The joint distribution of income and consumption in Italy: an in-depth analysis on statistical matching

Gabriella Donatiello[1], Marcello D'Orazio[1], Doriana Frattarola[1], Mattia Spaziani[1]

Abstract

This article presents an application of statistical matching methods to integrate the EU Statistics on Income and Living Conditions and the Household Budget Survey with the aim of creating a synthetic dataset that can permit an in-depth multidimensional analysis of households' economic poverty in Italy. The work takes stock of previous experiences done at the Italian National Institute of Statistics - Istat and proposes a modification of a well-known approach to the statistical matching of data from complex sample surveys. The re-designed method permits to create a synthetic dataset that preserves the marginal distribution of both the target variables. The proposed method is more complex than simpler donor-imputation methods and permits taking into account the final survey weights. The higher complexity requires some additional checks when validating the results of the whole application. Preliminary results, presented in this paper, are quite promising also because the work benefits from an accurate ex ante harmonisation strategy of the reference surveys and on the collection of useful data for the application of statistical matching methods.

Keywords: Data fusion, data integration, weights' calibration.

DOI: 10.1481/ISTATRIVISTASTATISTICAUFFICIALE_3.2022.03

1 Gabriella Donatiello (donatiel@istat.it); Marcello D'Orazio (madorazi@istat.it); Doriana Frattarola (frattarola@istat.it); Mattia Spaziani (mspaziani@istat.it), Italian National Institute of Statistics – Istat.

This work is part of the Project for the production of microdata relating to household Income, Consumption and Wealth (ICW project) at national and international level.
Although this article is the result of all the authors' commitment, the Sections are attributed as following: 5, 6 and 7 to Gabriella Donatiello; 3 to Marcello D'Orazio; 1 and 2 to Doriana Frattarola; 4 to Mattia Spaziani.

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

The authors would like to thank the anonymous reviewers for their comments and suggestions, which enhanced the quality of this article.

1. Introduction

In the last decades, the growing demand to provide data for measuring household economic well-being at the micro level has encouraged the production of integrated statistics on household income, consumption and wealth. In this context, the re-design of the social statistics framework at European and national level facilitated the use of integration techniques to exploit all information from existing data sources, with reduction in terms of costs for National Statistical Institutes and response burden for households. The micro integration is commonly performed by applying *statistical matching* (SM, also known as *data fusion*) methods that consist in an *ex post* integration of data from surveys referred to the same target population. This type of integration exploits variables observed and shared by different data sources for producing statistics on the relationship between variables not jointly observed in a survey. Starting from 2013, at the Italian National Institute of statistics - Istat we have investigated different matching techniques in order to produce joint statistics on income and consumption by integrating EU Statistics on Income and Living Conditions (EU-SILC) data, (from year 2012, with income reference year 2011, to year 2017), and the Household Budget Survey (HBS) data (from 2011 to 2016).

Application of statistical matching techniques is conditional to a series of prerequisites and assumptions. As the integration is based on a suitable set of variables shared by the two data sources, a key prerequisite is their *ex ante* harmonisation. This task was made easy by efforts undertook at national level in order to overcome the core social variables' reconciliation process required at European level. In the Italian EU-SILC and HBS, both conducted by Istat, the re-design of the survey questionnaires was the occasion to harmonise definitions and classifications of many shared variables such as demographic ones, household composition, family relationship, level of education, ILO labour status, *etc*.

The main obstacle to a successful application of statistical matching is the holding of the assumption underlying most of the SM methods, *i.e.* the fact that the relationship between the target variables not jointly observed - income and consumption in our application - could be fully explained by a subset of common variables shared by both the data sources (the *matching variables*). This statement corresponds to assume the independence between

income and consumption conditional on the chosen subset of common variables. Validity of this conditional independence (CI) assumption usually cannot be tested on the available data source, (an additional dataset where all the variables are jointly observed is necessary) and unfortunately it is seldom valid in real world applications. In particular, subject matter experts exclude that income and consumption can be independent conditional on a subset of socio-demographic variables shared by HBS and EU-SILC surveys.

Donatiello *et al.* (2014*a*) performed a statistical matching of EU-SILC and HBS at micro level by applying a *random hotdeck* procedure to impute the observed values for classes of consumption observed in HBS (donor dataset 2011 data) into the EU-SILC survey (recipient dataset 2012 data with income reference year 2011). In that occasion, the CI assumption could not be verified from the matched datasets, but it was assumed to hold as the set of matching variables included a proxy of the income variable (the household monthly income that could be reconstructed in the HBS data and directly observed in EU-SILC) that permitted to improve the accuracy of the SM results.

Donatiello *et al.* (2015) went ahead in the matching of SILC and HBS, focussing on the Renssen's weights calibrations procedure (Renssen, 1998). This is one of the few SM methods that can manage the survey weights and ensure a higher coherence of matching outputs in maintaining the distributions of the considered variables (more details on this method will be given in Section 3). This is a very appealing feature when analysing the data from complex sample surveys, where all the results are derived considering the final survey weights that reflect the probabilistic mechanism used in selecting the sample as well as weights' correction to compensate for coverage or non-response errors. Some interesting results were presented at the EU-SILC best practice workshop held in London in 2015 (Donatiello *et al.,* 2015) and at ITACOSM conference in June 2017 (Donatiello *et al.*, 2017).

In this paper, we investigate all the advantages and disadvantages related to the use of Renssen's SM method by applying it to HBS 2016 data and EU-SILC 2017 data (income reference year 2016). The objective is twofold: first to show further progress in applying the standard methods and a proposal for a modification that facilitates integration at micro level and secondly to highlight the impact of recent improvement done in re-designing the surveys in order to facilitate their *a posteriori* integration. This is the main

lesson learnt from previous matching applications, where all the *a posteriori* exercises of integrating surveys showed that a successful application of SM requires thinking at the integration in the survey design phase.

This work is structured as follows: the next Section will provide a brief description of the surveys and their recent changes also in view of integrating the data. Section 3 gives some insights of the Renssen's SM methods and of the modification introduced here to integrate household surveys at micro level. Section 4 presents the results achieved in applying the method to match HBS and SILC data and highlights the fact that the Renssen's procedure requires additional checks if compared to a "standard" assessment of the results of a SM application. Section 5 shows first results of multidimensional economic poverty and Section 6 makes the point of arrival of our work with some hints on future perspectives. Finally, Section 7 will present the concluding remarks.

2. An ex ante collection of data for micro integration purposes: EU-SILC 2017 data

In Italy, HBS and EU-SILC are carried out by Istat and cover the same population of private households. Both sample surveys use a stratified two-stage probabilistic sampling design. Primary sampling units (PSU) are the municipalities and second stage units (SSU) are the households. Inside each administrative region (estimation domain corresponding also to the primary strata), the PSU are further stratified according to their demographic size and, in order to guarantee self-weighting design, the total of residents in each stratum is approximately constant.

The Italian EU-SILC and HBS show a large number of common variables, and in year 2014, HBS experienced some important methodological improvements aiming at fostering data comparability at European level; in addition, the re-design also involved variables whose definition and observation were aligned as much as possible with that of EU-SILC.

The initial application of SM techniques in Istat (Coli *et al.*, 2005) highlighted the importance of using relevant auxiliary information to make the CI assumption holding and consequently improve the estimation of correlation between target variables. As the 2014 exercise showed (Donatiello *et al*. 2014*a*; 2016*a* and 2016*b*), including a proxy of income (a rough information of household income in classes collected in HBS) in the set of the matching variables, improves markedly the results if compared to application where this set does not include a strong proxy of one of the target variables. It should be noted that the income of Italian HBS was used by us in an experimental way, as it was not disseminated to users due to the well-known difficulties in collecting income data in a consumption survey. Considering that HBS income was not fully reliable we decided to overcome the problem by carrying out a matching based on a modified random hotdeck procedure that allowed to select donors in the same income classes or in those immediately preceding and following that of the recipient unit. Furthermore, in order to have proxy variables for SM purposes, we decided to collect some consumption variables in SILC. For this goal, on a voluntary basis, Istat implemented and tested the rolling module on Consumption & Wealth in EU-SILC 2017, agreed as part of the revision of EU-SILC within the new Framework Regulation on Social

Statistics (IESS)[2]. EU-SILC collection of variables on food consumption and transport, jointly with the already available data on housing costs, was seen as a way to provide enough information to derive a "strong" proxy of the total consumption that could be used in the SM applications. The design of the EU-SILC "Consumption & Wealth" module took stock of our previous SM exercise (Donatiello *et al.*, 2014b), where a detailed analysis of the structure of Italian HBS permitted to identify those consumption components representing good predictors for total consumption. Food and transport, as well as housing costs, were identified as the most important predictors of total consumption expenditures and for this reason these components were finally included into the 2017 EU-SILC Consumption & Wealth module[3].

2.1 Consumption variables in EU-SILC

Consumption & Wealth module in EU-SILC 2017 collected five consumption target variables: food at home; food outside home; public transport; private transport; regular savings. It is worth noting that EU-SILC annually collects an important amount of consumption expenditures, the "Total Housing Costs" (target variable HH070). This variable includes the costs of utilities (water, electricity, gas and heating) and all kind of expenses connected with the household right to live in the accommodation (for owners and tenants it includes mortgage interest payments and rent payments).

Annual target variable on housing costs together with the new variables observed with the module on food consumption and transport represent an important part of the total household expenditures in HBS (Consolini *et al.,* 2018*a* and 2018*b*). These variables gave us enough information to derive a synthetic variable that can be considered a good proxy of the household's total consumption.

2 Regulation (EU) 2019/1700 of the European Parliament and of the Council of 10 October 2019 establishing a common framework for European statistics relating to persons and households, based on data at individual level collected from samples.

3 The Italian module also included several voluntary variables, not provided for in the European Regulation but functional to the SM, such as the use of the monthly income, namely if the household spends the overall income for consumption or if it saves a part or it reduces saving. In fact, the same question was collected by HBS and Bank of Italy's SHIW. In addition, another variable helped to understand how the family finances expenditures that exceed their monthly income.

In order to compare SILC consumption variables with HBS data, we have set-up a list of "derived" consumption variables in HBS, using the same components that we collect in the new SILC module, in addition to total housing costs.

In HBS, most of the components included in the variable HH070 are collected with the exception of the tax on the main residence ("Imu" and "Tasi" taxes in year 2016) and the mortgage interest payments. In order to have a comparable measure of "Total Housing Costs", a modified variable of HH070 is derived in SILC (excluding the costs not covered in HBS) and likewise in HBS.

Then, we have estimated the correlations (Spearman and Pearson on the logarithmic of the two variables) between partial and total consumption. In both cases, we obtained a value of 0.80, remarkably close to one, which confirms a strong correlation between partial and total consumption in HBS.

This harmonised "Total Housing Costs" variable, together with food and transport expenses variables, was used to construct a variable of observed "SILC consumption", which turned out to be a very good predictor of the total consumption (Table 2.1).

Table 2.1 – Comparison between SILC and HBS of partial and total consumption
(Values in euros)

		min	Q1	mean	median	Q3	max
Partial consumption	HBS	0	561	961	862	1237	5317
	SILC	3	593	1001	881	1274	10903
Total consumption	HBS	93	1407	2482	2115	3189	18179

Source: Istat HBS 2016 and EU-SILC 2017

In our first exercises the CI assumption could not be verified from the available data, now with 2016 data on income and "partial" consumption expenditures, both available in EU-SILC 2017, allows us to roughly test the validity of this assertion. In particular, we estimated the Spearman correlation between income and "partial" consumption, controlling for matching variables used in some previous SM exercises; the results presented in Table 2.2, show correlation coefficients well far from 0 and therefore confirm that the independence between income and consumption conditional to some relevant common variables does not hold. In other words, the results of a SM exercise

integrating income and consumption data cannot be considered reliable unless the subset of matching results includes a strong proxy variable of income or consumption (or both).

Table 2.2 – Correlation coefficient between income and partial consumption by matching variables

Macro areas	Durable goods	Correlation
North	up to 5	0.40
	6	0.40
	7	0.48
	8	0.45
Centre	up to 5	0.34
	6	0.46
	7	0.42
	8	0.47
South and Islands	up to 5	0.27
	6	0.40
	7	0.39
	8	0.45

Source: Istat EU-SILC 2017

3. Statistical matching of data from complex sample surveys

As already mentioned, SM methods aim at integrating two distinct data sources, *A* and *B*, referred to the same target population with the objective of exploring the relationship between variables, *Y* and *Z*, not jointly observed in one of the sources. Integration is based on the variables (*X*) shared by the two data sources, and in particular on a suitable subset ($X_M;\ X_M \subseteq X$) of predictors of both *Y* and *Z*, denoted as *matching variables*.

A large part of the SM methods was developed to integrate data originating from simple random samples, where the values of (*X,Y,Z*) are independent random outcomes of the same (unknown) model which describes the relationship between the variables; in other words the observations in the data sources are *independent and identically distributed* (i.i.d.). Unfortunately, the data collected from National Statistical Institutes often originate from complex probabilistic sample surveys carried out on the same finite population that involve multistage and stratified cluster sampling designs, where the i.i.d. assumption is no longer valid. In fact, cluster sampling introduces dependence between units belonging to the same cluster; in addition, complex sampling designs may determine unequal units' inclusion probability.

In literature, there are relatively few statistical matching methods explicitly tailored to handle data from complex sample surveys; one of the most promising is the method suggested by Renssen (1998), based on *weights' calibration*. Calibration is a widespread practice usually employed in sample surveys to improve the precision of the final survey results (also to compensate for non-observation errors). It consists in deriving new weights, as close as possible to the starting ones, which fulfil a series of constraints concerning the totals of a set of auxiliary variables (usually known at population level).

3.1 Renssen's weights calibration procedure

Renssen's SM procedure is particularly suited to manage categorical *X*, *Y* and *Z* variables and is not primarily designed to integrate surveys at microdata level, as the main purpose is the estimation of the contingency table $Y \times Z$. This procedure is articulated in two steps; the first step consists in calibrating weights in both *A* and *B* to align the corresponding estimated

totals of the matching variables X_M (joint or marginal distribution) with known (or estimated) population totals; the second step estimates the two-way contingency table $Y \times Z$. Estimation can be done under the CI assumption:

$$\hat{P}^{(CI)}_{Y=j,Z=k} = \hat{P}^{(A)}_{Y=j|X_M=i} \times \hat{P}^{(B)}_{Z=k|X_M=i} \times \hat{P}_{X_M=i} \quad | \; i=1,\ldots,I \;, j=1,\ldots,J \;, \; k=1,\ldots,K \tag{1}$$

whereas the terms in the formula are obtained by considering the units' weights (w'), modified after the first harmonisation step, *i.e.*:

$$\hat{P}_{X_M=i} = \sum_{a=1}^{n_A} w'_a \, I(x_{Ma} = i) \tag{2}$$

Renssen's approach also allows exploiting a third additional data source *C* in which *Y* and *Z* are jointly observed. A first option consists in estimating $Y \times Z$ directly on *C* after a further calibration step performed on it, aimed at aligning the marginal distributions of *Y* and *Z* with the corresponding ones estimated in respectively *A* and *B* after the initial harmonisation (*incomplete two-way stratification*). In alternative, it is possible to perform the *synthetic two-way stratification*, which consists in "correcting" the estimate obtained under the CI assumption with the additional information provided by *C*. In such a case, *C* must also include the matching variables. In practice, also this alternative procedure consists in a series of calibration steps.

A very appealing feature of the Renssen's procedure is that the marginal distributions of the resulting contingency table $Y \times Z$ are aligned with those estimated on the starting datasets, but after the initial harmonisation step. This is a very important characteristic in official statistics where the coherence of the final statistical outputs is one of the key dimensions of the quality of statistics and plays a crucial role when integrating data from sources referred to the same target population.

Although the whole Renssen's procedure is designed with a macro purpose (estimating the contingency table $Y \times Z$) it also allows to perform imputation at micro level. Micro objective is pursued by generating the predicted values of the *linear probability models* that are fit across the whole procedure. A linear probability model assumes that the probability of an event (falling in a given consumption or income class in our case) can be expressed as a linear combination of a series of explanatory variables (matching variables in our application). These models are not the ones suited

to deal with this case[4], but they are used as "working" models because the corresponding predicted values (the estimated probability of assuming each of the categories of Z (Y) for every unit in A (B)) maintain the appealing feature of the whole procedure. In other words, when used to estimate the marginal distribution of Z (Y) in A (B) they return the same estimated distribution that is achieved by considering the data set B (A) (after the harmonisation). Unfortunately, the estimated probabilities provided by linear probability models should be used carefully, given that these models present some well-known drawbacks (estimated probabilities can be less than 0 or greater than 1; heteroskedastic residuals, *etc.*).

Practically, having the predicted probabilities at the end of SM may not be a viable option for the practitioner that would prefer having imputed categories for the target variable for easing the subsequent analyses. In this sense, a "direct" imputation would consist in adopting a randomised device that generates the imputed category by a random draw with probabilities proportional to the predicted probabilities; this would avoid the well-known negative consequences of getting the "most voted" category (the one with the highest predicted probability), as also shown by Donatiello *et al.* (2016a). Renssen (1998) suggests an "indirect" two-step imputation that consists in a mixed SM micro procedure; in practice, the estimated predictions are the input of a *nearest neighbour hotdeck* procedure (Singh *et al.*, 1993) where the final value to impute is the one observed on the closest donor according to the distance calculated on the predictions. This is one of the possible many variants of the SM mixed methods listed in Section 2.5 of D'Orazio *et al.* (2006). SM mixed methods are mainly developed for continuous target variables but can be easily adapted to handle predicted probabilities for the categories of a categorical target variable, as it will be shown in the next section.

It is worth noting that Renssen's SM allows the introduction of target continuous variables but in this case, some difficulties may arise in the various subsequent calibration steps. We are currently investigating the possibility of applying the micro "extension" of the original proposal to the case of continuous target variables and some preliminary results are quite satisfactory (Donatiello *et al.*, 2017), but they will not be presented in this

4 Some suggestions related to use of models in SM when the target variables is a categorical response variables are in de Waal (2015).

article as additional investigation is deserved. Essentially, in this work we consider continuous *Y* and *Z* target variables, but for the application of the Renssen's SM procedure we categorise them, although the procedure is designed to end up with a synthetic fused dataset with continuous values for the imputed missing variable to be fully used for validation and economic analysis.

3.2 Matching of EU-SILC and HBS

SM exercise aimed at imputing the household consumption variable (*Z*), originally observed in HBS (donor), in the SILC survey (households). This setting is not optimal as contradicts the common suggestion of using the smaller dataset as the recipient ($n_B = 15\,409$ households in HBS vs. $n_A = 22\,226$ in SILC). However, the difference is not so huge and not very relevant as the two-step mixed procedure is used instead of a "standard" SM *hotdeck* procedure. Specifically, two different variants of the Renssen's procedure are proposed for the final imputation of *Z* into SILC. The first step follows the Renssen's recommendations and is common to both the procedures; it consists in the calibration of the survey weights of both the data sources to reproduce the same estimated marginal distribution of the matching variables. Following the previous matching exercises, we identified few relevant matching variables: the geographical areas (5 categories, "North-West Italy", "North-East Italy", "Centre Italy", "South of Italy", "Islands of Italy") and the number of durable goods owned by the households (4 categories). In addition, to have the CI holding the proxy variable of household consumption is included in the set of matching variable; this variable, C^*, is a categorised version (8 categories) of the consumption variable observed in the new SILC rolling module on Consumption & Wealth; for matching purposes, the same consumption variable is derived in HBS using data collected from the survey.

Marginal distributions of the three matching variables are aligned to reproduce the fixed totals. In particular, the reference distributions of the Italian households by geographical areas are achieved by pooling the estimates provided by the starting data sources; the same procedure is used for the number of durable goods, while the reference distribution of households by classes of proxy consumption C^* is estimated on the HBS starting data.

The second step of the matching procedure consists in estimating the contingency table $Y^* \times Z^*$, being Y^* and Z^* the categorised versions of respectively *Y*, the household income provided by SILC, and *Z*, the household consumption expenditures observed in HBS. A first estimate of the $Y^* \times Z^*$ table can be achieved without integration at micro level by applying the expression (1), since the CI assumption can be considered valid as the set of matching variables includes a proxy oh the household consumption. As expected, this table has marginal distributions of both Y^* and Z^* that are equal to the ones provided by the starting data sources after the initial harmonisation step.

Then, as a by-product of the estimation of $Y^* \times Z^*$, we derive predictions of $\hat{p}(y^* = j)$ $(j = 1,2, \dots, J = 8)$ and $\hat{p}(z^* = k)$ $(k = 1,2, \dots, K = 8)$ for the units in both the data sources. Finally, for imputing the values of *Z* in SILC the following additional steps are proposed:

Step 3.1): imputation in SILC of the value of *Z* observed on the closest donor in HBS according to the following distance:

$$\hat{\Delta}_{a,b} = \frac{1}{2}\sum_{k=1}^{K} |\hat{p}(z_a^* = k) - \hat{p}(z_b^* = k)| \qquad (3)$$

In this expression $\hat{p}(z_a^* = k)$ is the predicted probability that the ath unit in *A* $(a = 1, \dots, n_A)$ gets an imputed category equal to *k* for the variable *Z*; similarly, $\hat{p}(z_b^* = k)$ is the same predicted probability for the *b*th unit in *B* $(b = 1, \dots, n_B)$.

The distance function (3) corresponds to the *total variation distance or dissimilarity index* and was chosen as the distance should be calculated on predicted probabilities. In fact, as also noted by de Waal (2015), this specific situation requires the adoption of suitable distance functions aimed at measuring dissimilarity between distributions of categorical variables (for a review of these distances see *e.g.* Cha, 2007). An alternative to the total variation distance can be the Hellinger's distance, but we opted in favour of the total variation distance because it corresponds to ½ of the Manhattan distance and this latter one is already implemented in many statistical packages.

Step 3.2): imputation in SILC of the value of *Z* observed on the closest donor in HBS according to the sum of the total variation distances related to predictions of both *Y* and *Z*:

$$\hat{\Delta}_{a,b} = \frac{1}{2}\sum_{j=1}^{J} |\hat{p}(y_a^* = j) - \hat{p}(y_b^* = j)| + \frac{1}{2}\sum_{k=1}^{K} |\hat{p}(z_a^* = k) - \hat{p}(z_b^* = k)| \qquad (4)$$

The steps (3.1) and (3.2) are alternative implementation of the final step of a SM mixed approach based on predictive mean matching, following suggestions in D'Orazio *et al* (2006). Such a mixed SM procedure joins the advantages of both parametric and nonparametric approaches; in particular, the final phase permits to exclude the matching variables from the computation of distance and is robust to model misspecification. In the presence of several potential donors at the minimum distance from the *a*th recipient unit, for matching purpose just one the donors is picked up completely at random.

As, in the end, the SM procedures described in this Section resemble a SM mixed approach it was decided to compare their results with those of a "standard" mixed approach that does not take into account the survey weights and the constraint of preserving the marginal distributions of the target variables. As shown in D'Orazio *et al* (2006, Section 2.5.1), several variants of the regression-based mixed approach exist when the target variables are continuous, for comparability purposes the choice fell on MM3 and MM5. Specifically, the regression step (step 1) of MM3 consists in fitting in HBS a linear regression model, where *Z* (log transformed) is predicted by the chosen matching variable (the log transformed version of the proxy variable of household consumption *C* is considered). The fitted model is then used to derive in SILC the "intermediate" values of *Z* ($\tilde{z}_a = \hat{z}_a + e_a$) obtained by summing the predicted values of *Z* in SILC ($\hat{z}_a$) with a random error term (e_a) generated from a gaussian distribution with mean 0 and residual standard deviation of the model. The matching step (2nd step) imputes in SILC the value of *Z* observed on the closest donor in HBS according to the distance $d_{a,b} = |\tilde{z}_a - z_b|$. It should be noted that the procedure MM5 also fits in SILC a regression model of *Y* (log transformed) vs the matching variables and then uses it to estimate the intermediate values of *Y* in HBS ($\tilde{y}_b = \hat{y}_b + e_b$). In the matching step of MM5 method the value of *Z* imputed in SILC is the one observed on the closest donor in HBS according to the distance $d_{a,b} = |y_a - \tilde{y}_b| + |\tilde{z}_a - z_b|$ (in the proposed MM5 procedure the matching step is constrained to use donors only once, but this is not possible in our application because SILC is larger than HBS).

Matching exercise was performed in R using the facilities of the **StatMatch** package (D'Orazio, 2022). The results of the proposed method and a comparison with the mixed approach that does not take into account the survey weights are presented in the next section.

4. Main results of the SM application

Typically, the assessment of results of a SM application at micro level consists in analysing whether the synthetic dataset, *i.e.* SILC with imputed household consumption, preserves the marginal distribution of the imputed variable and the relationship of this variable with the matching variables. These checks are necessary because it is not possible to assess the accuracy of the estimates involving the imputed variable obtained at the end of the whole SM procedure by estimating the MSE or just the variance. This is still an open problem of SM applications, where an indirect partial assessment of the variability associated to the final estimates can only be obtained when approaches based on assessment of SM uncertainty are applied (see *e.g.* Conti *et al.* 2012; Zhang, 2015).

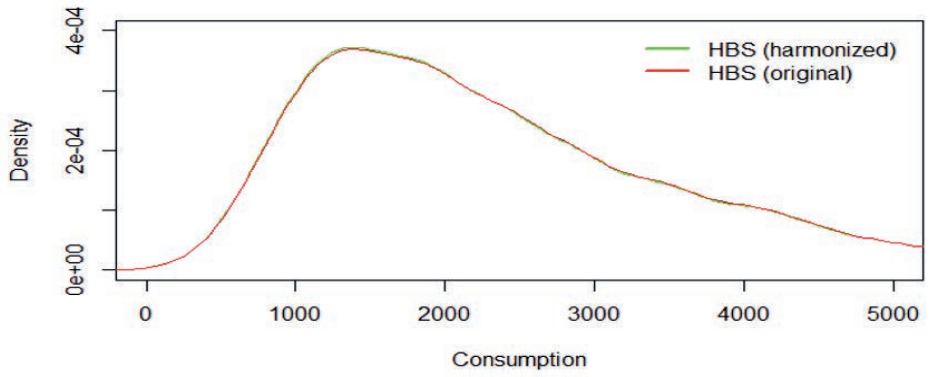
In addition, subject matter experts try to assess also the "plausibility" of the joint relationship between the imputed target variable, the total household consumption in our case, and the other target variable, *i.e.* the household income.

In our application, we believe that additional checks are required as the synthetic data set is the outcome of a complex SM procedure, whose first step modifies (calibration) the starting survey weights with the aim of harmonising the marginal distribution of the matching variables. These modified weights are then the ones to be used when analysing the data starting from the synthetic data set.

In this respect, 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.

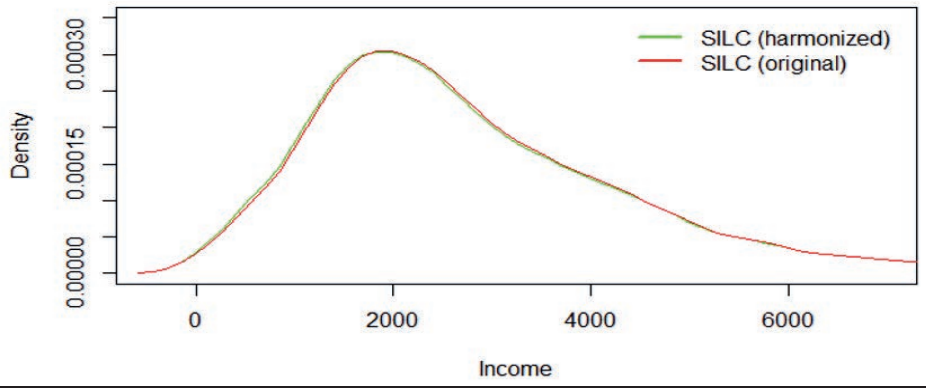
Figures 4.1 and 4.2 show that the estimated distributions of both the target variables remain almost unchanged considering both original and modified weights.

Figure 4.1 - Comparison of total consumption in HBS before and after the initial harmonisation step



Source: Istat HBS 2016

Figure 4.2 - Comparison of total income in SILC before and after the initial harmonisation step



Source: Istat EU-SILC 2017

An additional check consists in investigating the preservation of the joint distribution between the discretised target variable and some of the available common variables. Change in the joint distribution is measured by the Hellinger distance (HD) between the distribution estimated before and after the initial harmonisation step. As shown in Table 4.1, the distances are well below the 5% rule-of-thumb threshold, indicating that the modification of the weights does not affect markedly the joint distribution between the discretised target variable

and each of the considered common variables. The highest value for the HD, but still below the 5% threshold, is observed in EU-SILC for the joint distribution of discretised total income and discretised partial consumption estimated with the data collected in the new module. This is somehow expected since this partial consumption variable derived in SILC cannot be considered as accurate as the corresponding one observed in HBS[5]. For this reason, in the harmonisation step the weights were modified to return the distribution estimated from the HBS rather than, as usual, the pooled estimate.

Table 4.1 - Hellinger Distance of the joint distribution of income and consumption classes before and after the harmonisation step by common variables

	HBS	SILC
Partial consumption	0.5%	4.5%
Durable goods	1.8%	1.9%
Macro areas	0.3%	0.8%
Sex	0.3%	0.9%
Education	0.5%	1.0%
Citizenship	0.4%	0.9%
# people in household	0.4%	1.2%
Tenure status	0.4%	1.1%
# employed people	0.4%	1.0%
Household type	0.4%	1.2%

Source: Istat HBS 2016 and EU-SILC 2017

Finally, as HBS is the main reference for estimating the poverty, we have compared the relative and absolute poverty incidence estimated in HBS before and after the harmonisation step. Results (see Table 4.2) are quite satisfactory as there is a very slight change in the fraction of poor households. This result, however promising, is to be taken with caution, as HBS is the sole survey entitled for the dissemination of estimates on relative and absolute poverty in Italy. It should be noted that Table 4.2 also reports the estimate of absolute poverty derived on the synthetic data set at the end of the SM procedure. In this case, the estimates provided by step (3.1) show a small change compared to the reference ones estimated in HBS that however is about 0.2%.

Analysing more in detail the synthetic data set, Figure 4.3 shows the estimated density functions of total consumption imputed in SILC with respectively the

5 There are some relevant differences such as the method of data collection, in fact HBS uses a diary while the consumption collected in SILC is obtained through few direct questions. Furthermore, the HBS is a continuous quarterly survey while SILC is annual.

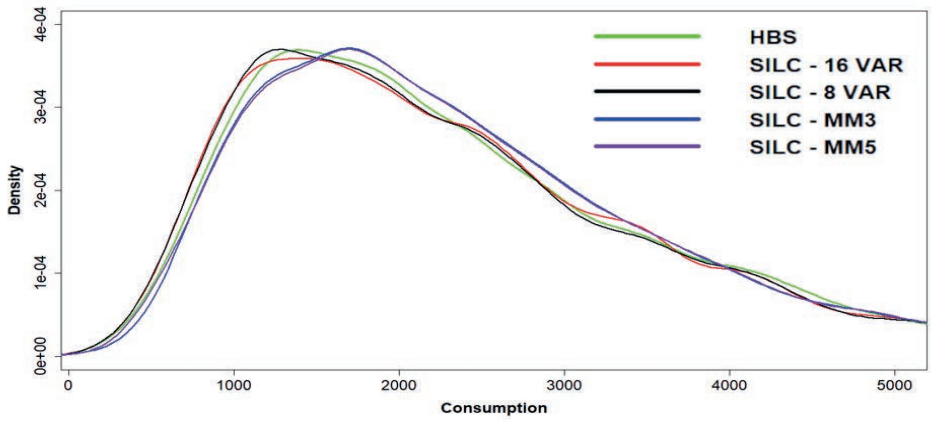
Table 4.2 – Poverty indicators by type of weight before and after the harmonisation step in the synthetic data set (percentage values)

	Origin survey weights	Weights after harmonisation	
	HBS	HBS	FUSED
Relative poverty	10.60	10.64	10.82
Absolute poverty	6.28	6.30	6.12

Source: Istat HBS 2016 and EU-SILC Fused 2017

proposed procedure with slight modification of the final step, (3.1) (nearest neighbour donor with distance calculated on the predicted probability of falling in one of the categories of the variable *Z*, denoted as "8 VAR" in Figures 4.3 and 4.4) and (3.2) (nearest neighbour donor with distance calculated on the predicted probabilities for both *Y* and *Z* variables; denoted as "16 VAR" in Figures 4.3 and 4.4), and the procedures MM3 and MM5. All these estimated distributions are compared to the consumption distribution measured in HBS. It is worth noting that the distribution estimated with the imputed values provided by step (3.1) is closer to the original distribution than the one estimated using imputed values given by step (3.2). On the contrary, the procedure MM3 and MM5 return an imputed consumption whose distribution is shifted toward the right, that tends to overestimate the overall consumption. These results clearly show that the additional step of harmonising the weights through the Renssen's method improves the final estimates.

Figure 4.3 - Comparison of original HBS and imputed total consumption in SILC by our method (distance function 8 or 16 dummies) **and mixed procedures** (MM3 and MM5)



Source: Istat EU-SILC 2017

Moreover, another "traditional" check to evaluate the outputs of the SM application consists in comparing the joint distributions of the imputed variable and each of the most relevant common variables.

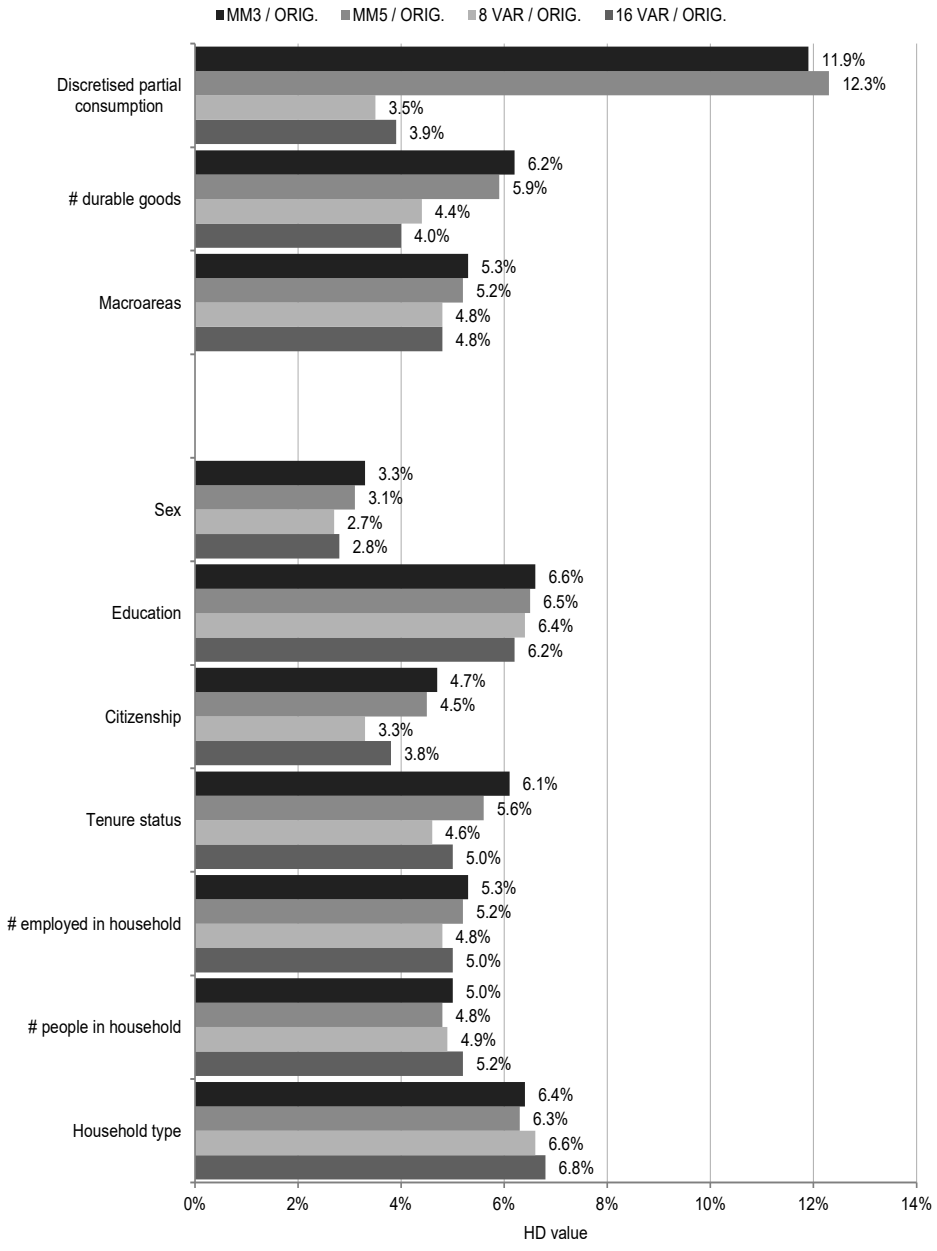
Figure 4.4 shows the HD between the distributions estimated respectively from the synthetic data set and from HBS for each combination of the quintiles of total consumption and some of the common variables (including the matching variables). Cases with the HD below 5% (the chosen threshold) indicate closeness between the distribution estimated on the fused dataset and the reference one (estimated on HBS). Distances over the threshold are observed when crossing the consumption quintiles with the household type, the level of education and with the proxy of consumption but only with procedures MM3 and MM5. In the first two cases the outcome is presumably due to the fact that the variables household type and education level were not used as matching variables as well as to differences in how they are observed in the two surveys. The failure of MM3 and MM5 in preserving the joint distribution of the overall consumption with its proxy in SILC, seems mainly explained by the some difficulties in imputing reliable low levels of overall consumption.

In general, Figure 4.4 shows that imputation using the proposed method with distance on the predicted classes of the discretised overall income (step 3.1) tends to perform better in preserving the relationship of consumption and some of the relevant common variables in SILC. On the contrary, methods MM3 and MM5 are often performing worse.

As the post matching checks show that the proposed procedure with final step (3.1) ("8 VAR" in Tables or Figures) tends to perform better that the other applied methods, all subsequent analysis will consider only the imputed SILC data set at the end of this procedure.

Table 4.3 presents means and medians of the imputed consumption in SILC at the end of step (3.1) and of the observed consumption in HBS by matching variables and some common variables; the relative differences between 5% and 10% are highlighted in light red and the ones over 10% in dark red. As expected, the larger differences are observed in correspondence of the same variables identified when calculating the HD between estimated and reference joint distributions (Figure 4.4). In general, the comparability is high with many differences under 5% and a difference for the total distribution close to 1%.

Figure 4.4 - Hellinger distance comparing estimated joint distribution crossing income quintiles and common variables in the synthetic data set and in HBS



Source: Istat HBS 2016 and EU-SILC Fused 2017

Table 4.3 - Comparison of imputed and original total consumption (Z) by common variables (Values in euros)

	Imputed Z in SILC		Z in HBS		Imputed Z / Z	
	Mean	Median	Mean	Median	Mean	Median
DURABLE GOODS						
up to 5	1370	1165	1448	1257	95	93
6	2063	1789	2151	1850	96	97
7	2774	2466	2743	2464	101	100
8	3563	3210	3581	3290	99	98
MACRO AREAS						
North	2589	2236	2713	2384	95	94
Centre	2535	2134	2554	2156	99	99
South and Islands	2144	1841	2028	1812	106	102
SEX						
Man	2652	2321	2642	2300	100	101
Female	2053	1663	2143	1792	96	93
EDUCATION						
Less Then Primary	1740	1400	1703	1447	102	97
Primary	2397	2037	2255	1956	106	104
Secondary	2706	2387	2722	2396	99	100
Post-Secondary or Upper	3005	2650	3465	3105	87	85
CITIZENSHIP						
Italian	2490	2137	2536	2185	98	98
Foreign	1807	1502	1640	1305	110	115
TENURE STATUS						
Owner	2059	1747	1857	1574	111	111
Rent	2617	2288	2674	2327	98	98
Usufruct	2015	1669	2130	1785	95	94
# EMPLOYED PEOPLE						
0	1805	1484	1991	1649	91	90
1	2487	2159	2410	2097	103	103
2	3340	3005	3391	3091	98	97
3 or more	3891	3478	3794	3584	103	97
# PEOPLE IN HOUSEHOLD						
1	1644	1378	1760	1467	93	94
2	2351	2060	2555	2196	92	94
3+	3182	2843	3023	2685	105	106
HOUSEHOLD TYPE						
Single	1644	1378	1690	1514	97	91
Couples Without Children	2372	2098	2528	2151	94	98
Couples With Children	3226	2875	2948	2607	109	110
Single Parent	2487	2179	2432	2119	102	103
Others	2755	2455	2598	2218	106	111
TOTAL	2437	2080	2471	2107	99	99

Source: Istat HBS 2016 and EU-SILC Fused 2017

Table 4.4 – Comparison of Propensity to consume by data source and common variables

	FUSED	HFCS
	APC	APC
DURABLE GOODS		
up to 5	0.78	-
6	0.86	-
7	0.86	-
8	0.83	-
MACRO AREAS		
North	0.81	0.73
Centre	0.82	0.78
South And Islands	0.91	0.80
SEX		
Man	0.82	0.75
Female	0.87	0.79
EDUCATION		
Less Then Primary	0.80	0.81
Primary	0.94	0.81
Secondary	0.86	0.74
Post-Secondary or Upper	0.71	0.69
CITIZENSHIP		
Italian	0.83	-
Foreign	1.06	-
TENURE STATUS		
Owner	1.09	-
Rent	0.79	-
Usufruct	0.88	-
# EMPLOYED PEOPLE		
0	0.82	-
1	0.93	-
2	0.77	-
3 or more	0.69	-
# PEOPLE IN HOUSEHOLD		
1	0.89	-
2	0.78	-
3+	0.85	-
HOUSEHOLD TYPE		
Single	0.89	-
Couples Without Children	0.75	-
Couples With Children	0.84	-
Single Parent	0.92	-
Others	0.80	-
QUINTILES OF INCOME		
1° quintile	1.61	1.21
2° quintile	1.03	0.90
3° quintile	0.93	0.83
4° quintile	0.83	0.75
5° quintile	0.59	0.64
TOTAL	0.84	0.76

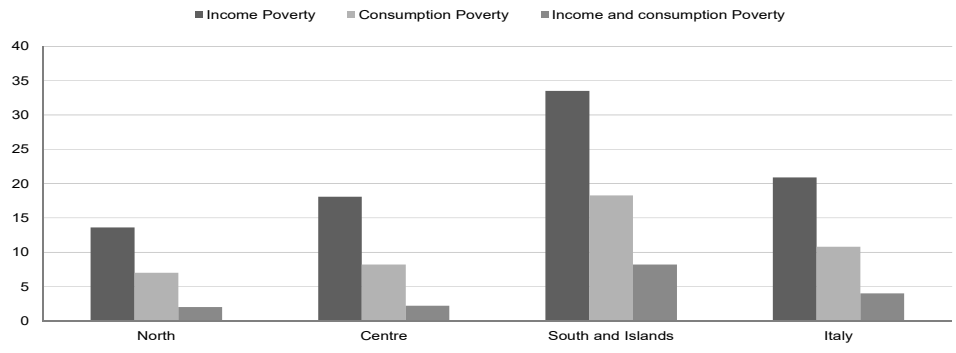
Source: Istat EU-SILC Fused 2017 and Bank of Italy HFCS 2016

Finally, to summarise the relationship between income and consumption in the synthetic data set we have estimated the propensity to consume by some common variables (Table 4.4). Results are between 0 and 1 for most cases with the overall propensity to consume equal to 0.84. This is a very indicative first result, to be considered rather as an exercise and not as a reliable estimate. In any case, as expected, the propensity to consume decreases as the income increases: when crossed with the quantile of income, the propensity to consume ranges from 1.61 for the first quintile to 0.59 for the richest quintile. Moreover, for households who have to pay the rent, the propensity to consume is significantly lower than that of the owners (0.79 *vs*. 1.09). Propensity to consume estimated by Bank of Italy with the European Central Bank's Household Finance and Consumption Survey (HFCS) is presented for comparison, as it represents the unique source in which household income and consumption expenses are jointly observed. Overall propensity is 0.76, therefore lower than our findings, but it is worth considering that comparability is affected by the different way of estimating the consumption based on the consumption expenses collected through the survey and the fact that usually HFCS provides lower consumption levels if compared to HBS, that provides the reference estimates.

5. Initial comparative analysis of poverty

In general, all the various checks done on synthetic data set obtained at the end of the Renssen's SM procedure, with modifications suggested in this paper, indicate that the results go in the desired direction of providing a reliable picture of joint distribution of income and consumption. For this reason, although at a very early stage, it is possible to draft some considerations on one of the main objectives that the analysts would like to achieve by analysing the joint distributions of income and consumption that is to have more insights on economic poverty[6].

Figure 5.1 - Income and consumption poverty by macro areas (percentage values)



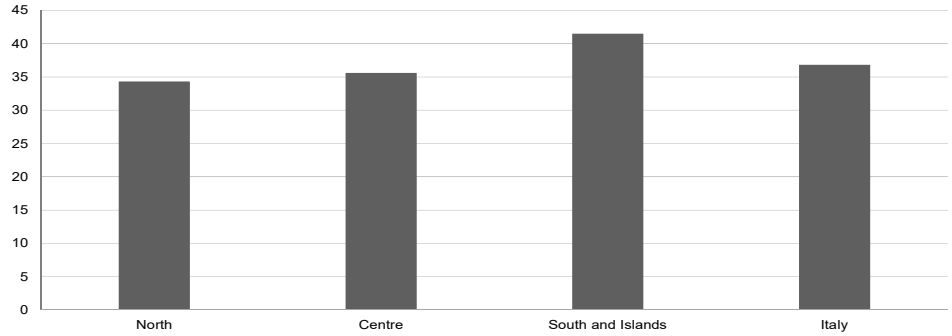
Source: Istat EU-SILC Fused 2017

Preliminary data suggest consumption poverty lower than those for income with overlap between income and consumption poverty around 4% of households (Figure 5.1).

Share of households with expenditure higher than income is an important indicator as the households that are unable to finance consumption entirely from income may face financial difficulties and may reduce their assets. Estimates calculated on the synthetic data source are in the Figure 5.2 and indicate that more than a third of households have consumption above income, especially in the South and Islands. These are findings to be interpreted very carefully as, especially for low-income households, it is known that income is more frequently underestimated while consumption is usually reported more accurately.

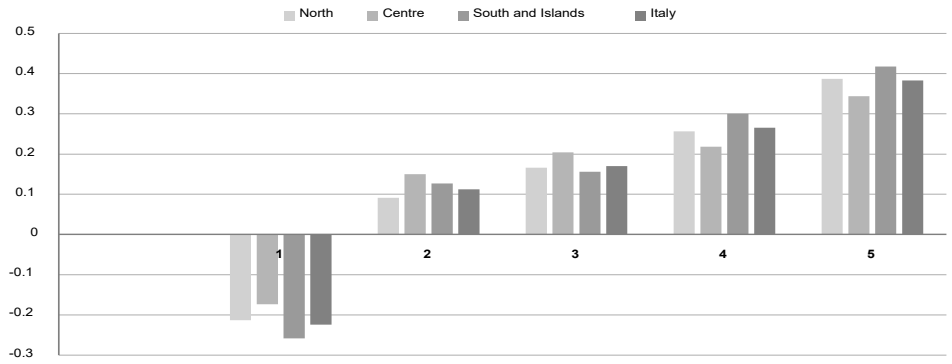
6 Scientific community has long agreed that income can be considered a good proxy for living standards but becomes a better measure when it is associated with data on household consumption and wealth. See: Stiglitz-Sen-Fitoussi Report 2009 and OECD 2013.

Figure 5.2 - Households with income less than consumption by macro areas (percentage values)



Source: Istat EU-SILC Fused 2017

Figure 5.3 - Median income saving rate by income quintiles and macro areas (percentage values)



Source: Istat EU-SILC Fused 2017

In Figure 5.3, data show a strong relationship between saving rates and income quintiles with low saving rates in Q1 that seem to reflect temporary low incomes but, as mentioned above, also potential under-reporting of low incomes/ high expenditures. A significant proportion of households have negative saving rates that, although in line with Eurostat's experimental estimates[7], need to be interpreted with caution.

7 Eurostat's experimental estimates on the joint distribution of income, consumption and wealth are available at: https://ec.europa.eu/eurostat/web/experimental-statistics/income-consumption-and-wealth.

6. Where we stand and the way forward

The construction of a data set containing the joint information on household income and consumption and the estimates provided by it are part of the project for the production of microdata relating to household income, consumption and wealth in Italy (ICW project). The current achievements reflect the work done over the past few years and take stock of all the findings of the various SM exercises done at Istat starting from 2005. Essentially, the proposed SM approach relies on three key choices: (*i*) an *ex ante* harmonisation of the social survey EU-SILC and HBS, (*ii*) the collection through an *ad hoc* module of information relating to the dimensions of consumption and wealth in the income survey EU-SILC, and finally (*iii*) the application of statistical matching techniques that take into account the survey weights.

Harmonisation of the Italian EU-SILC and HBS, started in our Institute in 2011, nowadays is quite extensive covering the sample design, concepts, definitions and consistent treatments and classifications. From the beginning, the harmonisation of the two surveys was also made to facilitate the application of micro-integration methods. Furthermore, the SILC modular approach allowed collecting variables relating to consumption and wealth, which, although limited in number and generally less accurate than data collected using an *ad hoc* survey, have proved to be important as hook variables in the matching procedures.

Statistical matching techniques we used are based on conditional independence assumption. It is well-known that such integration methods are considered a *second best* choice since it is not possible to fully capture the relationships between all the variables of interest conditional on a relatively small set of common matching variables. However, the SILC consumption and wealth module enabled us to test the efficacy of the use of proxy variables of the targets as matching variables capable of justifying the CIA.

As known, the results stemming out from micro integration techniques must be carefully evaluated in order to assess the validity and plausibility of the synthetic dataset, before they can be used for policy purposes and to design measures to fight poverty and material deprivation. Hence, for the validation of our results we used all the criteria suggested in literature (Rassler, 2002) and decided to add some additional checks.

Looking ahead, further efforts should be made to improve the SM method presented in this paper. In particular, care is needed in the initial weights calibration to harmonise the marginal (or joint) distributions of the chosen matching variables. This is not a straightforward step because it requires the estimation of the reference marginal distributions (reference totals) and may not end successfully, *i.e.* the calibration may not converge or may return negative weights (can happen with some calibration functions). This problem generally worsens by increasing the number of matching variables; therefore, also in this case, the main recommendation is to select few relevant matching variables. In addition, this initial weights' calibration may affect the results of the whole SM procedure if it introduces marked changes in the marginal distributions of the target variables. This issue often is not taken into account and we believe that it deserves special attention in the assessment of the results of this specific SM procedure. In our case, the initial harmonisation step came out to be a "cosmetic" weights' calibration which in fact has not introduced significant changes in the distributions of the relevant variables, as well as in the traditional poverty indicators. Furthermore, we have verified that the Renssen weight harmonisation step, together with the variants proposed by us, provide better results than the more traditional mixed methods. The comparison between our method and the MM3 and MM5 procedures clearly demonstrates that the distributions obtained with our method are closer to the reference one of HBS. The traditional mixed procedures, without the additional weight harmonisation step, tend to substantially overestimate the total consumption, as they likely fail in estimating the low incomes of the distribution.

In this work, we have essentially investigated the possibility of modifying the origin Renssen's method to create a synthetic dataset by imputing a continuous rather than a categorical variable. For this purpose, we passed through a discretisation of the target variables but we believe that in the future this step can be skipped, although some additional investigation is deserved.

Future plans also consider the possibility of producing a synthetic data set that includes the dimension related to the household assets. These data are well observed in the Bank of Italy's Survey on Household Income and Wealth but the exploitation of this additional survey in the SM exercise presents a number of methodological challenges. This is an unprecedented activity, which, however, can benefit of the experience accumulated in our previous matching exercises.

The final goal of our work is the production of microdata on the joint distribution of household income, consumption and wealth, which will represent the basis for producing experimental statistics, expected to be disseminated in aggregate tables but only after an accurate phase of validation of their reliability. We are aware that the latter issue is not straightforward due to complexity and underlying assumptions in the application of statistical matching techniques. For these reasons, particular attention will be paid to informing external users on the nature of the synthetic data produced, on the model applied and on the quality assessment.

7. Concluding remarks

Availability of joint micro data on income and consumption is fundamental to measure the poverty and the living conditions of households, overcoming the measures used up to now based on the observation of a single dimension (income or consumption). Statistical matching methods presented in this work seem particularly promising although show some methodological constraints and require additional checks for assessing the plausibility of the final estimates. It is therefore necessary to point out that our results are experimental and here are presented to highlight their potential advantages for the economic analysis. Although the general assessment is based on metrics recognised in the literature, we believe it is important to deepen the validation phase of the experimental statistical outputs.

Initial analysis of our results shows that the reproduction of the marginal distribution of the imputed consumption variables by the statistical matching turned out to be satisfactory. Comparison of the estimated probability density functions for the synthetic consumption variables and the original ones shows a good overlap especially on the tails of the distribution. In addition, we have also analysed the correlation structure between the target variables observed in the original and fused datasets. An effective match should lead to similar relationships between common and target variables in the donor and the matched file; the results obtained in our application show that the sign and order of magnitude of the correlation/association remain almost the same in both datasets. Moreover, the preservation of the joint distribution of the quintile of total consumption with each of the considered common variables show a level of comparability, which can be considered quite good. Synthetic dataset obtained at the end of the SM procedure permits a first multidimensional analysis of poverty and makes it possible to highlight areas of vulnerability for households. Data that have been analysed in this work refer to 2016 and, even with numerous caveats, allow us to make some initial analyses on economic poverty by studying jointly two dimensions of household conditions. The economic situation like the current one has certainly increased the request of multi-dimensional measures to assess the ability of households to support their living standards and to cope with important economic shocks. For this reason, we intend to apply the SM procedures presented here to the data referred 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 household wealth in mitigating the effects of the crisis.

References

Cha, S.-H. 2007. "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions". *International Journal of Mathematical Models and Methods in Applied Sciences*, Volume 1, Issue 4: 300-307

Coli, A., F. Tartamella, G. Sacco, I. Faiella, M. Scanu, M. D'Orazio, M. Di Zio, I Siciliani, S. Colombini, e A. Masi. 2005. "La costruzione di un archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine Istat sui consumi delle famiglie italiane e l'indagine Banca d'Italia sui bilanci delle famiglie italiane". *Documenti*, N. 12/2006. Roma, Italy: Istat. https://www.istat.it/it/archivio/219083.

Consolini, P., G. Donatiello, D. Frattarola, and M. Spaziani. 2018*a*. "The Consumption and Wealth data in IT-SILC 2017". *Proceedings of the Workshop on Best Practices for EU-SILC Revision*, Warsaw, Poland, 16th – 17th October 2018.

Consolini, P., G. Donatiello, D. Frattarola, and M. Spaziani. 2018*b*. "The IT-SILC measurement of the household finance, wealth and consumption". *Proceedings of the 35th IARIW General Conference*, Copenhagen, Denmark, 20th - 25th August 2018. http://old.iariw.org/copenhagen/consolini.pdf.

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.

de Waal, T. 2015. "Statistical matching: Experimental results and future research questions". Statistics Netherlands – CBS, *Discussion Paper* 2015/19.

Donatiello, G., M. D'Orazio, D. Frattarola, A. Rizzi, M. Scanu, and M. Spaziani. 2016*a*. "The role of the conditional independence assumption in statistically matching income and consumption". *Statistical Journal of the IAOS*, Volume 32, N. 4: 667-675.

Donatiello G., M. D'Orazio, D. Frattarola, A. Rizzi, M. Scanu, M. Spaziani. 2016*b*. "Statistical matching of EU-SILC and HBS at ISTAT: where do we stand for the production of official statistics". *Proceedings of the DGINS Conference of the Directors General of the National Statistical Institutes, Statistics on income, consumption and wealth*, Statistics Austria, Vienna, 26th – 27th September 2016.

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

Donatiello, G., M. D'Orazio, D. Frattarola, M. Scanu, and M. Spaziani. 2017. "Towards the production of integrated statistics on Income, Consumption and Wealth: pre-requisites and methodological challenges". *Proceedings of ITACOSM 2017, The 5th ITAlian Conference on Survey Methodology*, University of Bologna, Italy, 14th – 16th June 2017.

Donatiello, G., D. Frattarola, A. Rizzi, and M. Spaziani. 2015. "The Role of the Available Information in Statistical Matching It-SILC and HBS". *Proceedings of the Workshop on Best Practices for EU-SILC Revision*, Imperial College, London, UK, 16th–17th September 2015.

Donatiello, G., D. Frattarola, A. Rizzi, and M. Spaziani. 2014*b*. "Statistical Matching of IT-SILC and HBS: Some Critical Issues". *Proceedings of the Workshop on Best Practices for EU-SILC Revision*, Banco de Portugal, Lisbona, 15th – 17th October 2014.

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.

Eurostat. 2021. "Income, Consumption and Wealth". *Experimental statistics*. Luxembourg: Eurostat. https://ec.europa.eu/eurostat/web/experimental-statistics/income-consumption-and-wealth.

Istituto Nazionale di Statistica/Italian National Institute of Statistics – Istat. 2017. "Indagine sulle condizioni di vita (EU-SILC) - Dati Trasversali: File per la ricerca". *Microdati*. Roma, Italy: Istat. https://www.istat.it/it/archivio/212385.

Istituto Nazionale di Statistica/Italian National Institute of Statistics – Istat. 2016. "Indagine sulle spese delle famiglie: File per la ricerca". *Microdati*. Roma, Italy: Istat. https://www.istat.it/it/archivio/180341.

Organisation for Economic Co-operation and Development - 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.

Singh, A.C., H.J. Mantel, M.D. Kinack, G. Rowe. 1993. "Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption". *Survey Methodology,* Volume 19, N. 1: 59-79.

Stiglitz, J.E., A. Sen, and J.-P. Fitoussi. 2009. *Report by the Commission on the Measurement of Economic Performance and Social Progress*. Paris, France: French Government, CMEPSP.

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