An Outline of IT-SILC Data Processing

Clodia Delle Fratte and Stefano Gerosa

17 May 2018

Istat

Index

1. Objectives and structure of IT-SILC data processing sequence
2. Treatment of identification variables and definition of t+1 sample
3. Treatment of qualitative variables
4. Treatment of quantitative variables
5. Weighting procedure and weights calibration

Data processing in IT-SILC

The C&C process is the **set of techniques applied to raw data** (i.e. checking, editing, imputation) in order to:

- control and minimise the **non-sampling error**
- build **a coherent and complete set of microdata** (s.t. for each record/observation there are no missing values for every variable of interest and all rules of consistency [edits] are verified)
- obtain **more accurate estimates** of the parameters/measures of interest

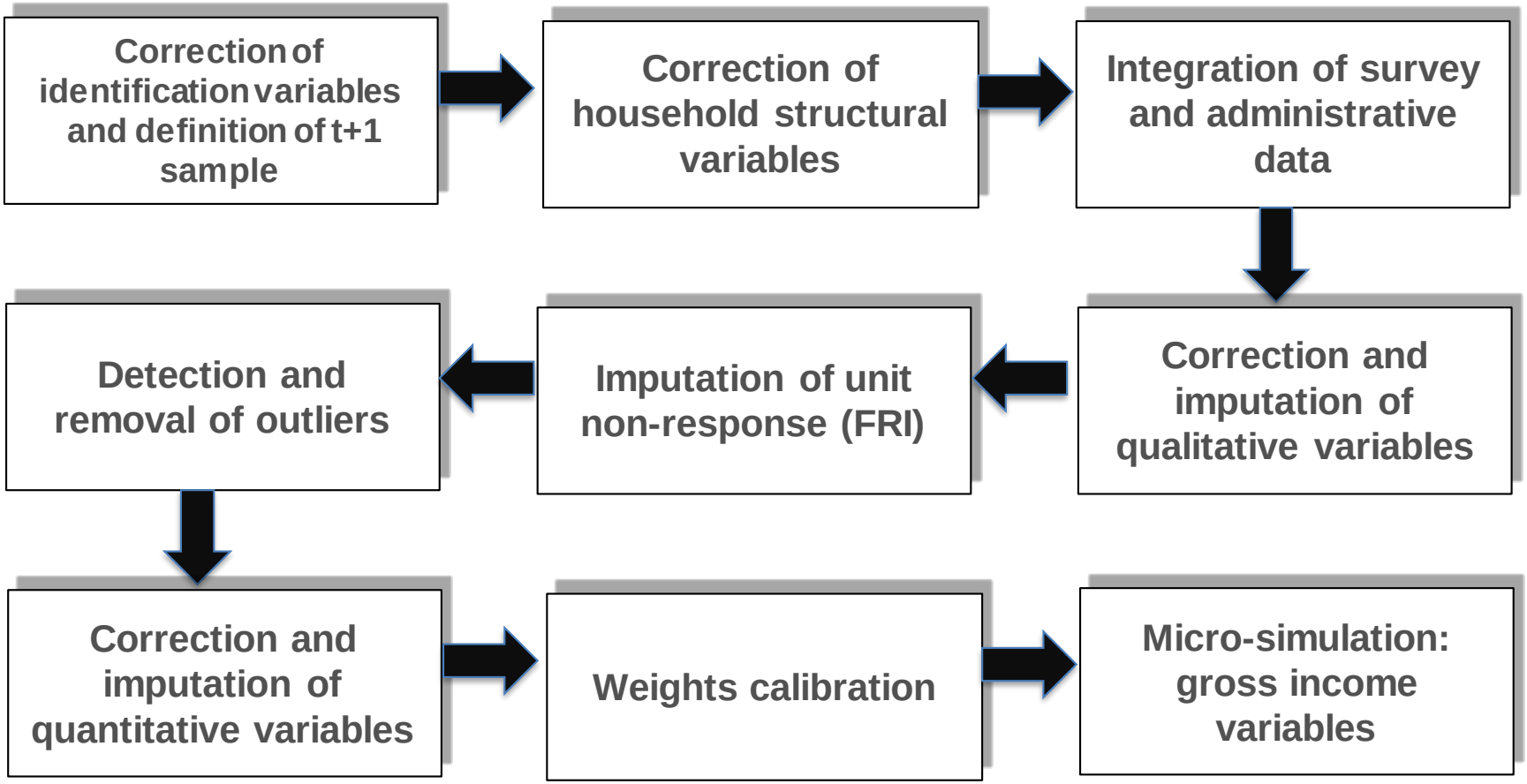
Non-sampling error

- **List errors** (incorrect omission/inclusion of sample units, duplication)
- **Processing errors** (i.e. incorrect registration)
- **Measurement errors**: observed sample value is not the «true» value (wording, interview technique,…)
- **Missing values**: unit and item non-response

The features of Eu-Silc survey (information at both the household and individual level, panel structure) is particularly demanding from a C&C perspective:

- **within-record** and **across-record** (i.e. intra-household) **consistency**
- **longitudinal consistency** (e.g. dynamic consistency of individual and household observed characteristics)

The IT-SILC Correction Process



Correction of identification variables and definition of t+1 sample

- **CAPI (computer-aided personal interview) technique and controlled registration**:
  - pre-loaded information from registers and past interview (dependent interviewing)
  - automatic control of questionnaire routes
  - soft and hard controls on variables (i.e. range of admissible values for some quantitative variables)

→ transition to CAPI **significantly reduced non-sampling error**

- **Control of household and personal identification numbers (DB030/RB030)**
  - Controls for ID duplication or inversion (i.e. different ID for individual with same identifying variables)

- **Correction of identifying variables:** name, sex, date of birth, place of birth.
    - Pre-loaded information from SIGIF (integrated archive of information from all ISTAT household surveys)
    - Dependent interviewing from 2011, with significant improvement in data quality.

UNIT NON-RESPONSE

- **Definition of household and individual survey outcome**

HH contact	HH outcome	Eligibile (16+)	Individual Outcome	Frequency	Percentage
Yes	Yes	Yes	Impossibility	287	0.6
Yes	Yes	Yes	Refusal	643	1.3
Yes	Yes	Yes	Positive	37,110	72.7
Yes	Yes	No	.	8,047	15.8
Yes	No	.	.	4,119	8.1
No	No	.	.	815	1.6

Correction of household structural variables

- **A generalized procedure for the C&C of household structural variables** :
  - Individuation and correction of inconsistencies among some demographic variables: **sex, age, relationship within the household, legal marital status**.
  - Identification of **couples** (partner ID) and **household type**

	Households	Corrected households	%
2010 (PAPI)	**19,149**	**3,854**	**20.1**
2013 (CAPI)	**18,693**	**1,119**	**6.0**

Integration of survey and administrative data

- **Record linkage:** Administrative (tax-code registers, pension registers) and survey data are integrated by **an individual 'matching-key', the tax code** (the personal ID number assigned to each person by the Italian tax authority);

- **Reclassification and alignment of income components in the matched dataset** :

  - formulation of hypotheses for the **removal of inconsistencies between income components and/or income values**;

  - **identification of income recipients**, for different types of income → income "flags" are then **constrained** for the rest of the correction process

Correction and imputation of qualitative variables

- **Fellegi-Holt methodology for automatic edit and imputation** (implemented in the software Concord-SCIA, developed in ISTAT):
  - **Principle of least change of original data:** identification of the minimum set of items to be corrected within a record to resolve an inconsistency
  - **Edit and imputation as a unique process**: imputation constraints form a "feasible region" within the complete edit set. Values imputed from the feasible region satisfies all edits with certainty.
  - **Retaining the structure of the data:** maximal preservation of marginal and joint distributions of original non-inconsistent data

Correction and imputation of qualitative variables

- **The complete edit set (i.e. the set of rules each record must satisfy) is made of :**
  - **Formal edits:** all rules logically implied by the questionnaire structure → we use **an automatic procedure** developed by P.P. Massoli (ISTAT) that generates all formal edits considering **the questionnaire as an acyclic oriented graph.**
  - **Substantial edits**: rules that identify **across-record inconsistencies** (e.g. an household with no children with an individual that receives a child allowance) or **inconsistencies generated by the use of administrative data** (e.g. an individual with a pension that has not answered to the relevant part of the questionnaire).

Correction and imputation of qualitative variables

- **The correction sequence:**
  - **Household questionnaire**
  - **Individual questionnaire:** since the complete edit set (the set of explicit edits derived from the intrsection of formal and substantial edits) is too large, the imputation is carried out in steps:
    - Sections on currently working/have worked in the past
    - Section on income flags
    - Section on working conditions
    - Income sections (employee, self-employed, pensioners, other income)
    - Education and health section

A measure of the impact of the C&C process – Qualitative variables (PAPI 2010 vs CAPI 2013)

PERCENTAGE OF CELLS* TREATED IN C&C PROCESS

IT-SILC 2010 (PAPI)

	Register form	Household interview	Personal interview	ALL
NOT MODIFIED	94.7	99.0	92.1	93.9
MODIFIED	0.9	0.1	1.0	0.8
IMPUTED	3.7	0.4	5.8	4.3
BLANKED	0.7	0.5	1.1	0.9
TOTAL	100.0	100.0	100.0	100.0

IT-SILC 2013 (CAPI)

	Register form	Household interview	Personal interview	ALL
NOT MODIFIED	99.5	99.8	94.9	96.5
MODIFIED	0.2	0.0	0.8	0.6
IMPUTED	0.1	0.0	3.6	2.5
BLANKED	0.2	0.2	0.6	0.5
TOTAL	100.0	100.0	100.0	100.0

* Each cell corresponds to a single sample unit and a single question item. Only qualitative variables are considered.

Imputation of unit non-response

- **Nearest-neighbour imputation methodology** (implemented in the SAS procedure FRI developed by P.P. Massoli (ISTAT)): the erroneous record (i.e. the record with no individual information, neither in the current survey or from administrative sources/previous survey) is replaced by an error-free record of a «close» donor.

  - **Definition of the donor set**: a set of **strata variables** are defined (geographical area, age, sex, citizenship, marital status household's "ability to make ends meet" answer) and their intersection form a group of subsets from which a donor is selected.

  - **Individuation of the donor**: a set of **match variables** (e.g. total household income) are defined and are used as arguments of a (euclidean) distance function. The donor is the record closest to the «missing» one according to the chosen distance.

Detection and removal of outliers

- **Hidiroglou-Berthelot methodology**

  - **Transformation of the original variable (Y)** to make it more symmetric wrt to the median (MEy):

$$HB_y = \begin{cases} \dfrac{Y - ME_y}{Y} & if\ 0 < Y < ME_y \\ \dfrac{ME_y - Y}{ME_y} & if\ Y \geq ME_y \end{cases}$$

  - **Calculation of minimum and maximum thresholds** outside of which observed values are considered outliers:

$$HB_{min}(k) = ME_{HB_y} - k(ME_{HB_y} - Q1_{HB_y}) \quad \text{for} \quad HB_y < ME_{HB_y}$$

$$HB_{max}(k) = ME_{HB_y} + k(Q3_{HB_y} - ME_{HB_y}) \quad \text{for} \quad HB_y \geq ME_{HB_y}$$

Detection and removal of outliers

- For some variables **a multivariate approach is necessary** (e.g. housing costs are considered in relation to the size of the house and the number of rooms; financial incomes are evaluated in relation to reported savings/wealth stock)

- Starting from the 2nd wave, **a similar method is also applied in a longitudinal direction**, considering the ratio between current and previous values of each variable, in order to control for excessive year-to-year variations

Outliers for the variable “employee income”

	Removed		Not modified		All	
	N	%	N	%	N	%
Right tail	15	0.2	7,784	99.8	7,799	100,0
Left tail	161	3.0	5,155	97.0	5,316	100,0
All	176	1.3	12,939	98.7	13,115	100,0

Correction and imputation of quantitative variables

- **A multivariate sequential regression model (implemented by IVEware software of the University of Michigan):**
  - Imputation is carried out variable by variable, under the **MAR (missing-at-random) hypothesis**, starting from that with the lowest number of missing values.
  - **Different regression models are used** (LRM, logistic regression,…) depending on the variable type (continuous, count,…)
  - An **iterative imputation scheme** is used, where the set of covariates is updated and previous imputed values may change in order to better preserve the correlation among variables.

Correction and imputation of quantitative variables

- for each variable, it is possible to select the relevant covariates, such as:
  - **Territorial characteristics**: geographical area, municipality (type and size);
  - **Individual characteristics**: gender, age, marital status, education, health condition, labour market status, profession;
  - **Household and dwelling characteristics***:* household size, type of dwelling,...;

- it is possible to **restrict the imputations** (e.g., imputing only a particular subset of cases) and to **assign upper and lower bounds** to the imputed values.

- it is useful for **transforming in precise values the approximate figures** indicated by those who do not remember the exact amount of their income. The precise values are imputed by IVEware within a band centered on the approximate figures given by the respondent.

Correction and imputation of quantitative variables

- **The imputation sequence**

Expenses, income and transfers **at the household level** were dealt with first, as they help to **establish the standard of living of the family**, and then to explain the level of other income components. The other sections were addressed starting with those with lower rates of missing values. The order was followed in the treatment of the different sections was therefore as follows:

  - Expenses, income and transfers at the household level
  - Employee income
  - Retirement income
  - Self-employment income
  - Financial Income
  - Real estate income
  - Other income components.

Correction and imputation of quantitative variables

	NUMBER OF CASES			MEAN			…
	Raw	Imputed	Total	Raw	Imputed	Total	…
ITALY	14398	356	14754	1.436	956	1.425	…
GEOGRAPHCAL DISTRIBUTION							
North-West	3653	100	3753	1.527	995	1.512	…
North-East	3955	66	4021	1.478	988	1.470	…
Centre	3285	103	3388	1.434	939	1.419	…
South	2505	57	2562	1.289	902	1.280	…
Islands	1000	30	1030	1.321	924	1.309	…
SEX							
Male	7701	168	7869	1.603	1.001	1.590	…
Female	6697	188	6885	1.245	916	1.236	…
AGE GROUPS							
15 - 34	3647	122	3769	1.135	885	1.127	…
35 - 44	4197	84	4281	1.415	921	1.405	…
45 - 54	4270	78	4348	1.569	1.078	1.560	…
55 - 64	2112	58	2170	1.719	1.015	1.700	…
<64	172	14	186	1.593	868	1.539	…
…	…	…	…	…	…	…	…

Weighting Procedure and Weight Calibration

- **The weighting procedure**
  - Each household in the sample is initially assigned **a design weight given by the inverse of its inclusion probability**, so that household j in the municipality i of stratum h has an initial weight:

$$p_{ji} = \frac{1}{\pi_{hi}} = \frac{P_h}{P_{hi}} * \frac{M_{hi}}{m_{hi}}$$

  - The base weights are then **adjusted for total non-response at the household level:**

$$\widetilde{p_j} = \frac{p_j}{\pi_j}$$

where **the response probability π_j is estimated by a logistic regression model** (main covariates: size of the municipality, household size, sex, age and nationality of the householder)

Weighting Procedure and Weight Calibration

- A first calibration step is then applied to assure **the same population structure of LFS with respect to education level and professional position** (since non-response is highly correlated with both variables).

- Final weights are then obtained applying **a calibration of household weights to external data sources (registers**), so that calibrated weights ψ_j are able to exactly reproduce a p_j set of known totals (distribution of the population by sex and 14 age-groups at the NUTS-1 level; distribution of the population by demographic size of the municipality at Nuts 1 level; distribution of non-national population at NUTS-1 level by sex and 2 age-groups; number of households at NUTS-2 level).

- The procedure also ensures that **each member of the same household has the same weight**.

Weighting Procedure and Tax Data

We are currently working on the use of tax data to improve our weighing **procedure**

- Current weights **do not correctly represent "fiscal population"**: number of income-tax payers and size of fiscal income aggregates (e.g. employee income and income from self-employment reported in fiscal) are **under-estimated** in the actual IT-SILC sample.

- Unit non-response is **correlated with both type and level of fiscal income**.

- Use of fiscal information both for estimates of response probability of household/individuals and for calibration **raises mean household income and the share of income from self-employment**, but **has minor effects on distributional indicators** (e.g. poverty rate and Gini index).

Micro-simulation of gross income variables

- **The construction of IT-SILC gross income variables**
  - The microsimulation model SM2/EU-SILC is developed and applied for the net-to-gross conversion of incomes
  - Survey and administrative data are jointly used in the micro-simulation process
  - Final microsimulation estimates are compared to register data at the micro-level, in order to assess the quality of the microsimulation.