How to measure income components: General perspective and the choices of Italy

Paolo Consolini – Istat
Presentation Overview

- Sources of household income statistics, data analysis and collection units, reference period, source of errors

- The choices of Italy: sample design, multisource data collection, errors correction process

- Joint use of survey and administrative data together: Why? What benefits?

- An outline of main It-Silc process phases, managing new administrative sources, critical factors, mapping

- Scheme of the micro-data integration process: data processing operations, requirements, classification and solution to inconsistencies, values reconciliation

- Concluding remarks and future developments
Sources of household/personal income statistics

- Most income distribution statistics rely on data collected in **household surveys**
- Some countries (i.e. Nordic Countries and Canada) use administrative data to this scope (**Register-based statistics**)
- Some income components are imputed by model (i.e. personal tax, social contribution, imputed rent)
Survey income statistics

• Income data are usually collected through sample surveys (designed household income surveys or multi-topic surveys)
• Household surveys generally collect information from usual residents of private dwellings
• The design of the sample is based on appropriate sampling techniques in order to obtain a fixed statistical precision
• The mode of data collection in household surveys may vary: usually face-to-face interview is applied (occasionally telephonic) together with Computer-assisted technique
Register-Based Statistics (RBS) on income

• There is a legal basis to use it for statistical purposes, personal income data from registers
• RBS provide total or near total population coverage and can therefore be used to produce more detailed statistics for small areas or subgroups of population
• RBS allow lower respondent burden and are generally a less costly means of producing statistics
• RBS are not subject to sampling and non-response errors
• Nevertheless RBS could be not relevant for statistical purpose and present constrains for timely releasing
Measurement units

• It’s important to distinguish between the **data collection unit** and the **data analysis unit**

• The first will depend on design of survey (households or dwellings) the second will depend on type of income (wages and salaries are best collected at the individual level whilst imputed rent should be collected at the household level)

• Private Household (EU-SILC): defined as a person living alone or a group of people who live together in the same private dwelling and share expenditures, including the joint provision of living essentials
Measurement units

*Household member* (EU-SILC)

Subject to the further and specific conditions shown below, the following persons must, if they share household expenses, be regarded as household members:

1. Persons usually resident, related to other members
2. Persons usually resident, not related to other members
3. Resident boarders, lodgers, tenants
4. Visitors
5. Live-in domestic servants, au-pairs
6. Persons usually resident, but temporarily absent from dwelling (for reasons of holiday travel, work, education or similar)
7. Children of the household who are being educated away from home
8. Persons absent for long periods, but having household ties: persons working away from home
9. Persons temporarily absent but having household ties: persons in hospital, nursing home, boarding school or other institution
Measurement units

**Household member (EU-SILC)**
Further conditions for inclusion as household members are as follows:

For categories (3) (4) (5):
Currently has no private address elsewhere or their actual or intended duration of stay is six months or more

For category (6):
Currently has no private address elsewhere and their actual or intended duration of stay is less than six months

For categories (7) and (8), irrespective of the actual or intended duration of absence, the person:
Currently has no private address elsewhere, is the partner or child of a household member, and continues to retain close ties with the household and considers this address to be his/her main residence

And for category (9):
The person has clear financial ties to the household and the actual or expected duration of absence from the household is less than six months
Reference period

The international standards state that household income statistics should relate to a full year to take into account seasonal variations in incomes. Annual income includes the income obtained from all sources over a period of a year.
Sources of errors (estimated value differs from true one)

Income distribution statistics are generally subject to two type of errors: non-sampling and sampling error

Sources of non-sampling error include:

- non-response errors
- errors in the frame of units sample (list errors)
- errors in reporting by respondents or recording of answers (measurement errors)
- errors in coding and processing the data
Non sampling errors

Table 3.2 Types of non-response

<table>
<thead>
<tr>
<th>Problem</th>
<th>Description</th>
<th>Common solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Unit non-response</td>
<td>Failure to obtain any information on a sample household, including the household interview and personal interviews in the household</td>
<td>Weighting</td>
</tr>
<tr>
<td>(2) Partial unit non-response</td>
<td>Failure to obtain a personal interview with a subset of the eligible adults in a household</td>
<td>Weighting or full-case imputation</td>
</tr>
<tr>
<td>(3) Item non-response</td>
<td>Failure to obtain some target variables in an otherwise completed interview (this generally affects non-income variables in register countries and all – especially income – variables in survey countries).</td>
<td>Imputation for missing items</td>
</tr>
<tr>
<td>(4) Partial item non-response</td>
<td>Refers to the situation when some but not all the information is obtained on a target variable. The most important case is that of detailed income components: a part of the component may be missing, and/or conversion may be required from the collected net to the required gross amount.</td>
<td>Imputation for the missing part</td>
</tr>
</tbody>
</table>

Source: Verma and Betti, 2010
Non sampling errors

It’s worth knowing:

- Non-response can affect the reliability of the results and can introduce a bias (= true – estimated value)
- The magnitude of this bias depends upon the level of non-response and the difference between the characteristics of people who responded to the survey and those who did not
Sampling errors

In **general terms** is a measure of the variability that occurs by chance because a sample, rather than the entire population, is surveyed.

In **statistical terms** is the difference between a sample statistic used to estimate a population parameter and the actual but unknown value of the parameter.

One measure of the likely difference is given by the standard error (SE). Another measure of the likely difference is the relative standard error (RSE => SE as a percentage of the estimate).

**Strategies to reduce the sampling error:** increase the sample size, improve the sample design (stratification)
EU-SILC: Italian choices (IT-SILC)

SILC is an output harmonized instrument

Degree of freedom in terms of:

✓ Data collection strategy: CAPI/CATI & Admin Data (multisource)

✓ Sampling: Two stages and 4 waves rotational sample design

✓ Questionnaire design: Contact-Register-Household-Individual Form

✓ Errors correction strategies: editing - imputation - calibration
Two stage sampling design:
✓ First stage units (PSU): the municipalities
✓ Second stage units (SSU): the households

PSU are stratified according to their demographic size (i.e. # residents)
Stratification is carried out inside each administrative region
PSU are selected in each stratum with probability proportional to their size (systematic sampling method by Madow, 1949)

Secondary sampling units (i.e. households) are not stratified
Households are selected with equal probability by systematic sampling in each selected municipality from population register lists (LAC).
No substitution of unit non-response has been applied
IT-SILC Sample size

- the **actual sample size** which is the number of sampling units selected in the sample (8,000 - till 2010 => 10,000 – 14,000 between 2014-2018)
- the **achieved sample size** which is the number of observed sampling units (household or individual) with an accepted interview (about 21,000 households and 41,500 individuals (16+))
- the **effective sample size** which is defined as the achieved sample size divided by the design effect with regards to the at-risk-of poverty rate indicator

The minimum effective sample size is 7,250 households (according to Eurostat). Being the sampling design effect about 2 (e.g. 2.1904 in 2011), the minimum sample of households to meet the requirements of the regulation is 15,880. We choose a larger sample size to provide reliable estimates at regional level
Questionnaire

- Register form
  - Individual information (sex, date of birth, country of birth, citizenship, marital status....)
  - Household members relationships (with reference person – 20 possible relationships)
  - Child-care target variables
Questionnaire

- **Household information**
  - Dwelling (dwelling characteristics and amenities, durables, utility bills and other costs…)
  - Rented accommodation (tenant(s), rent costs…)
  - Owned accommodation (owner(s), mortgage..)
  - Economic situation (material deprivation items, household income, savings, support network…)
  - Income of household members younger than 16
  - Children economic situation (Child-specific material deprivation items, e.g. Two pairs of properly fitting shoes, One meal with meat, chicken or fish or vegetarian equivalent at least once a day)
Questionnaire

- Individual information
  - Education
  - Health
  - Employment (à la ILO)
  - Main Job characteristics
  - Previous job experience
  - Calendar of activities (income reference year)
  - Income from dependent employment
  - Income from self-employment
  - Other job-related income
  - Pensions (from individual private plans & from foreign bodies)
  - Other income information (savings and rent from properties, alimonies and other private transfers)
  - Individual material deprivation (e.g. Spend a small amount of money each week on yourself, Replace worn-out clothes by some new ones)
The IT-SILC Scheme of Errors Correction Process

1. Correction of identification variables and definition of t+1 sample
2. Correction of household structural variables
3. Integration of survey and administrative data
4. Correction and imputation of qualitative variables
5. Correction and imputation of quantitative variables
6. Weights calibration
7. Imputation of unit non-response (FRI)
8. Detection and removal of outliers
9. Micro-simulation: gross income variables

The IT-SILC Scheme of Errors Correction Process
The EU-SILC gross income components

Total gross household income (HY010) is computed as:

- The sum for all household members of gross personal income components
- Gross employee cash or near cash income (PY010G),
- Company car (PY021G),
- Gross cash benefits or losses from self-employment (including royalties) (PY050G),
- Pensions received from individual private plans (other than those covered under ESSPROS) (PY080G),
- Unemployment benefits (PY090G),
- Old-age benefits (PY100G),
- Survivor's benefits (PY110G),
- Sickness benefits (PY120G),
- Disability benefits (PY130G),
- Education-related allowances (PY140G):
- Plus gross income components at household level
  - imputed rent (HY030G),
- Income from rental of a property or land (HY040G),
- Family/children related allowances (HY050G),
- Social exclusion not elsewhere classified (HY060G),
- Housing allowances (HY070G),
- Regular inter-household cash transfers received (HY080G).
The EU-SILC disposable household income

**Total disposable household income (HY020)** can be computed as:

The sum for all household members of gross personal income components:
- Gross employee cash or near cash income (PY010G),
- Company car (PY021G),
- Gross cash benefits or losses from self-employment (including royalties) (PY050G),
- Pensions received from individual private plans (other than those covered under ESSPROS) (PY080G),
- Unemployment benefits (PY090G),
- Old-age benefits (PY100G),
- Survivor's benefits (PY110G),
- Sickness benefits (PY120G),
- Disability benefits (PY130G),
- Education-related allowances (PY140G);

Plus gross income components at household level

Minus
- Regular taxes on wealth (HY120G),
- Regular inter-household cash transfer paid (HY130G),
- Tax on income and social insurance contributions (HY140G).
The choice of Italy
Multiple data collection strategy: Why? What benefits?

• **for data providers:** *lower* response burden

• **for producers:** Decreasing budgets, improving cost efficiency, sharing and building on good practice

• **from user's point of view:** *higher* data quality and *more detailed* information
How is possible to combine administrative and survey data?

By selecting an individual “matching-key” able to link the same unit across different data-sources.

Which is the best “matching-key” to our scope?

The “tax number” (i.e. the personal identification number assigned to each person by the Italian Tax Authority)
Phases of IT-Silc multisource data collection process

- Sample frame (Object selection)
- Questionnaire Revision (Variable selection)
- Administrative data collection
- Tax Authority and Social Security Body
- Data reading and cleaning
- Data matching
- Micro-data integration
- Respondent (individuals, reference household person)
A simplified scheme of the exploitation of new Admin sources

Identification of unexplored administrative data sources → Meta-data analysis → Data acquisition (data cleaning & reading) → Feasibility studies

Assessment of data quality (validation with external benchmarks)

rejected → Redesigning the data acquisition action plan

accepted → Inclusion on data integration It-Silc process
Critical factors to extensive use of administrative data

- The adequacy of concepts, classifications, nomenclatures and variables of administrative data for statistical needs (relevance & harmonization)
- The quality of the administrative data itself (coverage)
- The access conditions and terms/times of transmission of administrative data
- The stability of the administrative source and its contents (legislative changes)
- IT Capacity of the NSI in the management of an Administrative Data Repository
Information security
Istat policy of protecting personal data

Authentication, authorization and accounting (AAA) system for tracking user activities on Istat Data Warehouse

Identifier elements need to be stored separately from the "research data"

All personal information will be destroyed once it is no longer needed for the statistical purpose

Protection of confidentiality in data dissemination (statistical disclosure control)
Mapping administrative sources for statistical purposes

**Income components:** (Target variables)

- Employee cash or near-cash income (PY010G/PY010N)
- Cash benefits/losses from self-employment (PY050G/PY050N)
- Pensions for different functions (old age, etc.) (PY100N-PY110N-PY130N)
- Non-pension cash benefit for different functions (PY090N-PY100N-PY110N-PY130N)
- Interest, dividends, profit from capital investment (HY090N)
- Incomes from rental of a property or land (HY040N)

**Administrative sources:**

- CU Withholding Tax Register
- 730 Tax Return Register
- U.P.F. Tax Return Register on self-employed
- Pensioners Register
- Non Pension cash beneficiaries Register
- Social Insurance Employee and Payroll Register (Uniemens)

**Administrative sources** (INPS)

**Incomes from rental of a property or land (HY040N)**
The Integration Process of Survey & Admin Source:

1. Input data (survey): **Personal tax numbers** and personal items (used to generate them) are retrieved from SIGIF (Population register)

2. Input data (administrative sources): **Personal tax numbers** are checked and corrected. Information coming from multiple records and relating to the same person is organized in order to avoid duplications or double counting

3. Exact matching procedure and harmonization: survey and administrative sources are matched using the **Personal tax number** as the key variable. The result is a matched file

4. Detecting and solving inconsistencies on income in the matched file: an inconsistency arises whenever survey and administrative sources assign a different type of income to the same person

5. Reconciliation of incoherent income values: even in presence of a consistent structure of incomes between administrative and survey data, they could display incoherent income levels
Data processing operations – Part 1

- Selection of units and variables (tax code, etc.) form SIGIF
- Generator of “missing” tax codes (matchig-key)
- Release of tax codes file (It-Silc sample)

SIGIF Frame of IT-SILC sample

Population Register (LAC) (Istat)

Pension Register (Istat-Inps)

Istat

Exact matching (tax codes)

Tax codes registers (Revenue Agency)

Sogei (IT-partner)

Istat Data processing operations – Part 1

- Exact matching (tax codes)

Selection of units and variables (tax code, etc.) form SIGIF

Generator of “missing” tax codes (matchig-key)

Release of tax codes file (It-Silc sample)
**Data processing operations – Part 2**

(Istat)

Exact matching (It-Silc-Pension Register) → Units linked → Pensioners and pensions File (It-Silc) → Detection and correction of duplications/outliers → Administrative Income Database (cleaned & integrated data)

Not linked units File → Units not linked

Exact matching (It-Silc-Tax Registers) → Units linked → Tax forms Registers (CU, 730, Upf) → (Revenue Agency)

Tax statements File (CU, 730, Upf) → Reading procedures, detection and correction of duplications/outliers

Solution of inconsistencies and Reconciliation of incoherent values

Survey Income Database (Capi/Cati)

Integrated database of income data (output)
Graphical representation of Admin Income Database

AID is a “integrated statistical database for selected objects (sample units)” that combines several administrative sources

\[ \Omega' = A \cup B \cup C \cup D \cup E \cup F \cup G \cup H \cup I \cup L \cup M \]
The classification/reclassification of the income components

The framework:

1) Primary classification of the income components based on CU tax withholding register

2) Secondary classification if the income components based on auxiliary information: mod.730 e Upf tax returns

3) Reclassification of the income components based on Pension Register
Primary classification of the income components

Cu tax withholding

Variables employed: flag on kind of taxable incomes, deductions, type of withholding agents, social contributions

Tax statements
File: Cud/770

Tax form CU
Part B (fiscal data)

Employee income
Pensions
Not pension cash benefits
Other incomes (otherwise classified)
Secondary classification of the income components based on auxiliary information: mod.730/Upf tax returns (sets $A$ and $E$)

$\Omega = A \cup B \cup C \cup D \cup E$

**LEGENDA**

- Only 730 tax returns
- Only CU tax withholding
- Only UPF tax returns
- CUD and 730 tax declarations
- CUD and UPF tax declarations
Reclassification & reconciliation of the income components based on Pension register and Tax registers

- It-Silc units matched with Pension register
- It-Silc units matched with Tax Register
- It-silc units matched with Pension register and Tax register
- It-silc units matched with Tax register but not linked with Pension register
- It-silc units matched with Pension register but not linked with Tax register

(7.47% pensioners Eu-Silc'06)
(91.62% pens. Eu-Silc’06)

“Old age and Survivors Annual pensions” of PR are close to “Annual gross Pension” of Tax Register

DREL1_PEN \leq 5% YES

DREL2_PEN \leq 5% NO

DREL3_PEN \leq 5% YES

NO

“Old age, Survivors and Disability Annual pensions” of PR are close to Annual gross Pension” of TR

Old Age, Survivors, Disability Pension (Calculated as Monthly amount X nr payments) of PR is close to Annual gross pension of TR

It-Silc frame matched with both Pension and Tax registers for which are applied other decisional rules

(0.91% pensioners Eu-Silc'06)

(14.33% pensioners Eu-Silc'06)
Micro-integration of Pension and Tax Registers

Units of the It-Silc Frame matched with the Pension Register

Micro-integration

id (Tax code) = matching variable
Gross Annual Pension = comparison variable

Units of the It-Silc Frame matched with the Tax Registers

Units of It-Silc Frame matched with Pension Register but not linked with Tax Registers

Lack of information: Tax

Units with no-taxable pensions
- Net income = Gross income

Units with taxable pensions
- Net income computed by micro-simulation model

Units of It-Silc Frame matched with both Pension Register and Tax Registers

Full information

Pensions reported in CU
- (C)

Pensions reported in 730
- (D)

Pensions reported in UPF
- (E)

Units of It-Silc Frame matched with Tax Registers but not linked with Pension Register

Lack of information: Pension type

Assignment of the pension type

Survivor’s pensioner

Old age’s pensioner

(A)  (B)  (C)  (D)  (E)  (F)
The micro-data integration process of different data-sources involves the satisfaction of the following requirements (P. van der Laan 2000):

Part 1

a. harmonization of units: are the statistical units defined uniformly in all sources? (comparability in space and time)

b. harmonisation of reference periods: do all data refer to the same period or the same point in time?

c. completion of populations (coverage): do all sources cover the same target population?

d. harmonisation of variables/classifications: are corresponding variables defined (classified) in the same way? (comparability in space and time)
The integration process of different data-sources involves the satisfaction of the following requirements (P. van der Laan 2000):

Part 2

e. adjusting for measurement errors (accuracy): after harmonising definitions, do the corresponding variables have the same value?

f. adjusting for missing data (item non-response): do all the variables possess a value?

g. derivation of variables: are all variables derived using the combined information from different sources?

h. checking overall consistency: do the data meet the requirements imposed by identity relations?
Inconsistencies analysis on incomes

- A first type of inconsistency arises from the presence of a specific component of income in the survey (declared during the interview) that is not reported in the Tax registers.

- A second type of inconsistency occurs when a type of income is recorded in Tax registers but not in the survey.
Solution of inconsistencies:

- **Scope:** assigning the “most likely” income pattern to each unit

- **Strategy:** it depends on the income type (generally the administrative source is assumed to be more reliable than survey data - except for self-employment income). Hierarchical approach: firstly employee income, secondly self-e. incomes, ...

- **Analysis:** comparison of the respondent’s professional status (or economic activity) between “survey” and “administrative” sources; identification of potential classification errors in income structures from the survey; inclusion of incomes from underground economy

- **Actions:** removal-replacement-integration of income components

- **Advantages:** avoiding misclassification of income components, reducing double counting (the same amount classified in different way) and removing interviewer falsification.
Reconciliation of self-employment income values is based on:

**Maximum Rule:** impose the maximum value between interviews and tax file:

\[ Y_{se} = \max(\text{Net income reported by the interviewees or imputed*}, \text{After tax income}) \]

\[ H_0: \text{both data-sources underestimate the true value on self-employment income.} \]

\[ \Rightarrow \text{the measurement error is “minimised” by using the maximum value between the two data-sources.} \]

Note(*): multivariate sequential regression approach based on the information collected on individuals/households during the interview.
Adjusting for measurement errors: 1st case

1st case: **Under-reporting > Tax avoidance**

=> “After tax income” > Reported “Net income”

=> “After tax income” is closer to true “Disposable income”
Adjusting for measurement errors: 2\textsuperscript{nd} case

2\textsuperscript{nd} case: **Tax avoidance > Under-reporting**

=> Reported “Net income” > ”After tax income”

=> Reported “Net income is closer to true “Disposable income”

**Survey data**

**Administrative data**
Some specific problems to MR approach:

The definition of *taxable* self-employment income in Italy is quite complex. There is not a set of simple, wide-ranging rules to convert the accounting profits: normal income of professional, deduction of profits retained and reinvested in the business, etc.

Interviewers have to refer to "*money drawn out concept*" when they ask for self-employment income (no to taxable income):

self-employment activity leads to a revenue if the individual or her/his family has got from this activity an amount of money that has been used for personal/household expenses, save or investment
Reconciliation of employee income is based in the following decision rules:

- The survey if reported only in the questionnaire but not in tax data;
- The tax data if the interviewee declares an employee income in a tax statement but does not report this income in the questionnaire (cold-deck imputation that selects donors from tax file – record linkage);

Survey/tax source according to this decision rule:

- if \( \text{tax value} > \text{survey value} \Rightarrow \text{tax data} \)
- If \( (\text{tax value} + 10\% \times \text{tax value}) > \text{survey value} > \text{tax value} \Rightarrow \text{survey data} \)
- If \( 2(\text{tax value}) > \text{survey value} > (\text{tax value}) \times 110\% \Rightarrow \text{tax data} \)
- If \( \text{survey value} > 2(\text{tax value}) \Rightarrow \text{closeness respect to the predicted value estimated by a multiple regression (only questionnaire data)} \)

In this way it is possible include in the final record the undeclared income (pay hidden) and at the same time to avoid the inclusion of outliers (errors produced by interviewees that report annual employee income instead of monthly employee income).
Concluding remarks

Integration of Survey & Administrative data (ISAD)

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ accuracy</td>
<td>▪ relevance</td>
</tr>
<tr>
<td>▪ coherence</td>
<td>▪ timeliness</td>
</tr>
<tr>
<td>▪ completeness</td>
<td></td>
</tr>
<tr>
<td>▪ cost and response burden</td>
<td></td>
</tr>
<tr>
<td>(C&amp;RB) reduction</td>
<td></td>
</tr>
</tbody>
</table>
Concluding remarks

Advantages:

- **Accuracy**: ISAD reduce or remove selecting non-response and measurement errors (memory/telescoping effects)
- **Coherence**: is improved through the use of common definition, classification and conceptual frameworks
- **Completeness**: through ISAD is possible to increase the coverage rate on the target population
- **C&RB reduction**: by eliminating items from the data collection when administrative information is available
Concluding remarks

Drawbacks:

- **relevance**: items conform to administrative (not statistical) concepts few concessions to statistical needs

- **timeliness**: generally slow, constrains for timely releasing of administrative data because of bureaucracy, time-consuming cleaning-procedure carried out by data producers
Micro-simulation of gross income variables

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

- **Response burden reduction**: replacement of questionnaire items with administrative information

- **Timeliness**: a new release schedule will be applied (2020 edition?). Effort will be made to anticipate the transmission deadline of full dataset to Eurostat by 11 months (from November T+2 to December T+1), a revised file will be produced later. The administrative data acquisition program will be re-negotiated with data providers, new strategies will be designed (statistical models)

- **Coverage enhancements**: a massive use of social security’s database in order to cover utmost information on non-pension cash benefits and an exploitation of new fiscal databases in order to estimate the income from capital assets
References

consolin@istat.it

Thank you for your attention!