2014 INTERNATIONAL WORKSHOP AND CONFERENCE ON COMPARATIVE EU STATISTICS ON INCOME AND LIVING CONDITIONS

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

Gabriella Donatiello, Doriana Frattarola, Antony Rizzi, Mattia Spaziani

Lisbon, 15-17 October 2014

Istat

Overview

1. Introduction
2. Harmonization of common variables
3. Household income in HBS
4. The matching procedure
5. Predicting consumption with HBS data
6. Asking consumption in EU-SILC
7. Some concluding remarks

Statistical matching provide joint information on variables not collected through a single survey, as income (IT-SILC) and consumption (HBS).

As well known, **matching income and consumption is not a simple task**:

- Different modes of collection data
- Different definitions of variables
- Complex sample surveys (involving two stages of selection of the sample units)
- Impossibility to assume CIA

Overstep this critical aspects require strong prerequisites of coherence of data sources

- **Harmonization of sources and common variables**
- **Reconciliation and harmonization efforts beyond core social variables**
- **Introducing new shared variables**

HBS and IT-SILC surveys cover the same population and are based on a two-stage sampling design.

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

	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

It is necessary to choose a set of common variables that have to be **comparable**.

✓the analyses are done at **household level**
✓some variables are aggregated from the **individual level** (each one refers to the reference person and in both survey it is the holder of the registry form)

Matching variables have to satisfy two criteria:

i. there must be **homogeneity** in **distribution** across the two surveys (average HD 3,62%);
ii. they must be **good predictors of** both **income and consumption**.

HH070

IT-SILC **Total Housing Cost**

- costs of utilities (water, electricity, gas and heating)
- expenses connected with the household right to live in the accommodation (mortgage interest payments or and rent payments)

HBS Most of the components included in HH070 are collected (except the expenses for municipal solid waste and sewer services).

A new variable is created in each survey by adding rent payments for tenants and subjective rent for non-tenants.

- There are significant differences in the ways each survey collects these information
- The analysis (T Test and Kolmogorov-Smirnov Test) on the reconstructed variable confirms a good degree of comparability among HBS and IT-SILC.

Identify consumption components that are good predictors for total consumption in HBS to enhance reliability of the matching estimates similarly to the use of auxiliary information on income in HBS

Two steps:

i. **check the structure of total consumption** and compare the shares that different items have across the classes of income
ii. investigate the **explanatory power** of each amount **with a statistical model**

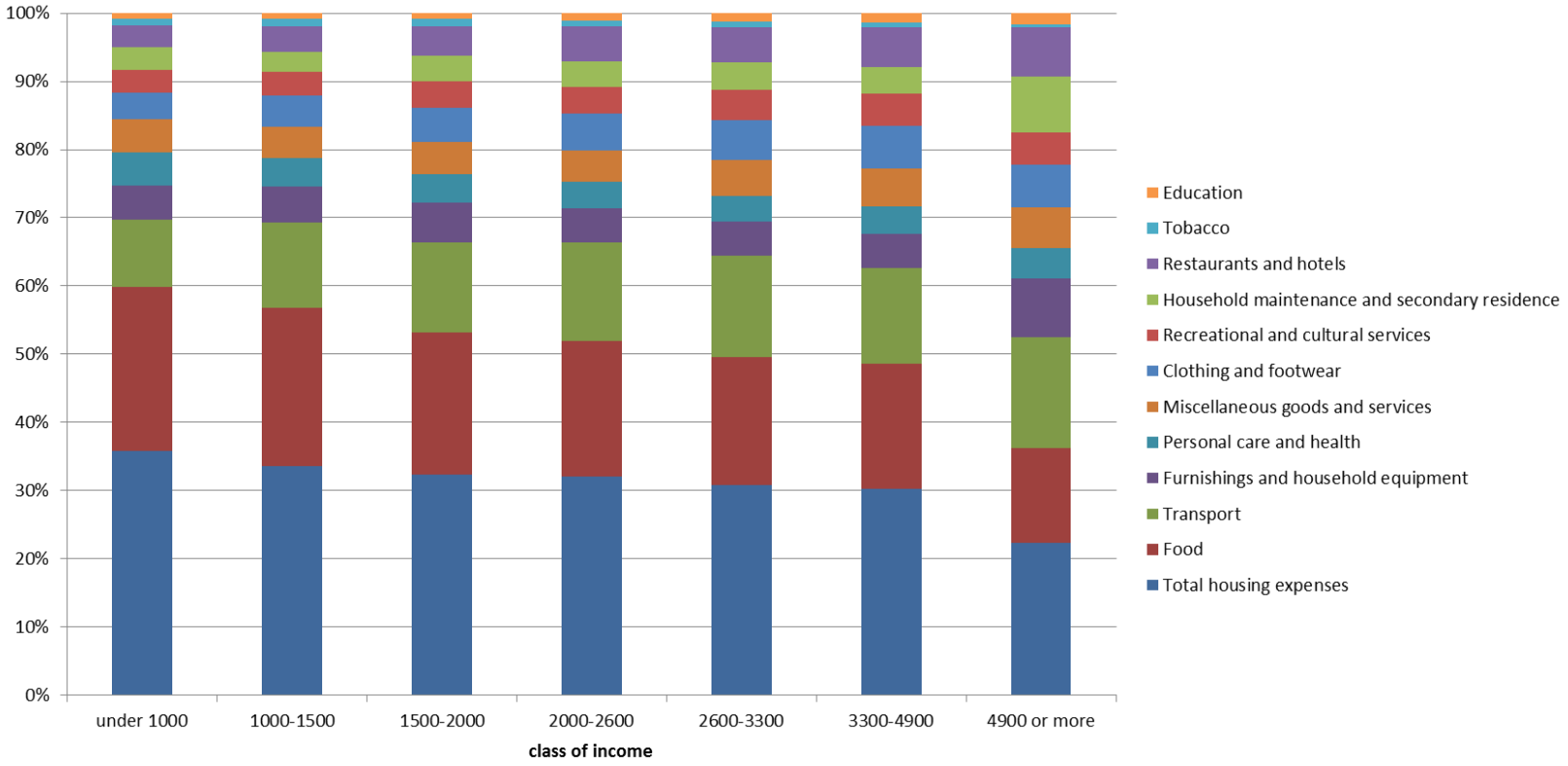
Method: stepwise regression
Dependent variable:

- logarithmic transformation of monthly total consumption

*Covariates**:***

- Socio-demographic characteristics of household and r.p.
- synthetic class of income
- all main consumption components

HBS main consumption components by income classes



Well-known trend of food costs: the first class of income reserve 24% while the richest class reserve 14%

3. Household income in HBS: Creation of a new variable 1/3

Income is collected in different ways in IT-SILC and HBS.

HBS

Income is observed at **household level**.

There is an ad hoc section about income and savings that includes:

1 A **multi-response question** about the average household income (in classes);

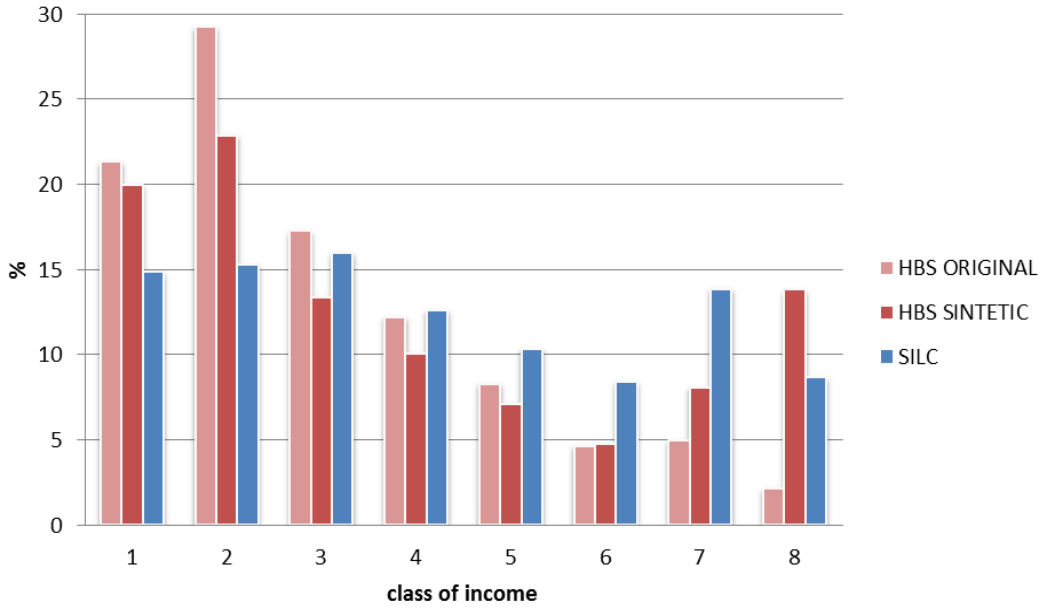
2 A question on **the use of income**: the household has to indicate if all the income is spent in household or, instead, **if there is a saving**;

3 If there is a saving, a **last question** permit to declare **the total amount.**

3. Household income in HBS: Creation of a new variable 2/3

The additional information from the HBS income section has been used to estimate a new income variable by preserving 81% of original distribution.

The **synthetic variable** has decreased the household income's underestimation in HBS.

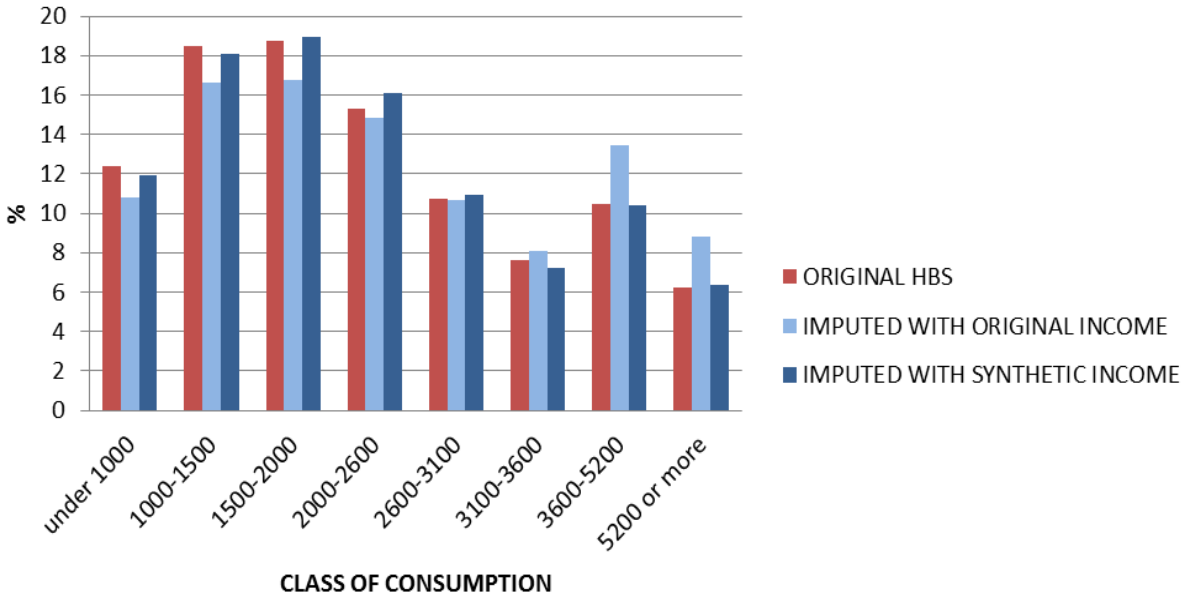


The Hellinger distance slightly decrease from 17,2% to 12,9%.

3. Household income in HBS: Difference between imputed classes of consumption 3/3

The auxiliary information was used in the matching procedure in further restricting the subset of potential donors.

It is possible to consider the difference between the **imputed consumption** in IT-SILC using the original HBS income or using synthetic income as auxiliary information.



The Hellinger distance decreases from 5,1% to 1,5%.

Different matching approaches

1. **Non-parametric (random hot deck) under CIA**

2. **Exploration of uncertainty**

3. **Auxiliary information to relax the CIA**

1. CIA = independence between income and consumption given some common information in both the data sources. UNREALISTIC

2. Average width of uncertainty bounds = 7.8% TOO WIDE

3. Household monthly income collected in HBS as aux info. Random hot deck with restricted subset of potential donors

- living in the same macroarea
- having the same number of durable goods
- same or in the upper/lower class of income.

Suitable results

1.unlikely assignments between classes of consumption and income is limited (classes of consumption that differ more than three from the respective class of income)

2.household typology (not selected as matching variable) presents a **similar distribution to the original in HB.**

Promising starting point for the distributional analysis on the propensity to consume by main socio-demographic variables

		Consumption	under 1000	1000-1500	1500-2000	2000-2600	2600-3100	3100-3600	3600-5200	5200 or more	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**

The Hellinger distance of the joint distribution of imputed and observed consumption is 5%

Some simulation on HBS data, using different methods of classification

i. **multinomial logistic regression**
ii. **classification trees**
iii. **random forest**

Dependent variable:

- household monthly consumption expenditures divided into seven classes using the same monetary thresholds of income classes.

*Covariates**:***

- different set of variables

Test each set individually and each combination with the common variables

5. Predicting consumption: comparison between covariates 2/2

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

Comparing the overall classification error between models and covariates, every models identify the same set (the union of 1 and 2).

The combination of common variables and most predictive ones classifies correctly the 56,3% of total households in HBS survey.

Multinomial regression

confusion matrix between observed and predicted class of consumption

OBSERVED	PREDICTED							
	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%
								56,3%

Despite the low percentages for the highest classes, prediction in classes non-contiguous to the diagonal is very limited.

6. Asking consumption in EU-SILC: Ad-hoc module 1/4

Task Force on the revision of the EU-SILC legal basis :

- ✓ Short-term fixed every 3-years modules (on children, health, housing conditions and labour)
- ✓ **Rolling module every 6-years**. Topics proposed:

  Quality of life, social and cultural participation;

  Over-indebtedness, wealth, consumption

  Access to services, social transfers in kind
  Intergenerational transmission of disadvantages

The background

Browing et al (2002): “food at home” and “food outside home” are two predictors that explain a good part of non-durable expenditure

INSEE (French National Institute of Statistics and Economic Studies)
twelve questions were added in the monthly consumer confident survey:

- three about expenditure on food (at home and outside) and utilities
- eight questions to collect information on household regular expenses (clothing, public transport and other, binary variables)

A similar set of questions was added in the **Household Wealth Survey run between 2009 and 2010.** (good match with HBS 2010 data)

INSEE did not ask the amount of transport expenditure but only two binary variables were collected (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.

Food, housing and transport expenditures are three good predictors of classes of consumption. (ESTAT, ISTAT)

According to the structure of total consumption, asking for food at home and food outside home make an increase in the explained total variability of consumption is achieved (from 63% to 66%).

Some evidences from HBS's data:

- one third of the total amount of transport expenditure is performed by expenses for gasoline or other fuel for cars and bicycles

- one fifth is represented by car or bicycle's insurance

- 15% regards to expense for buying a new car

A possible structure for the new module:

- ✓ Food at home (amount)
- ✓ Food outside home, including restaurants, at-work restaurants, bar.. (amount)

- ✓ Public transport, as train, bus, plane, subway and taxi (amount)

- ✓ Private transport:

  Purchase of a new car (if yes amount)

  (if household have a car or a motorcycle)
  Gas expenses for household cars or motorcycles(if yes amount)
  Assurance for household car or motorcycle(if yes amount)

7. Some concluding remarks

- There is a good degrees of comparability among HBS and IT-SILC, but a better harmonization of housing costs variable is advisable

- The role of HBS income variable is valuable for overcoming CIA in matching procedures; it is essential to underline the presence of a question about savings that has allowed us to reconstruct HBS income (decreasing the difference with IT-SILC and improving the quality of the matching process).

- New shared consumption variables in IT-SILC have a great potential and explanatory power; the new module can be a source of potential auxiliary information