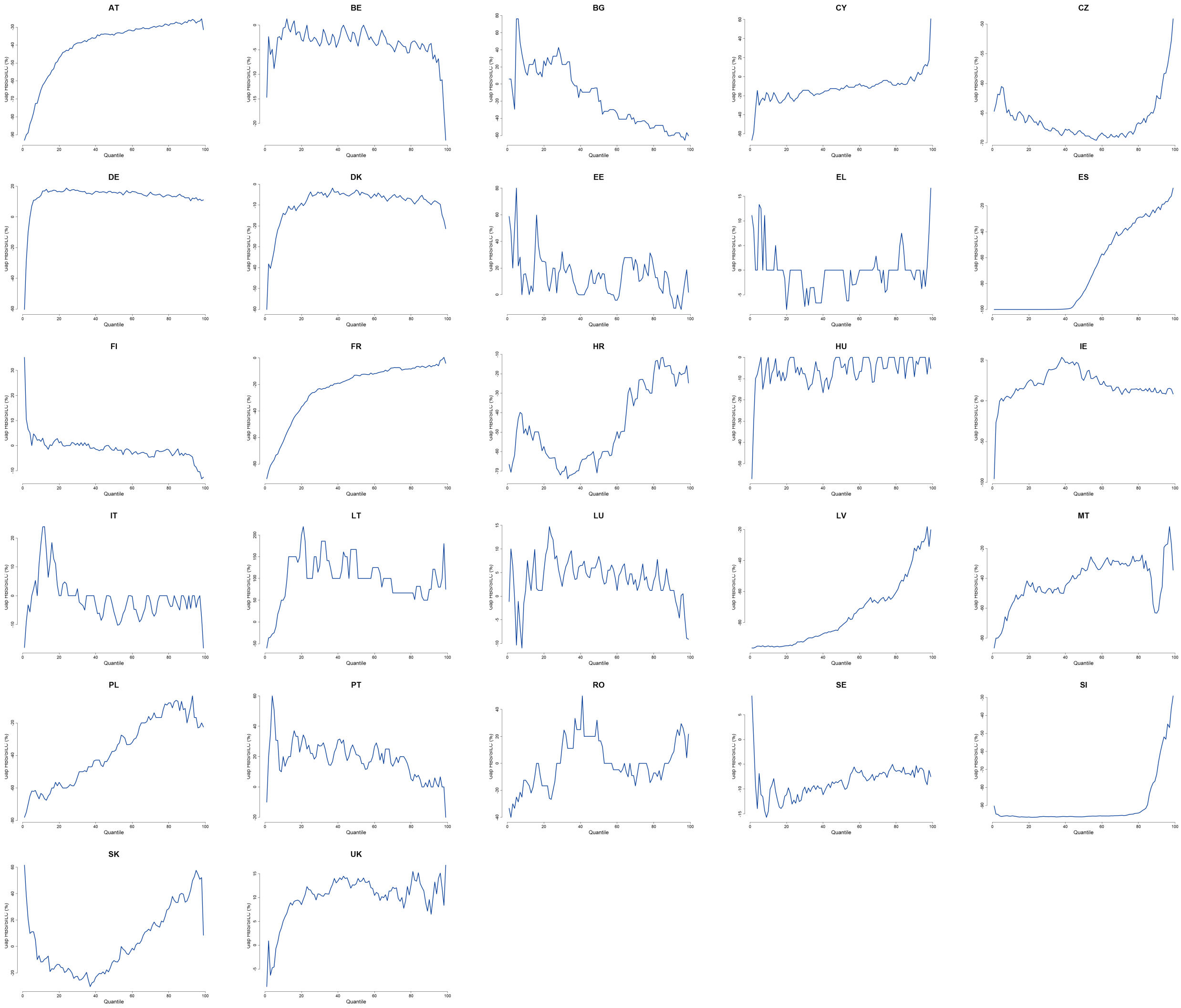


Figure 7: Q-Q plot for the rents paid by tenants in HBS as compared to EU-SILC

B.8 Income

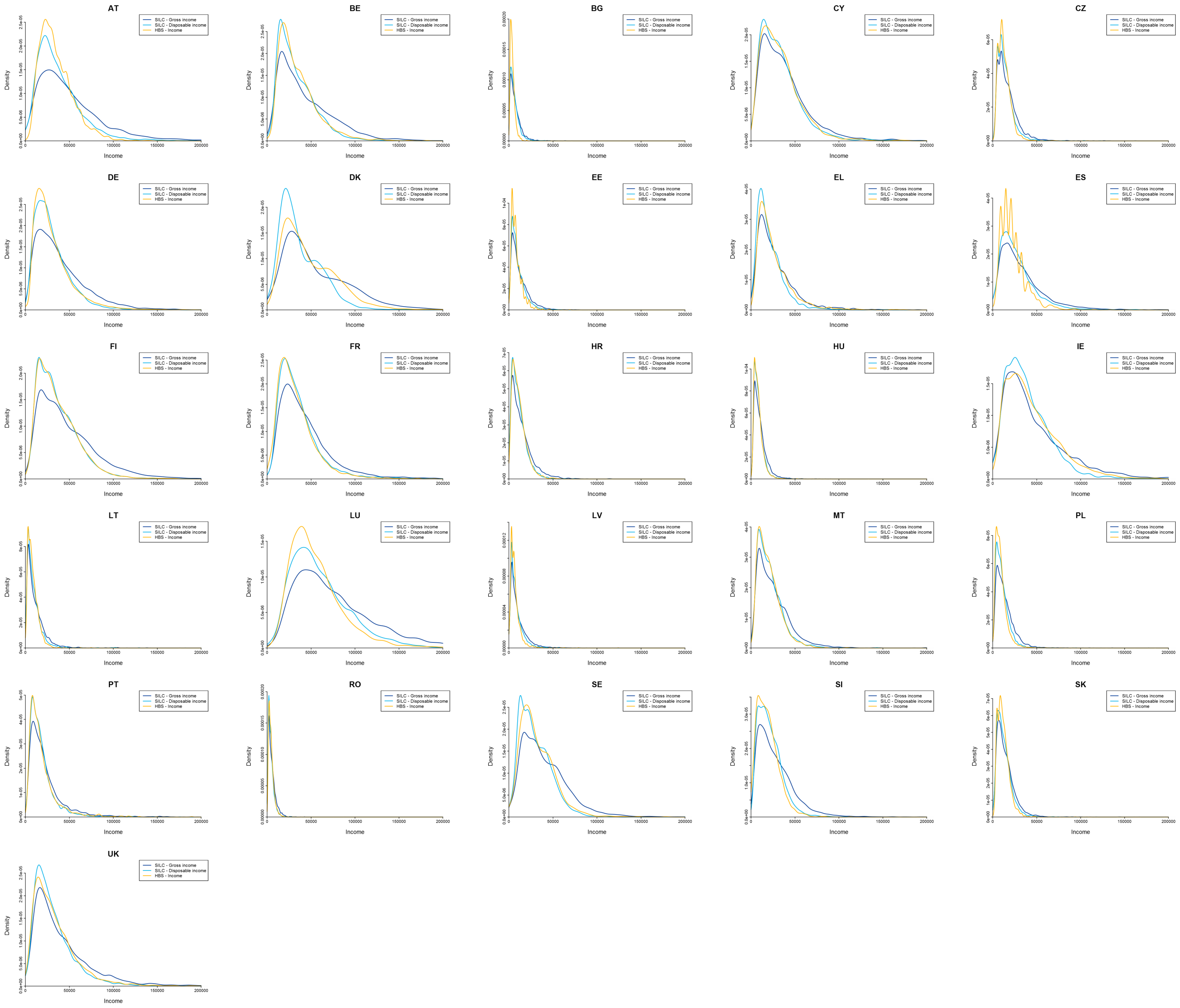


Figure 8: Densities of income in EU-SILC (gross and disposable) and in HBS

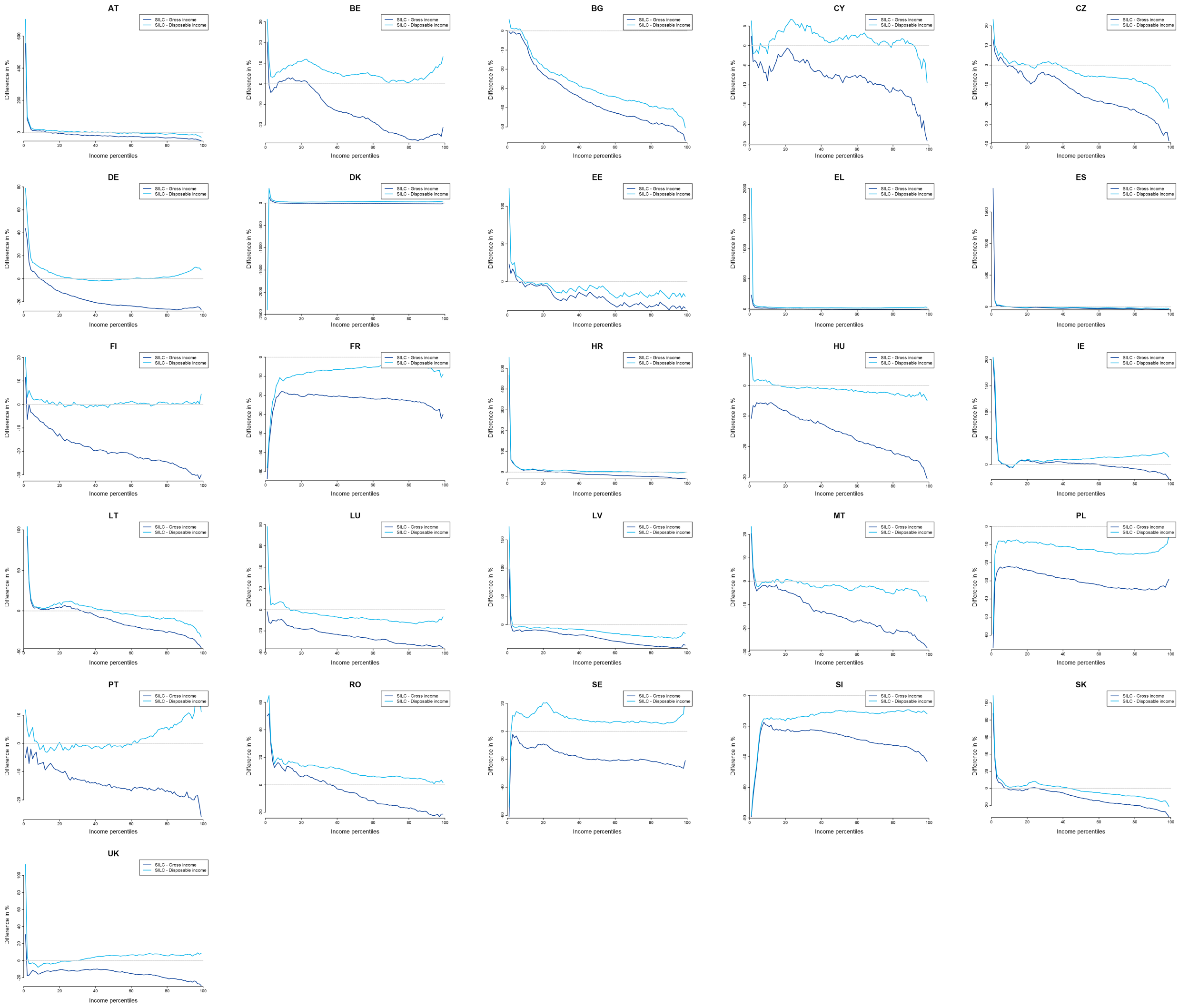


Figure 9: Q-Q plots for income in EU-SILC (gross and disposable) as compared to HBS

B.9 Gross vs. disposable income

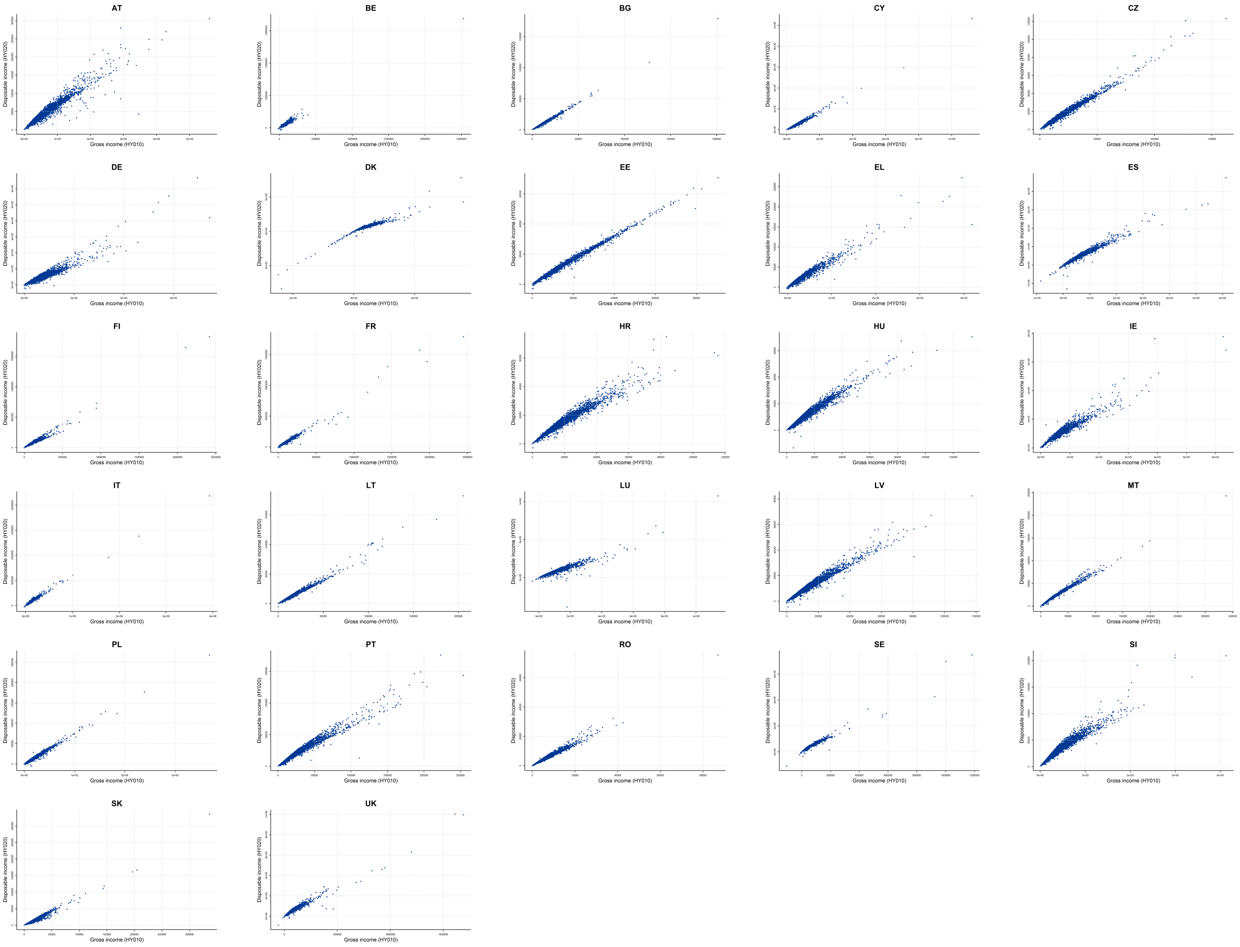


Figure 10: Scatter plot of gross and disposable income in EU-SILC

In this section, we display two different types of graphs; the first group (see Figure 10) shows the values simultaneously taken by gross income (variable `HY010`) and disposable income (variable `HY020`) in the EU-SILC sample. The representation does not account for weights and reflects basically the (unweighted) correlation between the two variables (as scatter plots usually do). The second group (see Figure 11) rather focuses on the link between the ranking provided by the two variables and accounts for weights in the survey. For each variable, the target population is normalized to 1; we denote X_i and ω_i respectively the value taken for household i by the variable of interest (either gross or disposable income) and the weight allocated to this household. $X_{(i)}$ and $\omega_{(i)}$ denote statistics sorted according to X, *i.e.* $X_{(1)} < ... < X_{(i)} < ... < X_{(n)}$. Along with these notations, we compute the rank r_i^X associated to the variable X as follows:

$$r_i^X = \sum_{k=1}^{i} \omega_{(k)} \tag{11}$$

We compute this weighted rank for both gross and disposable income. The Figure 11 represents a scatter plot of this rank for the sample in EU-SILC. In other words, Figure 10 is a graphical representation of Pearson's ρ, while Figure 11 is a representation of weighted Spearman's ρ. In an economic perspective, since we compare for each household income before and after taxes, these figures also display the effect of the fiscal system both on the level and on the rank according to income.

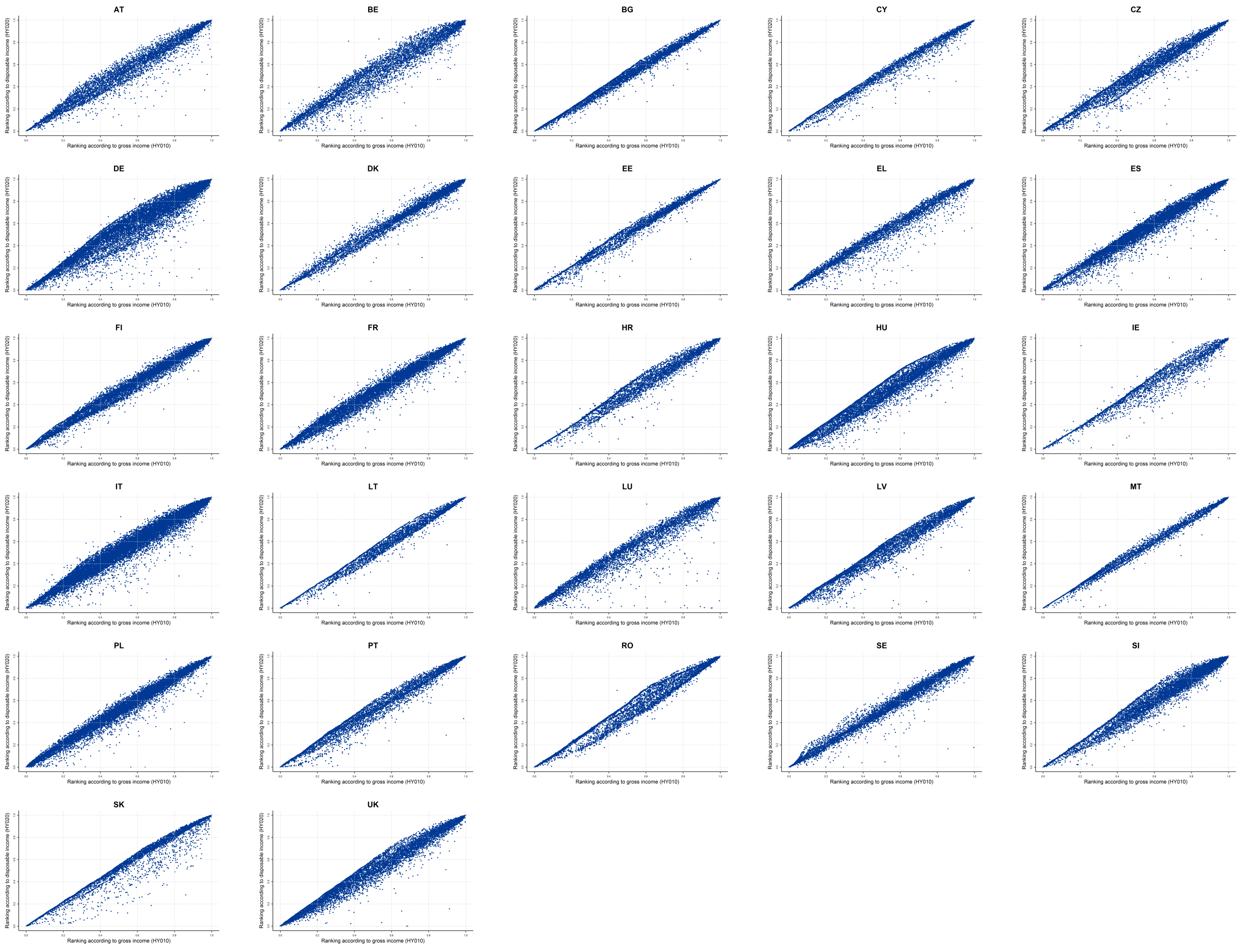


Figure 11: Scatter plot of the weighted ranking according to gross and disposable income in EU-SILC

References

[1] D'Orazio M, Di Zio M, Scanu M. Statistical matching: Theory and practice. John Wiley & Sons; 2006.

[2] Leulescu A, Agafitei M. Statistical matching: a model based approach for data integration. Eurostat-Methodologies and Working papers. 2013;Available from: http://ec.europa.eu/eurostat/documents/3888793/5855821/KS-RA-13-020-EN.PDF/477dd541-92ee-4259-95d4-1c42fcf2ef34?version=1.0.

[3] Baldini M, Pacifico D, Termini F, et al. Imputation of missing expenditure information in standard household income surveys. Universita di Modena e Reggio Emilia, Dipartimento di Economia" Marco Biagi"; 2015. Available from: http://155.185.68.2/campusone/web_dep/CappPaper/Capp_p116.pdf.

[4] Renssen RH. Use of statistical matching techniques in calibration estimation. Survey Methodology. 1998;24:171–184.

[5] Rubin DB. Multiple imputation for nonresponse in surveys. vol. 81. John Wiley & Sons; 2004.

[6] D'Orazio M. StatMatch: Statistical Matching; 2017. R package version 1.2.5. Available from: https://CRAN.R-project.org/package=StatMatch.

[7] D'Orazio M, Di Zio M, Scanu M. Statistical matching for categorical data: Displaying uncertainty and using logical constraints. Journal of Official Statistics. 2006;22(1):137. Available from: https://www.istat.it/it/files/2014/04/jos-2006-221.pdf.

[8] Conti PL, Marella D, Neri A. Statistical matching and uncertainty analysis in combining household income and expenditure data. Statistical Methods & Applications. 2015;p. 1–21. Available from: http://EconPapers.repec.org/RePEc:bdi:wptemi:td_1018_15.

[9] Rässler S. Data fusion: identification problems, validity, and multiple imputation. Austrian Journal of Statistics. 2004;33(1&2):153–171. Available from: http://www.ajs.or.at/index.php/ajs/article/download/vol33%2C%20no1%262%20-%209/380.

[10] Bhattacharya N, Mahalanobis B. Regional disparities in household consumption in India. Journal of the American Statistical Association. 1967;62(317):143–161. Available from: http://library.isical.ac.in:8080/jspui/bitstream/10263/1397/1/JOTASA-62-317-1967-P143-161.pdf.

[11] Pyatt G. On the interpretation and disaggregation of Gini coefficients. The Economic Journal. 1976;86(342):243–255. Available from: http://www.jstor.org/stable/2230745.

[12] Lambert PJ, Aronson JR. Inequality decomposition analysis and the Gini coefficient revisited. The Economic Journal. 1993;103(420):1221–1227. Available from: http://www.jstor.org/stable/2234247.

[13] Eurosystem Household Finance and Consumption Network. The Eurosystem Household Finance and Consumption Survey-Results from the First Wave. European Central Bank; 2013. Available from: https://www.ecb.europa.eu/pub/pdf/other/ecbsp2en.pdf?2180f869d12ccc366869c9419b3da32e.

[14] Eurosystem Household Finance and Consumption Network. The Eurosystem Household Finance and Consumption Survey-Methodological Report for the First Wave. European Central Bank; 2013. Available from: https://www.ecb.europa.eu/pub/pdf/other/ecbsp1en.pdf?c5295916d8521d593c30abc97ef9fc58.

[15] Browning M, Crossley TF, Weber G. Asking consumption questions in general purpose surveys. The Economic Journal. 2003;113(491):F540–F567. Available from: http://socserv.mcmaster.ca/sedap/p/sedap77.pdf.

[16] Lamarche P. Can your stomach predict your total consumption? IFC Bulletins chapters. 2015;39. Available from: http://www.bis.org/ifc/events/7ifcconf_lamarche.pdf.

[17] Lamarche P. Estimating consumption in the HFCS: Experimental results on the first wave of the HFCS. ECB Statistics Paper Series. 2017 May;Available from: https://www.ecb.europa.eu/pub/pdf/scpsps/ecb.sps22.en.pdf?e8a63d8194ca58f6d3dcee9b37a7b61b.

[18] UNECE. Canberra Group Handbook on Household Income Statistics. on Household Income Statistics CG, editor. United Nations; 2011. Available from: https://www.unece.org/fileadmin/DAM/stats/.../Canbera_Handbook_2011_WEB.pdf.

[19] Emad A, Bailey P. wCorr: Weighted Correlations; 2016. R package version 1.8.0. Available from: https://CRAN.R-project.org/package=wCorr.

[20] Lumley T. leaps: Regression Subset Selection; 2017. R package version 3.0. Available from: https://CRAN.R-project.org/package=leaps.

[21] UNECE. Expert group on household income statistics: final report and recommendations. Canberra Group; 2001. Available from: http://www.lisdatacenter.org/wp-content/uploads/canberra_report.pdf.