April 26, 2023, 2:00-3:30 pm (CEST)

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS IASS
WEBINAR SERIES

THE PRODUCTION OF MICRODATA ON HOUSEHOLD INCOME, CONSUMPTION AND WEALTH AT ISTAT: EXPERIENCES, METHODS, PERSPECTIVES

GABRIELLA DONATIELLO
ISTAT| ITALIAN NATIONAL INSTITUTE OF STATISTICS

Acknowledgments

- This presentation is part of the ISTAT Project for the production of microdata on household Income, Consumption and Wealth in Italy, for which I have been responsible since 2013.
- The Bank of Italy participated in the ISTAT project since 2018 as part of the joint research activities and international activities at Eurostat and the OECD.
- This presentation is therefore the result of the work of all the colleagues who contributed to the project that I would like to thank.

- *The views and opinions expressed in this presentation are those of the author and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics - Istat.*

Outline

- Production of official statistics on the joint distribution of income, consumption and wealth (ICW) at the micro level
- Use of micro-integration techniques of social surveys: pre-requisites and actions taken to fill the gaps in term of data requirements and methodological issues
- ISTAT experience for the production of a joint distribution of income and consumption and for the imputation of household wealth
- Data matching of complex sample surveys and Renssen's weight calibration approach
- Insights on quality assessment of the final estimates

Integrated statistics on income, consumption and wealth

- Multidimensional analysis of poverty and living conditions
- Indicators/parameters on the joint distribution of income, consumption and wealth - complementary indicators (e.g. wealth poverty)
- International standards for the production of integrated statistics
- - OECD ICW Framework 2013
- - UNECE Canberra Group Handbook 2011
- - Eurostat-OECD Expert Group on the Joint Distribution of Income, Consumption and Wealth at Micro Level (EG ICW)

Integrated statistics on income, consumption and wealth

- Integrated statistics on income, consumption and wealth are quite difficult to collect:

  ✓ **Include stock data (wealth held at any given time) and flow data (income, consumption, financial assets)**

  ✓ **NSIs' budget constraints for conducting new surveys**

  ✓ **significant reporting burden on respondents**

- Better exploitation of existing data sources:

  ✓**combining sample and administrative sources to obtain data on income or wealth (first best)**

  ✓ **using statistical matching (SM) techniques as an additional tool (second best)**

Statistical matching: some pre-conditions 1/2

Ex-ante harmonization of social surveys

- SM procedures require strong pre-requisites of coherence of different data sources (EU Statistics on Income and Living Conditions – EU-SILC, Household Budget Survey - HBS, Labour Force Statistics - LFS)
- From 2011, ISTAT carried out a deep process of standardization of concepts, statistical units and variables
- Especially SILC and HBS have been reconciled to harmonize:

  - ✓**demographic variables**
  - ✓ **household composition**
  - ✓ **family relationship**
  - ✓ **level of education**
  - ✓ **ILO labour status**
  - ✓**dwelling facilities**

Statistical matching: some pre-conditions 2/2

***Ex-ante* collection of data for micro integration purposes**

- New shared variables with stand-alone value and as “hooks” to facilitate the links across different data sources
- The Consumption & Wealth module in EU-SILC 2017 collects data on:
  - ✓ **food consumption (at home and outside home)**
  - ✓ **transport (public and private)**
  - ✓ **regular savings**
  - ✓ **value of main residence**
  - ✓ **value of second (more) residence(s)**
- Variables able to capture some aspects of consumption/wealth with high predictive value

Statistical matching framework

- Main goals:

  ✓ **Achieve a micro data file from different sources that do not contain the same units or the same unit identifier**

  ✓ **The object is to investigate the relationship between variables not jointly observed in a single data source**

- Model-based techniques that are able to get timely results with reduction of costs and response burden
- Several methodological issues involved in the matching process

Statistical matching framework

- In the basic SM framework, the surveys to integrate share a set of variables X, while the variable Y is observed only in A, and the variable Z is observed in B (D'Orazio *et al.,* 2006)

SURVEY A		SURVEY B
Variable	Common variables	Variable
Y	X	Z

- The aim is to explore the relationship between Y and Z

Statistical matching framework

- Data integration at micro level is not necessary in case of estimation of parameters (correlation coefficient between Y and Z; regression coefficients, contingency table)
- Data integration is necessary when the final goal is a fused or synthetic data set which contains all the variables (X,Y,Z)
- SM techniques are mainly based on methods developed for imputing missing values:
  - ✓ **parametric (e.g. regression imputation)**
  - ✓ **nonparametric (hot deck imputation)**
  - ✓ **mixed (methods based on predictive mean matching)**

Statistical matching framework

- Hot deck procedures (random hot deck or nearest neighbor) are frequently used
- The underlying assumption:
  - ✓ **(i) conditional independence assumption (CIA) of the target variables given the common variables (i.e. Y and Z are independent once conditioning on X variables)**
  - ✓ **(ii) the observations in the samples are independent and identically distributed (i.i.d.) (i.e. the sample is a simple random sample)**

Statistical matching framework

- The CIA is a very limiting and rarely holds in practice
- It can be avoided:
  - ✓ **with auxiliary information concerning the relationship between Y and Z (estimates of a correlation coefficient, additional data sources observing jointly the target variables)**
  - ✓ **approaching SM in terms of uncertainty**
- The i.i.d. assumption is difficult to be maintained in case of complex sample surveys involving two or more stages of selection of the sample units

Statistical matching framework

Uncertainty approach to SM

- From the available datasets it is possible to estimate just the pairwise distributions (X,Y) and (X,Z), with the same X marginal distribution (Donatiello *et al.*, 2016)

- All the parameters on the (Y,Z|X) distribution are uncertain. This uncertainty reflects in a set of equally plausible estimates for the (Y,Z|X) distribution

- When the X, Y and Z are categorical, the probability of the (Y,Z|X) contingency table cell $\boldsymbol{\theta}\_(\boldsymbol{yz}|\boldsymbol{x})$ lies in the interval

$$\boldsymbol{max}\left(\boldsymbol{\theta}_{\boldsymbol{y}.|\boldsymbol{x}} + \boldsymbol{\theta}_{.\boldsymbol{z}|\boldsymbol{x}} - \mathbf{1}; \mathbf{0}\right) \leq \boldsymbol{\theta}_{\boldsymbol{yz}|\boldsymbol{x}} \leq \boldsymbol{min}\left(\boldsymbol{\theta}_{\boldsymbol{y}.|\boldsymbol{x}}; \boldsymbol{\theta}_{.\boldsymbol{z}|\boldsymbol{x}}\right)$$ (the so called Fréchet bounds)

Statistical matching framework

SM of data from complex sample surveys

- SM methods that explicitly take into account the sampling design and the sampling weights are:

  ✓**(a) Renssen's approach based on calibrations of the weights** (Renssen, 1998)
  ✓**(b) Rubin's file concatenation** (Rubin, 1986)

- The Renssen's approach seems more suitable when the interest variables (X, Y and Z) are categorical and it consists of two steps:

1. **harmonization of the distribution of the matching variables in the two data sources**
2. **estimation of the contingency table under the CIA or by exploiting auxiliary information**

Statistical matching at ISTAT

FIRST PHASE

- Provide a synthetic data set containing joint information on household income and consumption in Italy
- Imputation of total consumption of HBS (donor) into SILC (recipient) for the years 2012, 2013, 2014 and 2016
- Goals:

  ✓ **apply the most suitable SM methods to produce fused data**

  ✓**analyze the propensity to consume, to save, the material deprivation for sub-groups of population**

- Matching exercises were performed in R using the **StatMatch** package (D'Orazio, 2022)

Statistical matching at ISTAT 2/2

SECOND PHASE

- Enhance the fused SILC/HBS dataset with wealth information, by imputing assets and debts from Bank of Italy's survey
- Imputation of net wealth from SHIW 2016 (donor) into SILC Fused 2017 (recipient)
- Goals:

  ✓**study asset-based poverty and wealth inequality**

  ✓**measure the ability of households to support their living standards and cope with important economic shocks**

- R package StatMatch is used

Statistical matching main steps

- The integration process consists of several steps (D'Orazio *et al.,* 2006):
  - ✓ **(i) preliminary analysis of data sets**
  - ✓ **(ii) harmonization and reconciliation of data sets**
  - ✓ **(iii) selection of matching variables**
  - ✓ **(iv) selection and application of matching methods**
  - ✓ **(v) evaluation of the accuracy of the final estimates**

(i) Preliminary analysis of EU-SILC 2013 - HBS 2012 1/3

- Some critical factors in matching income and consumption
- Income and consumption are complex concepts:
  - ✓ **usually collected with specific surveys**
  - ✓ **different data collection techniques (diary)**
- Some surveys that look at income and consumption at the same time:
  - ✓ **Bank of Italy Survey on Household Income and Wealth - SHIW (HFCS of the ECB)**
  - ✓ **Israel (consumption survey)**
  - ✓ **Canada (Financial Security Survey)**
  - ✓ **Denmark (HBS expanded)**
- **In this type of survey, data on income and consumption have different levels of accuracy**

(i) Preliminary analysis of EU-SILC 2013 - HBS 2012 2/3

- In both surveys: the reference population is made up of households residing in Italy (Donatiello *et al.*, 2014)
- The sampling design is stratified in two stages municipalities-households, with stratification of municipalities based on demographic size

	HOUSEHOLDS		PERSONS	
	HBS	**EU-SILC**	**HBS**	**EU-SILC**
Sample Size	22.933	18.487	56.118	44.622
Population*	25.383.757	25.486.842	60.450.044	60.568.804

* Estimates with new inter-census weights from 2011

- The differences in the total population are due to the different reference periods used (quarterly and annual data for HBS, 31 December of the income reference year for SILC)

(i) Preliminary analysis of EU-SILC 2013 - HBS 2012 3/3

- There are many common socio-demographic variables
- Both surveys present partial information on the target variables: in HBS there is a section on income and in SILC the housing expenses are collected in detail

<table>
<tr><th></th><th>CONSUMPTION</th><th>SOCIO-DEMOGRAPHIC VARIABLES</th><th>INCOME</th></tr>
<tr><td>HBS</td><td colspan="2">COLLECTED</td><td>PARTIAL</td></tr>
<tr><td>SILC</td><td>PARTIAL</td><td colspan="2">COLLECTED</td></tr>
</table>

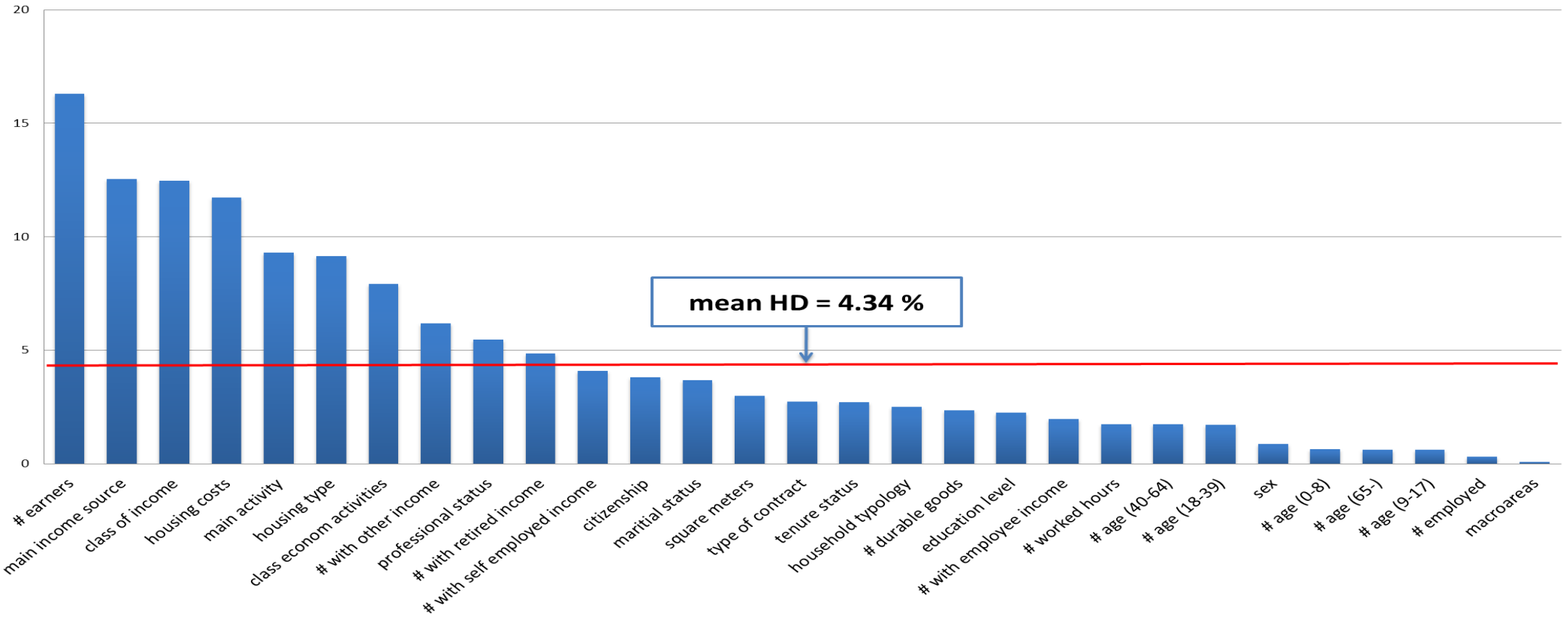
(ii) Harmonization and Reconciliation SILC 2013 - HBS 2012 1/2

- The high number of common variables required an intense phase of recoding of those variables that have a different definition in the two surveys

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

(ii) Harmonization and Reconciliation SILC 2013 - HBS 2012 2/2

- The application of empirical rules such as the Hellinger Distance (HD) has shown a good level of comparability of the harmonized variables. Variables whose relative cell frequencies differ by more than 5% are excluded from the subset of potential matching variables



(iii) Selection of matching variables SILC 2013 - HBS 2012 1/2

- Choice of the subset of matching variables
- Variables must generally meet two minimum criteria:
  - ✓**present homogeneous distributions in the two surveys**
  - ✓**act as good predictors of the target variables**
- Predictor selection methods:
  - ✓**OLS regression with stepwise selection method / GLM / Logistic regression weighted data**
  - ✓**Classification trees / Random Forest method / Conditional inference trees**

(iii) Selection of matching variables SILC 2013 - HBS 2012 2/2

- In the selection methods, the following were considered as response variables:
  - ✓ **for HBS, the logarithm of the average monthly expense**
  - ✓ **for EU-SILC, the logarithm of monthly income**

- Predictors with greater explanatory power:
  - ✓ **Consumption: number of durable goods, macro areas, level of education, number of employed persons, marital status**
  - ✓ **Income: number of durable goods, macro areas, level of education, sqm**

(iv) Selection and application of statistical matching methods

1. Non-parametric imputation (random hot deck) under CIA of HBS consumption classes in SILC
2. Uncertainty analysis
3. Exploitation of the available information to avoid CIA and/or to improve the estimates produced with consolidated SM techniques
4. Data matching of complex sample surveys:

   ✓**application of the Renssen's weight calibration procedure with:**

   (i) categorical variables

   (ii) continuous variables

(iv) 1. Non-parametric imputation (random hot deck) under CIA 1/2

- **Hot deck procedures** allow to impute missing values in the receiving data set using the other data set as a donor:
- Donation is based on the common variables
- Each record is imputed with a class of donors randomly extracted from a subset (number of durable goods and macro areas)
- Procedure suitable for categorical variables

(iv) 1. Non-parametric imputation (random hot deck) under CIA 2/2

- When the income-consumption matching operates on the basis of the common variables (X), independence between income and consumption is implicitly assumed conditionally on X
- This means that the Xs are able to fully explain the income-consumption relationship. This is an assumption that is not always valid
- Can the CIA represent an appropriate model for income-consumption matching?

(iv) 2. Uncertainty analysis

- The CIA can be “relaxed”:

  ✓**using an ancillary data source in which the income-consumption relationship is observed**

  ✓**addressing the SM problem in terms of uncertainty, due to the estimation of the parameters of the joint distribution of variables that are not observed jointly**

  ✓**by analyzing the uncertainty, a punctual estimate of the probability of interest is not obtained, but rather a range of plausible values: the higher the correlation between the common variables and the target variables, the narrower the range and the lower the uncertainty**

(iv) 2. Uncertainty analysis

- In the case of SILC and HBS it is possible to apply SM methods under the CIA if at least **one matching variable (e.g. income) is highly correlated with the target variable (consumption)**
- If there is perfect correlation, the target variable can be estimated with a simple regression function
- By relaxing the hypothesis of perfect correlation, it can be assumed that the CIA between the target variable and the matching variable is a non-unrealistic hypothesis
- We analyzed the available information to obtain a matching variable highly correlated with the target variable

(iv) 3. Exploitation of available information 1/5

- «Income and Savings» section present in the Italian HBS
- Data that is not disclosed to users and that we used experimentally
- This section contains information relating to:
  - ✓ **Number of income or pension recipients in the household**
  - ✓ **Average monthly household income in classes**
  - ✓ **Use of income in the year (spends everything on consumption or saves part of it)**
  - ✓ **Possible annual savings (punctual or in classes)**

(iv) 3. Exploitation of available information 2/5

- We corrected the HBS household income distribution using data from the ad hoc section (Donatiello *et al.*, 2014)
- The difference between the declared income class and the consumption class was analysed
- Only for households with an income class lower than the consumption class and who declare savings, a new income variable was calculated as the sum of consumption and savings
- Underlying assumption: in HBS the consumption class is more reliable than the reported income class

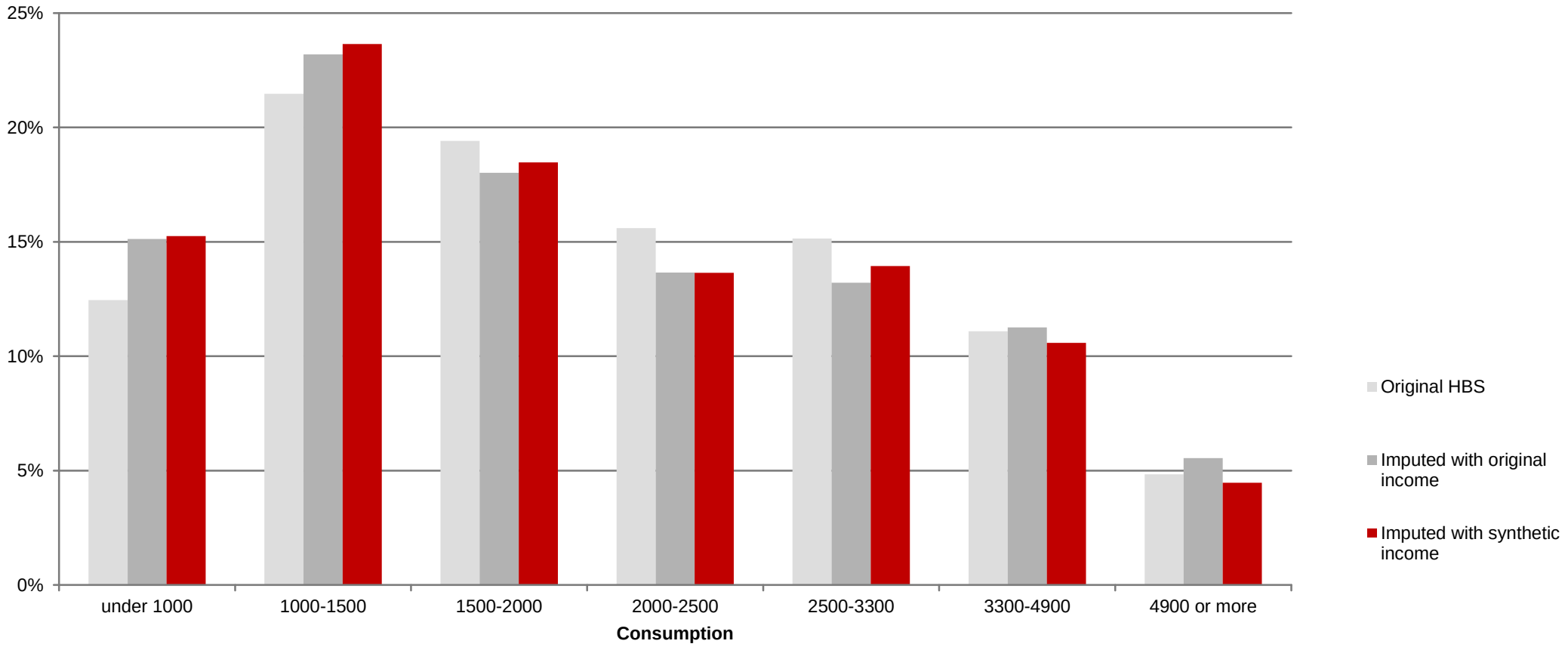
(iv) 3. Exploitation of available information 3/5

- The new HBS income reduced the underestimation of HBS compared to SILC especially in the extreme classes. It was used:
  - ✓ **as a matching variable**
  - ✓ **to narrow down the set of potential donors**

- HBS donor subpopulation for Random Hot Deck, household in:
  - ✓ **same macro areas**
  - ✓ **same number of durable goods**
  - ✓ **same class or adjacent income class**

(iv) 3. Exploitation of available information

- Comparison of consumption classes imputed in SILC with observed and synthetic HBS income



(iv) 3. Exploitation of available information 5/5

- ISTAT tested on a voluntary basis the C&W module 2017, as part of the revision of EU-SILC legal basis within the new European Framework Regulation on Social Statistics (IESS)
- New variables less accurate than those collected in specific surveys but important in the matching process
- The C&W module enabled us to test the efficacy of the use of proxy variables of the targets as matching variables capable of justifying the CIA

✓**C&W "partial consumption" as proxy of total consumption**

✓**C&W "value of main residence" as proxy of wealth**

(iv) 3. Some positive spin-offs

- At national level, we suggested to our colleagues in the HBS Unit to reintroduce the "income and savings" section in the new Consumption Expenditures survey (2014), which had been resized during the restructuring phase

- At European level, we recommended the introduction of a few variables on the type of use of income (consumption and savings) in HBS useful for:

  - ✓ **improving quality of income variable**
  - ✓ **income-consumption integration purposes**

(iv) 4. Renssen's Weight Calibration Procedure

- Many of the SM methods have been developed to supplement simple random sample (i.i.d.)
- The data matching of complex sample surveys presents a further difficulty due to the treatment and harmonization of sample weights
- The Renssen's procedure is one of the few methods that explicitly take into account the sampling design and the relative weights

(iv) 4. Renssen’s Weight Calibration Procedure 2/7

- We applied Renssen's method to categorical variables (Donatiello *et al.,* 2016)
- The weights in data sets are calibrated so as to harmonize the marginal distributions of the matching variables (macro areas, number of durable goods, consumption classes)
- The income/consumption joint distribution is obtained by estimating the income/consumption contingency table under CIA with linear models
- The consumption classes are imputed in SILC using weights and harmonized variables

(iv) 4. Renssen's Weight Calibration Procedure 3/7

- PROS:

  - ✓ starts from available data and weights
  - ✓ harmonizes marginal (joint) distributions of the matching variables
  - ✓ provides a synthetic data set that preserves the marginal distribution of the imputed variables and the joint distribution of the matching variables
  - ✓allows to easily introduce auxiliary data sources

- CONS

  - ✓negative probabilities
  - ✓heteroskedasticity and non-normally distributed residuals

(iv) 4. Renssen's Weight Calibration Procedure

- Imputation of consumption classes

  ✓ **The procedure makes it possible to estimate the probability of belonging to a consumption class for each household in HBS**

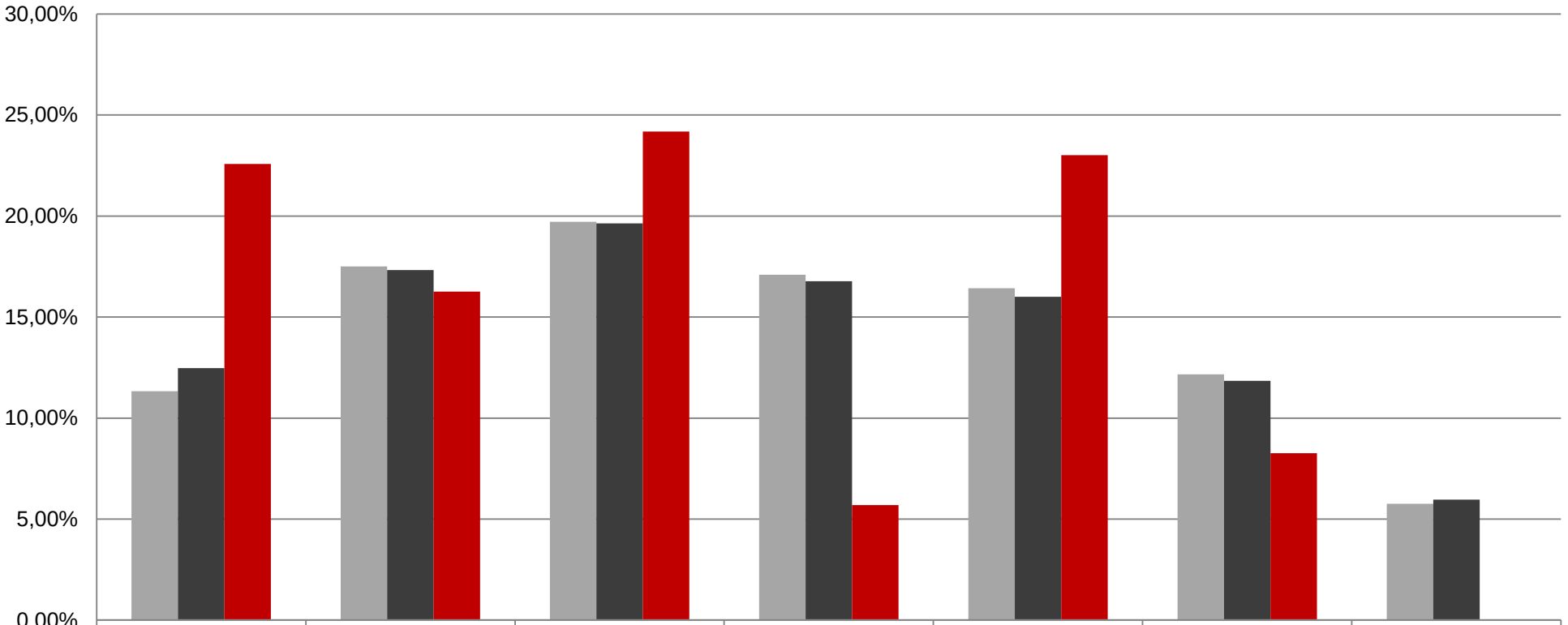
  ✓ **The probabilities are then used to provide an estimate of the consumption classes in SILC, we applied 2 methods:**

  (a) imputation of the class of consumption which corresponds to the highest probability

  (b) random selection of the consumption class starting from the estimated probabilities

(iv) 4. Renssen's Weight Calibration Procedure

- Comparison of consumption classes observed and imputed with random prediction and with selection of the highest probability



	Under 1000	1000-1500	1500-2000	2000-2600	2600-3300	3300-4900	4900 or more
HBS	11,34%	17,51%	19,71%	17,09%	16,43%	12,17%	5,76%
IMPUTED AT RANDOM	12,48%	17,32%	19,63%	16,77%	16,00%	11,84%	5,96%
IMPUTED WITH MAX	22,58%	16,26%	24,18%	5,70%	23,02%	8,26%	0,00%

(iv) 4. Renssen's Weight Calibration Procedure 6/7

- Imputation of HBS consumption 2016 in EU-SILC 2017 (income reference year 2016)
- We **Modified Renssen's Weight Calibration Procedure** to impute at the micro level a continuous variable to overcome some drawbacks of categorical variables (Donatiello *et al.*, 2022)
- MIXED TWO-STEP PROCEDURE

  ✓ **a consumption value is estimated in both SILC and HBS with a linear model with harmonized weights**

  ✓ **the observed HBS consumption value with the Nearest Neighbor Distance Hotdeck is imputed into SILC and the distance is calculated considering the predicted consumption value in both data sets**

  ✓**matching variables: macro areas (5), number of durable goods (4) consumption proxy (observed in the C&W module of SILC; reconstructed in HBS)**

(iv) 4. Renssen's Weight Calibration Procedure 7/7

- The consumption proxy as a matching variable:

  ✓ **guarantees the validity of the assumption of independence between income and consumption conditionally on the matching variables which is the basis of the method used**

- The modified Renssen's Weight Calibration Procedure:

  ✓ **uses the survey weights (after re-calibration to harmonize the distribution of the matching variables)**

  ✓ **returns an imputed variable (imputed consumption in SILC) whose marginal distribution is practically identical to that observed in the donor (HBS) after the weight harmonization**

(v) Evaluation of the accuracy of the final estimates 1/7

- The quality assessment of the matching results is a very complex step as it involves some critical issues in measuring the accuracy and reliability of the final estimates (Donatiello *et al.,* 2022)
- It is not possible to assess the accuracy of the imputed variable obtained at the end of the SM procedure by estimating the MSE or just the variance
- An evaluation of the whole matching process would be required but it depends on:

  - ✓ **quality and coherence of the data sources**
  - ✓ **the explanatory power of the common variables**
  - ✓ **characteristics of the algorithm used for the matching**
  - ✓ **method used to create the synthetic data set (correlation/regression coefficients, contingency tables, etc.)**

- An indirect partial assessment of the variability associated to the final estimates can be obtained through approaches based on assessment of SM uncertainty (Conti *et al.* 2012; Zhang, 2015)

(v) Evaluation of the accuracy of the final estimates

- There are several methods proposed in the literature for assessing the quality of estimates
- Rässler (2002) suggests a multi-level procedure:

1. **ability to reproduce the values of the input variables (Z) in the receiving file**
2. **ability to reproduce the joint distribution of common variables (X) and target variables (Y,Z)**
3. **ability to keep unchanged the correlation/association structure of the joint distribution (X, Y, Z) and for the marginal distributions of X-Y and X-Z**
4. **ability to preserve the marginal distributions of Z and of X-Z**

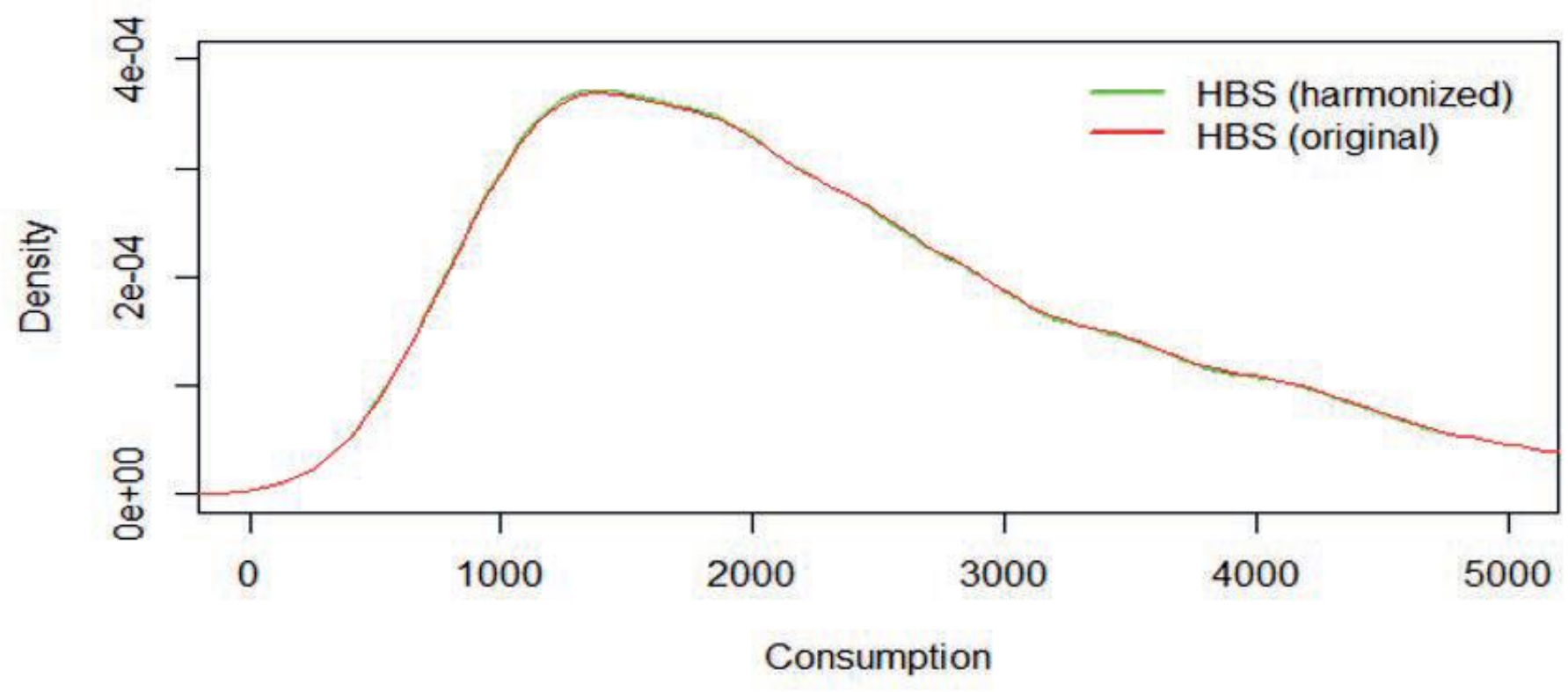
- The first 3 points can be verified through simulation studies, while the ability to preserve marginal distributions of the imputed variables can be verified using descriptive measures and statistical tests

(v) Evaluation of the accuracy of the final estimates 3/7

- In our application, the synthetic data set is the outcome of a complex SM procedure, whose first step (calibration) modifies the starting survey weights with the aim of harmonizing the marginal distribution of the matching variables
- A first check consists in assessing whether the initial calibration of the survey weights introduces significant changes in the marginal distributions of the target variables
- As a result, the estimated distributions of both the target variables remain almost unchanged considering both original and modified weights (Donatiello *et al.*, 2022)

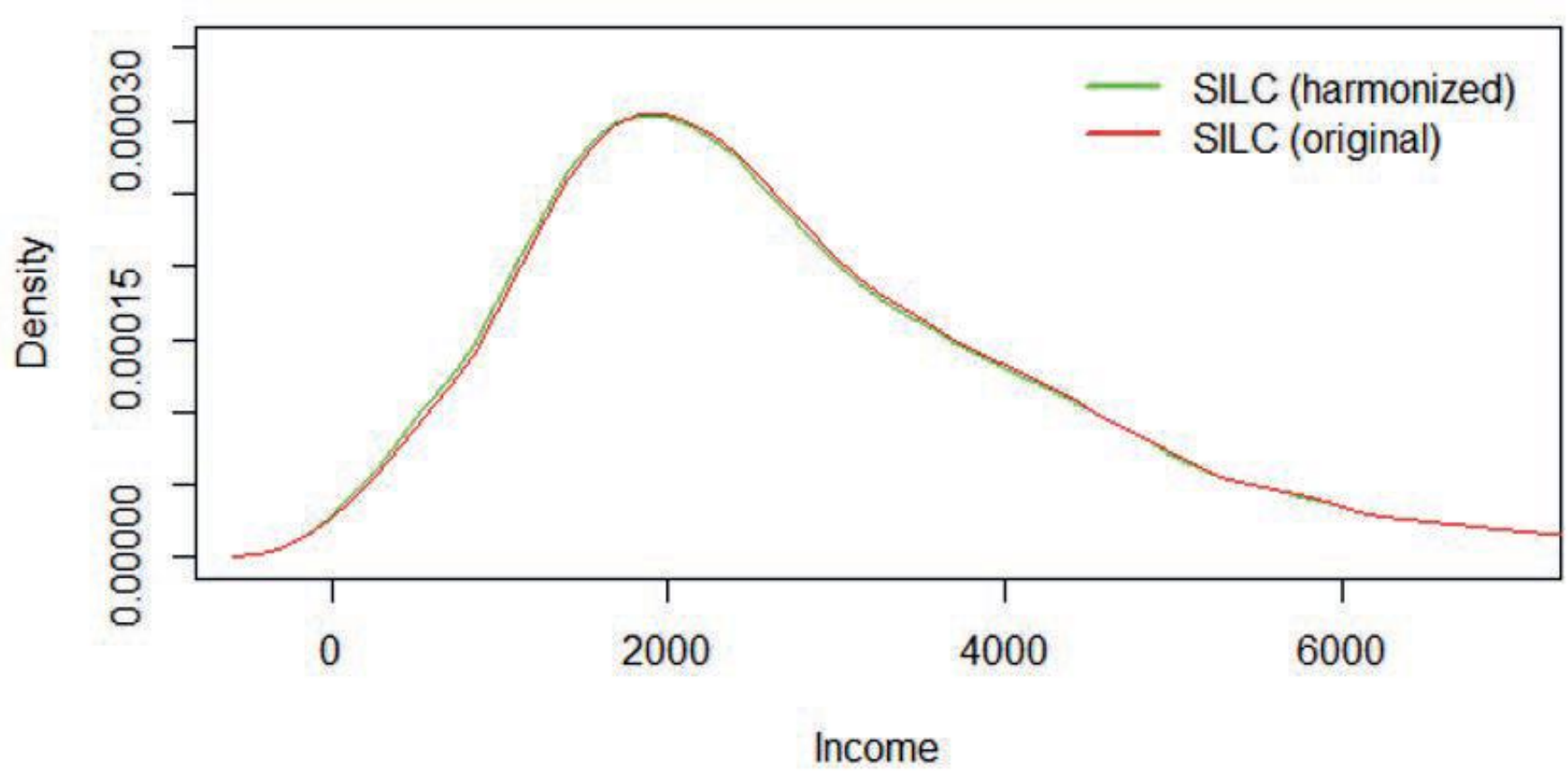
(v) Evaluation of the accuracy of the final estimates

Comparison of total consumption before and after the initial harmonisation step (HBS 2016)



(v) Evaluation of the accuracy of the final estimates

Comparison of total income before and after the initial harmonisation step (EU-SILC 2017)



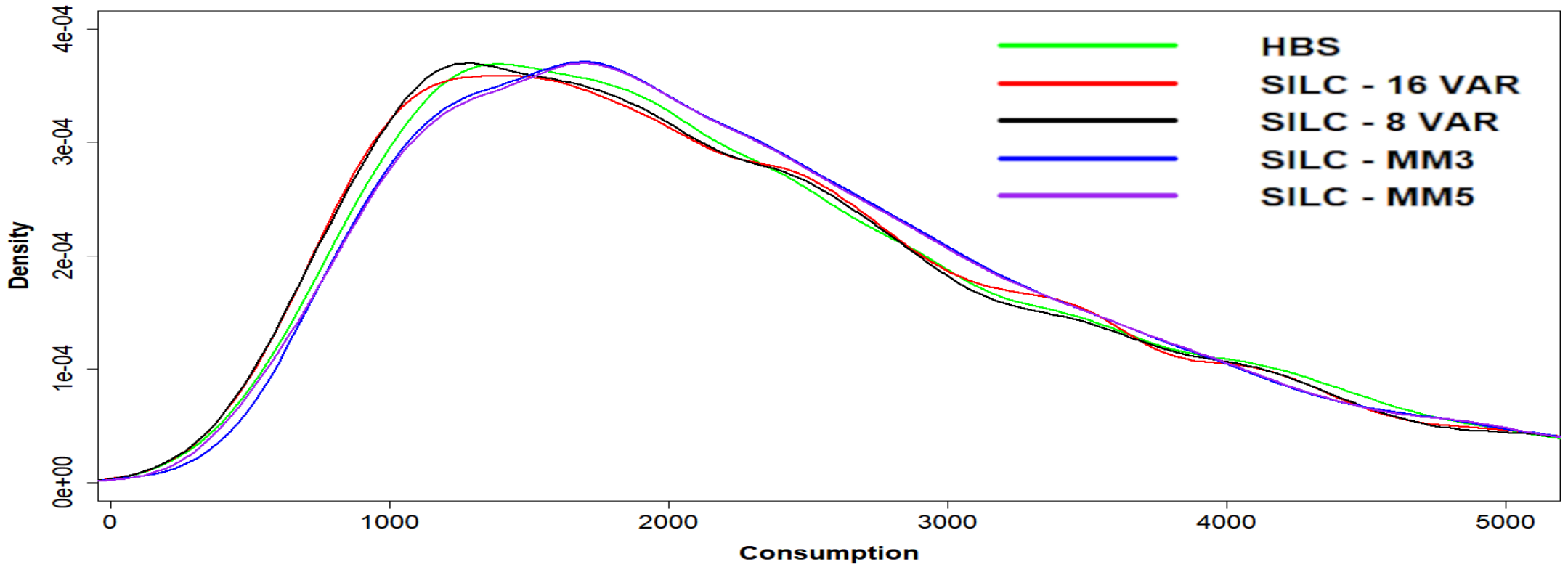
(v) Evaluation of the accuracy of the final estimates

- Poverty indicators by type of weight before and after the harmonization step SILC Fused 2017 (Percentage values)

	Survey weights	Weights after harmonisation	
	HBS	HBS	SILC FUSED
Relative poverty	10.60	10.64	10.82
Absolute poverty	6.28	6.30	6.12

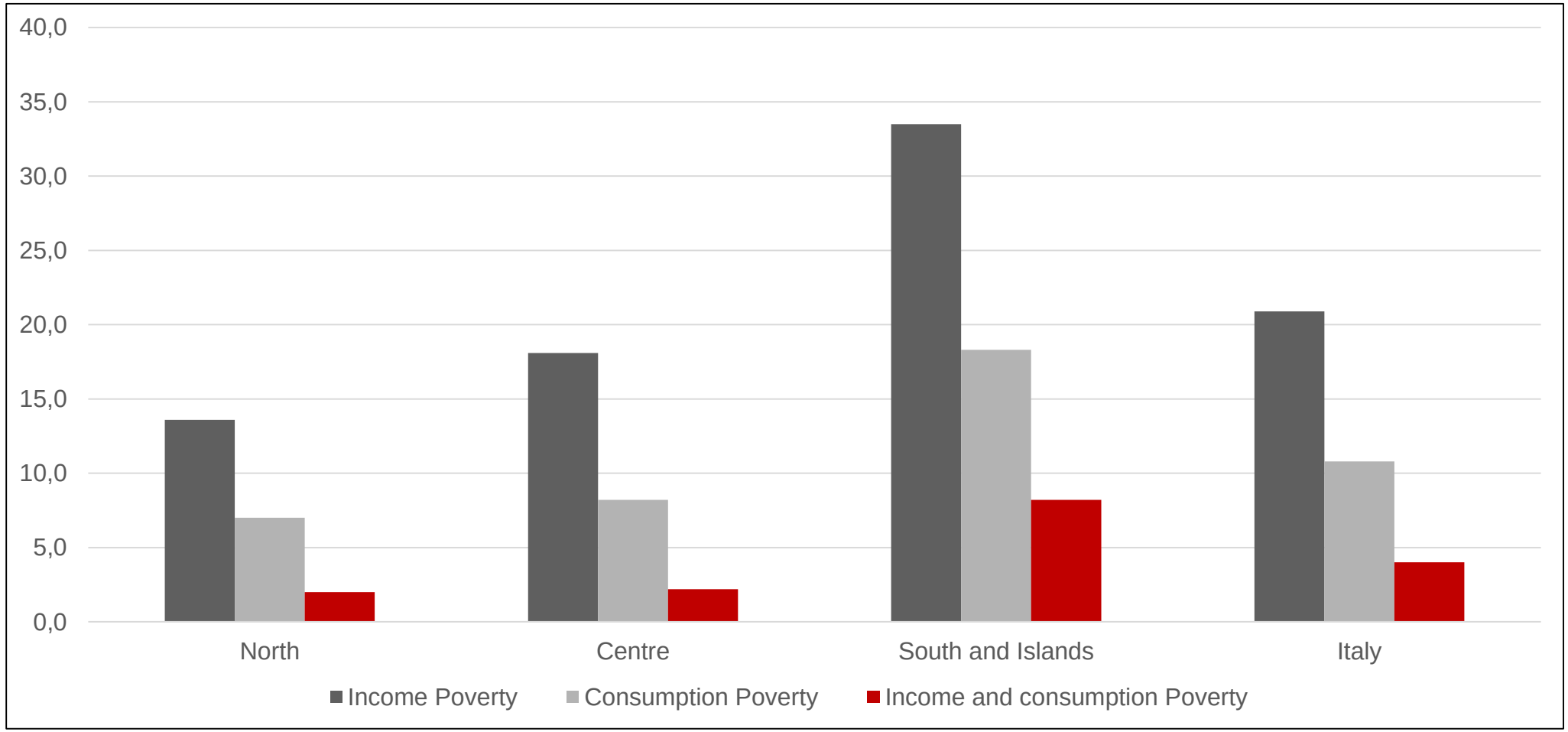
(v) Evaluation of the accuracy of the final estimates 7/7

- Comparison of total consumption of HBS and imputed consumption in SILC with our method (distance function with 8 or 16 dummies) and standard methods (MM3 and MM5)



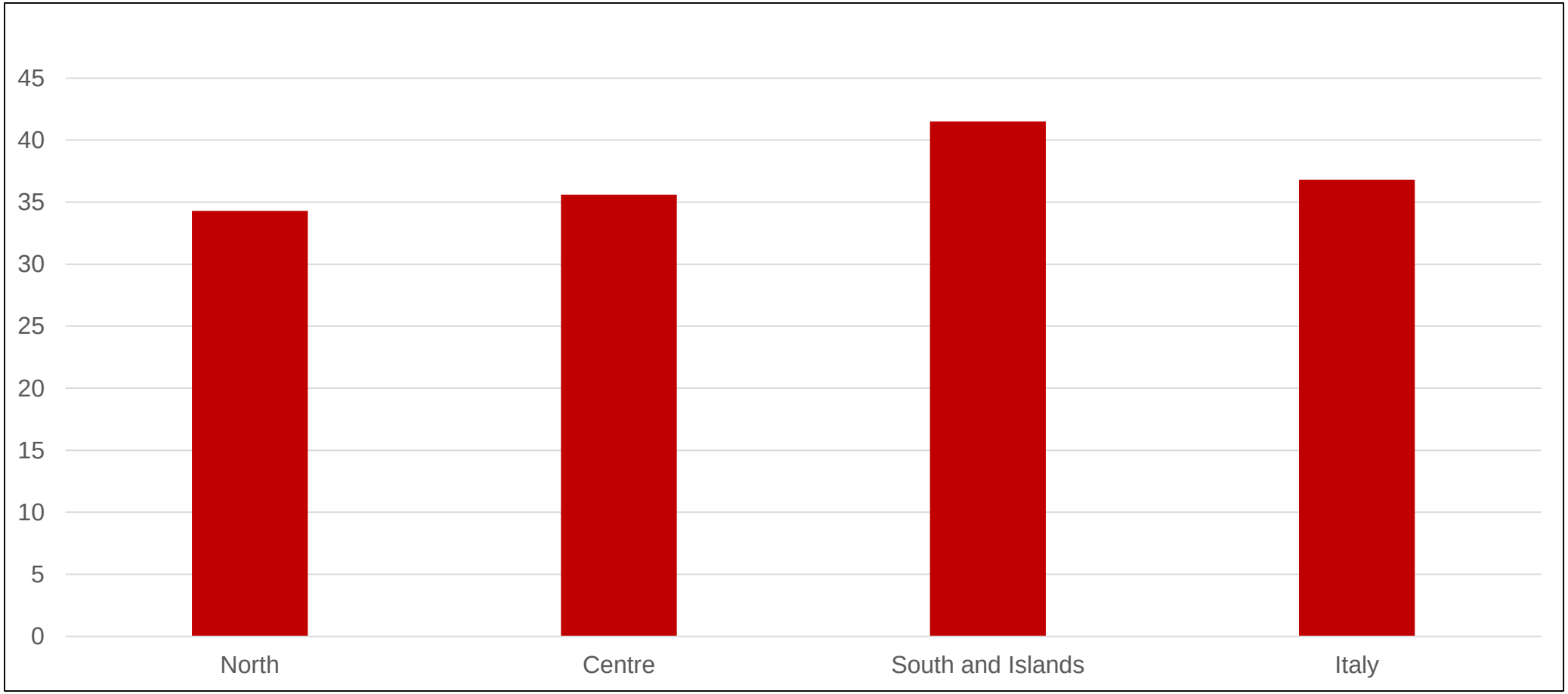
Joint distribution of income and consumption 2016 1/3

- Relative poverty by type and macro areas SILC FUSED 2017 (Percentage values)



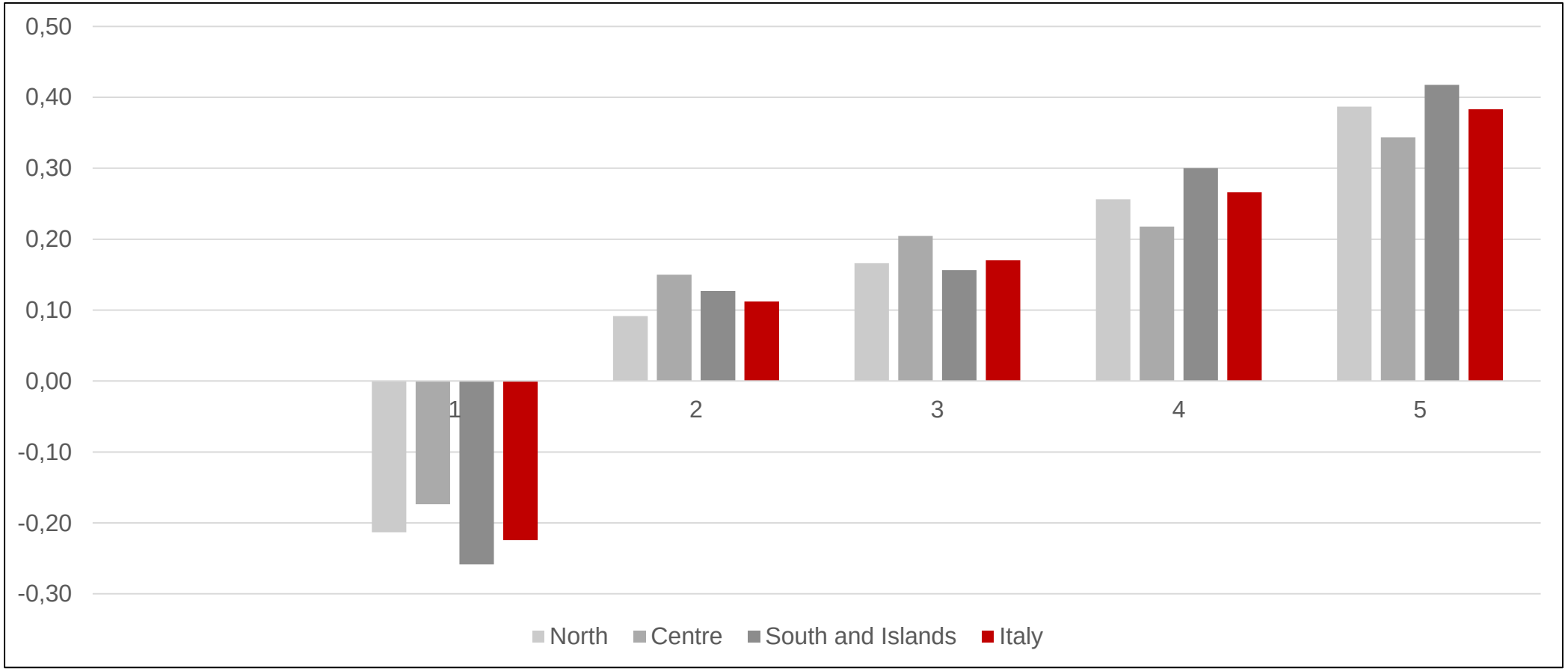
Joint distribution of income and consumption 2016 2/3

- Households with income less than consumption by macro areas SILC FUSED 2017 (Percentage values)



Joint distribution of income and consumption 2016 3/3

- Median saving rate by income quintile and macro areas – SILC FUSED 2017 (Percentage values)



The wealth dimension

Bank of Italy Survey on Household Income and Wealth - SHIW

- Preliminary analysis of SHIW 2016 and EU-SILC 2017 (income reference year 2016) for checking the target population
- Both surveys are based on a two-stage sampling design
- Both surveys include a longitudinal component
  - ✓ **Rotational panel scheme for EU-SILC (4 years)**
  - ✓ **Split panel survey for SHIW (panel households interviewed on previous surveys, even since 1989)**
- Different weighting procedures
  - ✓**Cross sectional and longitudinal weights for EU-SILC**
  - ✓ **Final weights including attrition-adjusted weights for panel units in SHIW**

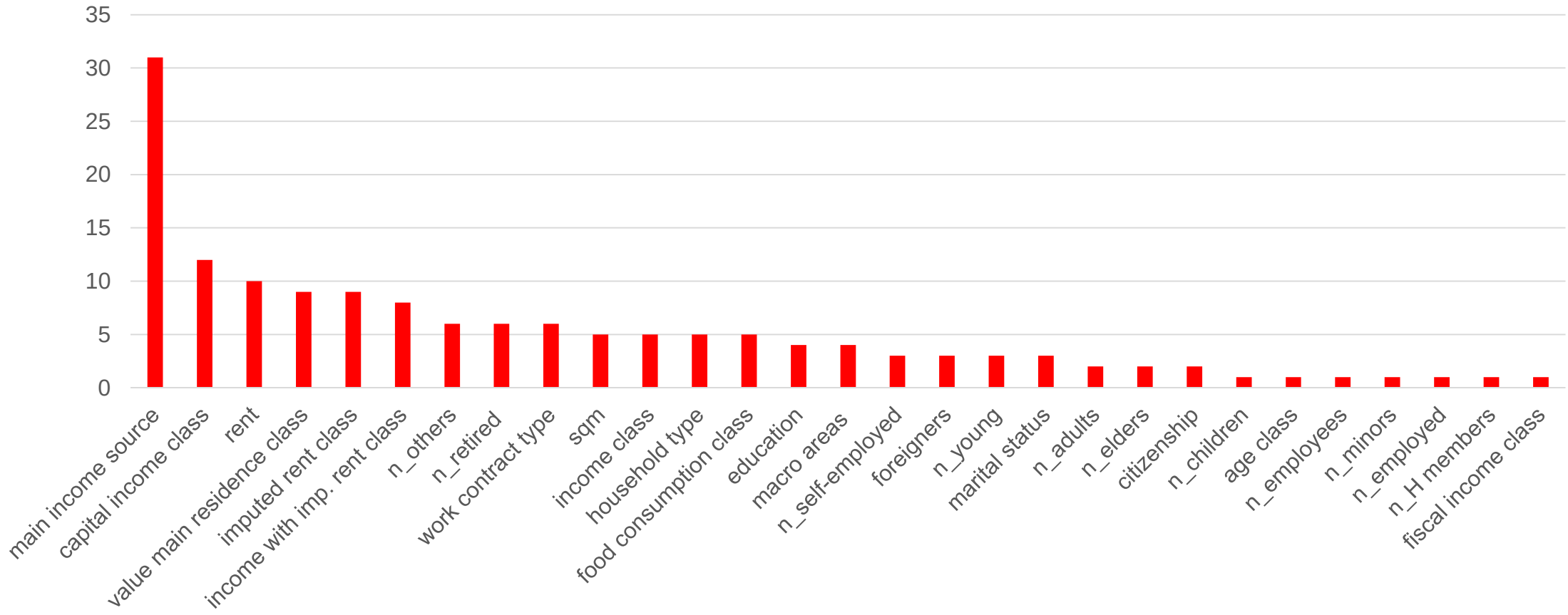
The wealth dimension

- Reconciliation of the definitions and classifications of the shared variables of SILC and SHIW 2016

	COMMON VARIABLES
HOUSEHOLD REFERENCE PERSON	**Sex, Marital status, Age, Educational level attained, Main activity status, Main income source**
HOUSEHOLD STRUCTURE	**Number of children (0-8), Underage people (9-17), Younger people (18-39), Adults (40-64), Elderly people (65-)**
INCOME	**Number of employed people, Number of income earners, Employee income, Self-employment income, Capital income, Pension income, Income Quintiles, At-risk of poverty rate, Ability to make ends meet**
HOUSING CONDITIONS	**Macro areas, Square meters, Tenure status, Imputed rent**
HOUSING-RELATED EXPENSES	**Rent, Mortgage, Utility bills, Arrears on utility bills, Rent arrears, Arrears on loan payments**
	CONSUMPTION AND WEALTH VARIABLES
CONSUMPTION	**Food at home, Food outside home**
WEALTH	**Value of main residence, Real and financial capital income**

The wealth dimension

- The Hellinger Distance (HD) for analyzing the dissimilarity of the estimated distributions shows a quite good level of comparability of harmonized variables

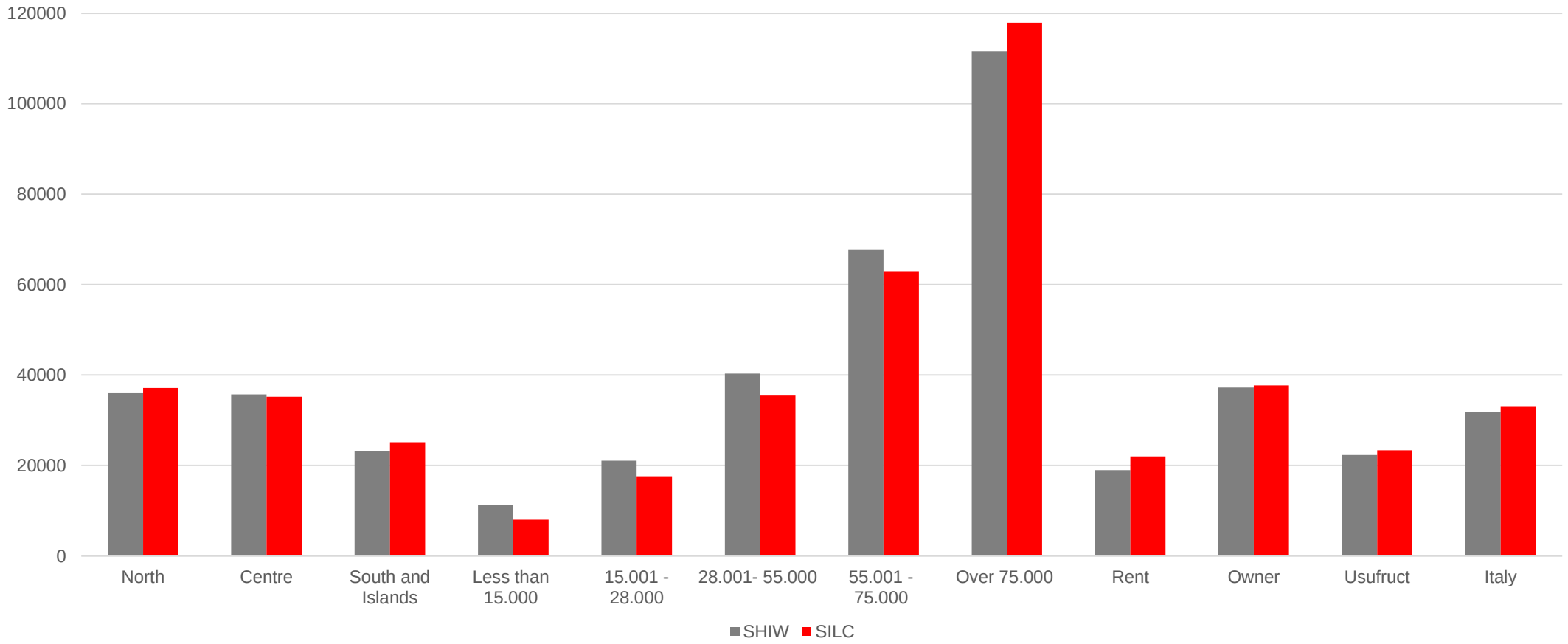


The wealth dimension

- We are currently imputing net wealth from SHIW into SILC FUSED
- This is a new and rather complex exercise with methodological challenges
- Activity that can benefit from the experiences accumulated at ISTAT and Bank of Italy
- To have a perfectly comparable common variable between SILC and SHIW, fiscal income from ADMIN sources was added using record linkage
- SM Method: Modified Renssen's Weight Calibration Procedure without discretizing continuous variables
- Matching variables: tenure status and fiscal income

The wealth dimension

- Comparison of gross fiscal income by macro areas, income brackets and tenure status - Year 2016 (Average in euros)



The wealth dimension

- Preliminary results of net wealth imputation into SILC
- Matching variables: tenure status and fiscal income

	MIN.	1ST QU.	MEDIAN	MEAN	3RD QU.	MAX.	SD	CV	IQR
SHIW with SHIW weights	1	15.000	125.757	206.608	252.768	8.492.227	343.793	1,6640	237.768
SHIW with harmonized weights	1	15.275	125.761	206.338	252.737	8.492.227	342.282	1,6588	237.462
Imputed in SILC with SILC's harmonized weights	1	17.000	137.616	217.913	265.000	8.492.227	351.942	1,6151	248.000

The wealth dimension

- Preliminary results of net wealth imputation into SILC
- Matching variables: fiscal income, tenure status, sqm, household type

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	CV	IQR
SHIW with SHIW weights	1	15.000	125.757	206.608	252.768	8.492.227	343.793	1,6640	237.768
SHIW with harmonized weights	1	15.000	124.388	200.598	250.000	8.492.227	334.309	1,6666	235.000
Imputed in SILC with SILC's harmonized weights	1	16.865	129.000	205.584	250.000	8.492.227	355.908	1,7312	233.135

Concluding remarks

- The production of statistics on the joint distribution of income, consumption and wealth at the micro level continues to present many challenges
- The quality assessment of the final estimates produced with statistical matching techniques is rather complex as based on strong assumptions
- The synthetic data produced are still experimental and it is important informing external users on the model applied and on the validity and plausibility of the final estimates
- Next goals: apply the SM procedures to year 2020 to analyse the impact of the pandemic on the resilience and saving capacity of households examining joint information on income, consumption and the role of wealth in mitigating the effects of the crisis

References

Bank of Italy (2018), Survey on Household Income and Wealth – 2016, https://www.bancaditalia.it/pubblicazioni/indagine-famiglie/bil-fam2016/index.html

Conti, P.L., D. Marella, and M. Scanu. 2012. “Uncertainty Analysis in Statistical Matching”. Journal of Official Statistics - JOS, Volume 28, N. 1: 69-88.

Donatiello G., M. D’Orazio, D. Frattarola, M. Spaziani. 2022. “The joint distribution of income and consumption in Italy: an in-depth analysis on statistical matching”. Rivista di Statistica Ufficiale, Review of Official Statistics n. 3, DOI**:** 10.1481/ISTATRIVISTASTATISTICAUFFICIALE_3.2022.03. https://www.istat.it/it/archivio/279980

Donatiello, G., M. D’Orazio, D. Frattarola, A. Rizzi, M. Scanu, and M. Spaziani. 2016. “The role of the conditional independence assumption in statistically matching income and consumption”. *Statistical Journal of the IAOS*, Volume 32, N. 4: 667-675. http://content.iospress.com/articles/statistical-journal-of-the-iaos/sji1000

Donatiello, G., M. D’Orazio, D. Frattarola, A. Rizzi, M. Scanu, and M. Spaziani. 2014. “Statistical Matching of Income and Consumption Expenditures”. *International Journal of Economic Sciences*, Volume III, Issue 3: 50–65.

D’Orazio, M. 2022. “StatMatch: Statistical Matching or Data Fusion”. R package version 1.4.1. https://CRAN.R-project.org/package=StatMatch

D’Orazio, M., M. Di Zio, and M. Scanu. 2006. Statistical Matching: Theory and Practice. Chichester, UK: John Wiley & Sons.

OECD. 2013. OECD Framework for Statistics on the Distribution of Income, Consumption and Wealth. Paris, France: OECD Publishing.

Rässler, S. 2002. Statistical Matching. A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. Cham, Switzerland: Springer, Lecture Notes in Statistics.

Renssen, R.H. 1998. “Use of Statistical Matching Techniques in Calibration Estimation”. Survey Methodology, Volume 24, N. 2: 171-183.

Rubin, D.B. 1986. “Statistical matching with adjusted weights and multiple imputations”. Journal of Business and Economic Statistics, 4, 87-94

Zhang, Li-C. 2015. “On Proxy Variables and Categorical Data Fusion”. Journal of Official Statistics - JOS, Volume 31, N. 4: 783–807.

THANK YOU

GABRIELLA DONATIELLO| gabriella.donatiello@istat.it

Istat | Istituto Nazionale di Statistica