COMITATO CONSULTIVO PER LE METODOLOGIE STATISTICHE
ADVISORY COMMITTEE ON STATISTICAL METHODS

The statistical matching of EU-SILC and HBS at ISTAT: where do we stand for the production of official statistics

Gabriella Donatiello (donatiel@istat.it), Marcello D'Orazio (madorazi@istat.it), Doriana Frattarola (frattarola@istat.it), Mauro Scanu (scanu@istat.it), Mattia Spaziani (mspaziani@istat.it)[1]

[1]Istat

Key words: data fusion, uncertainty, conditional independence assumption, Fréchet bounds

Preface

The following paragraphs were produced as a paper presented to the last DGINS 2016 (meeting of the Director Generals of the National Statistics Institutes at Eurostat) and describe one of the latest statistical matching applications in Istat. The statistical matching application had the objective (sponsored by Eurostat at European level) to estimate the joint distribution of household expenditures and income from data observed in two distinct samples, respectively the Household Budget Survey (HBS, containing reliable and detailed information on household expenditures) and the European Survey on Income and Living Conditions (EU-SILC, that includes reliable information on household income).

The idea to replicate this paper to the attention of the Advisory Committee on Statistical Methods is due to the latest developments in the dissemination of data obtained by means of non-standard approaches by National Statistical Institutes in dedicated webpages named Experimental Statistics. In general, Experimental Statistics are "…series of statistics that are in the testing phase and not yet fully developed" (Office for national Statistics), "…which are published in order to involve users and stakeholders in their development and as a means to build in quality at an early stage" (UK Statistics Authority) and "…are compiled from new data sources and methods" (Eurostat). As a matter of fact, statistical matching applications comply to all the different aspects of experimental statistics shown before, and both the ONS and Eurostat web-sites on experimental statistics already include a section that contains data on income and expenditures obtained by statistical matching applications. The Istat web-site section on experimental statistics could also appear by the end of 2018, and could potentially include data obtained by statistical matching. Before submitting the already obtained data to the Istat experimental statistics web-site, we believe it is extremely important to discuss preliminarily all the pros and cons of a publication of micro and/or macro data obtained by statistical matching, as well as to pinpoint all the methodological limits of this kind of data. The results of this discussion will be an extremely important source of information to give to potential users of this data, and potentially should complement the dissemination of this kind of data in the experimental statistics web-site. This is a list of important issues to take in mind before reading the following paragraphs.

1. Reliable joint observations on household income and expenditures are not currently available. In order to avoid a heavy respondent burden, two distinct and independent sample surveys are conducted. Although in principle a household can potentially be observed in both surveys, this should be considered as an extremely rare event and the set of linked records observed in the two files that refer to the same households, if non-empty, would never be a representative sample.

2. Economists are eager of a sample where joint information on household income and expenditures can be treated by traditional statistical methods.
3. Usually, traditional statistical matching applications construct a file by imputing data from one sample to the other by means of common variables, imposing a specific model to the variables of interest. For instance, imputing expenditures on EU-SILC would reproduce the independence of income and expenditures given the matching variables. Note that the model is not justified by data but imposed by the imputers, hence it should be considered as not appropriate for any kind of statistical inferences on the joint relationship between income and expenditures. This kind of limit seems to affect data already disseminated as experimental statistics on the ONS and Eurostat experimental statistics web-sites.
4. The conditional independence assumption is just one of the relationship models between income and expenditures that are compatible with the observed data samples: the wider the set of models that are compatible with the data at hand, the most uncertain is the actual relationship model between income and expenditures. Uncertainty measures, usually based on Fréchet inequalities and bounds, detect how large is this uncertainty and in case it is possible to introduce constraints and restrictions, how effective are these restrictions in reducing uncertainty. Hence, if one imputed data set needs to be provided it should be important to declare if that data set can be considered valid for inferences because uncertainty is limited.
5. There is one case when the conditional independence assumption between income and expenditures given the matching variables is appropriate. If there exists a variable X* perfectly correlated to either income or expenditures, imputation of the correlated variable given X* becomes a constant and consequently conditional independence is respected. In the statistical matching approach defined in the next paragraphs we will use information on household income observed on HBS (highly unreliable and affected by underreporting) as a source of information for developing X*: X* will be based just on the orderings of the households according to the reported income, instead of the actual unreliable observed variables. The hypothesis imposed in this case is that, although income investigated with just one question is usually underreported, poor income families will still be represented by low income values while richer income households will report higher values. This is still an untestable assumption, although it can be justified by the fact that the shape of the income distribution in the two data sources is very similar but shifted. As a matter of fact, correlation between household income and X* will not be perfectly 1, especially when X* is reconstructed as a categorical variable. Anyway it can be declared as an approximation.

1. Introduction

The growing demand to provide data for measuring households' economic well-being at the micro level is encouraging the production of integrated statistics on household income, consumption and wealth (ICW). Nonetheless this goal represents a major challenge for NSIs as the setting up of new surveys is almost unfeasible because of budget constraints and a significant reporting burden on respondents. The redesign of the European social statistics framework towards a better integrated system of social surveys looks at the integration methodologies as a good opportunity to enhance the potential information of the existing data sources. In this context the statistical matching techniques are essentially recognized as a valid tool for producing statistics on variables not jointly collected in a single survey. For this aim, a project of integration of household income, consumption and wealth started in 2012 in ISTAT, as part of the national

modernization program of social statistics and in response to Eurostat invitation to share the best practices in the field of statistical matching. The work done so far in Istat has focused on providing an integrated data set containing joint information on household income and consumption. We used data observed on the EU Statistics on Income and Living Condition (EU-SILC: abbreviated as SILC) 2014, with income reference year 2013, and Household Budget Survey (HBS) 2013. In the future, the next step will be the strengthening of the integrated SILC data with information on household wealth, by imputing assets and debts from Bank of Italy survey. For providing a fused data set on income and consumption, different statistical matching methods were tested. Both the more traditional statistical matching techniques (nonparametric imputation; see [12]) that the most innovative methods dealing with data of complex surveys were investigated [11]. In this paper we present the updated methodological choices made, with the latest data available, and the main achieved results. One of our main research goals was to improve the matching outputs of traditional statistical matching techniques by appropriately assessing implicit underlying assumptions. Moreover, how important is to exploit all the available information in order to improve the accuracy of the final estimates provided by integrated data sets is also underlined. It should be noted that a greater and effective use of the matching techniques is actually limited by the current state of harmonization of SILC, HBS and other important social surveys. For this purpose, Istat has set up an extensive process of harmonizing national social surveys, as part of the modernization of the social statistics framework promoted at European level. It was a strong effort that clearly went beyond the core social variables and concerned the statistical units, the concepts, the definitions and the classifications of all the common variables of HBS and SILC, in order to fulfil those pre-conditions essential for data matching purposes and micro-integration.

2. Statistical matching: some pre-conditions

Statistical matching (hereafter denoted as SM) or data fusion procedures usually refer to a broad range of techniques that generally aim to investigate the relationship between variable not jointly observed in single data source. In some cases such objective is achieved by creating a micro data file by integrating the available data sources in absence of units' identifiers. The integration is performed by exploiting the relevant information shared by the data sources at hand, i.e. a suitable subset of the common variables (usually denoted as matching variables). Unfortunately, this way of working implies a strong assumption: the conditional independence (CI) of the target variables given the common variables used as matching variables. This assumption is very limiting and rarely holds in practice. It is worth noting that most of the SM methods (see for instance [10]) assume that the observations in the available samples are independently and identically distributed (i.i.d.). This additional constraint is also difficult to be maintained when matching data arising from complex sample surveys. In such a case, relatively few SM can be applied [8], the one suggested by Renssen [11], based on the calibration of the survey weights, seems particularly

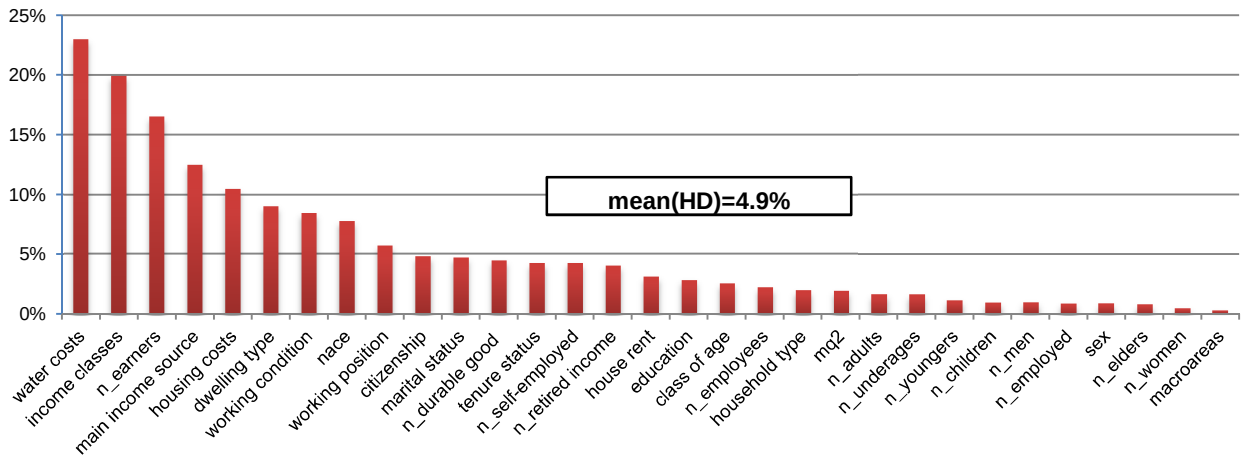
suited to deal with data from complex sample surveys carried out on households. In case no assumption are considered (i.e. CI), the estimation of the joint distribution of the target variables income and expenditures, given the common variables in the two surveys, as well as of their association parameters, consists of sets of equally plausible distributions (or parameter values). This is perfectly described in case a contingency table should be estimated, but only estimates of the marginal distributions are available, so that cells inside the table can accept any values in the so called Fréchet bounds. This approach, described firstly in [10], has been developed lately in a number of papers that seem promising [2], [1]. The approach based on the Fréchet bounds, also called uncertainty bounds in statistical matching, given to the additional source of uncertainty due to the absence of joint observations on the target variables, proves useful for validation purposes, as shown in Section 3. Integration of data via SM requires performing a series of preliminary activities which may result time-consuming: (i) check whether the target population underlying each survey is the same; (ii) identify the variables shared by both the data sources (usually denoted as X); (iii) select the subset of the common variables to be used in the matching (matching variables). Then the chosen SM method can be applied and the corresponding results can be evaluated. Step (i) and (ii) are very important; in fact statistical matching procedures largely depend both on the quality and coherence of the surveys, as well as on the common variables. Unfortunately, inconsistencies in data sources that need to be harmonized and reconciled can only be partially overcome by an ex–post integration technique. These incoherencies can arise at different levels of the statistical process and, basically, depend on differences in data collection (e.g. dissimilar definitions and different variables measuring comparable concepts) and in survey methods (e.g. sampling design, weighting, calibration, and treatment of missing values). In our SM experiments, the use of the matching methods was limited by state of harmonization of SILC, HBS and other important social surveys. A better integration and coordination of surveys requires a preliminary harmonization of common variables, statistical units and concepts. This need of an ex-ante identification and incorporation, at the design stage, of some pre-conditions of micro integration is rapidly spreading. An essential point in the success of ex-post matching procedures is the existence of a set of common variables in different data sources that are homogeneous in their statistical content. EU-SILC and HBS show a large number of common variables, mostly related to demographics, household composition, dwelling, labour, income, whose quality and coherence are in general quite good. Nonetheless in our work, the selection and harmonization of the common variables has resulted in an intense phase of reconciliation of classifications and definition of units, with a re-coding of several variables in order to have the same degree of detail [5]. However this step will become easier when the data of the renewed Italian HBS will be available. As regard consumption and income, HBS could be an appropriate source of information but the quality and coherence of income data (few questions on net monthly income at household level) are not comparable with EU-SILC, where income is more extensively and better collected. Nonetheless this rough information

on income should not be completely disregarded as far SM application is concerned, as it will be shown in the next section.

3. Statistical matching of HBS and SILC

In order to apply a statistical matching procedure it is necessary to choose a set of common variables that can be comparable. The common variables need to be harmonized across the two datasets by comparing their definitions and afterwards by harmonizing the corresponding answer categories, when different. As a measure of coherence, given that all the variables are categorical the Hellinger Distance (HD) has been used for analyzing the dissimilarity of the estimated distributions (using survey weights) across the two data sets (Figure 3.1). The water costs, the income classes and the number of earners present the highest values of HD. The large discrepancy of income classes and number of earners was expected since in HBS the latter variables have not the same quality and level of detail as in SILC. Marginal distributions which have HD distance below 5% (the chosen arbitrary threshold) are considered roughly coherent.

Figure 3.1 - Hellinger distance of the common variables (values in percentage)



2.1. Selection of the matching variables

In performing SM, not all the common variables will be used but just the most relevant ones. The selection of the most relevant Xs, the so called matching variables, should be performed by consulting subject matter experts and through appropriate statistical methods. This decision could influence the results of the matching exercise; parsimony should be the guiding principle given that choosing too many matching variables can introduce undesired additional noise in the final results. The method used for choosing the matching variables is a compromise between intersection and union of those variables that are good predictors of the target variables, respectively the monthly household income in SILC (Y) and the monthly household consumption expenditure in HBS (Z); both transformed using logarithm. The common variables that are more statistically significant in jointly explaining both the response variables are: (i) the

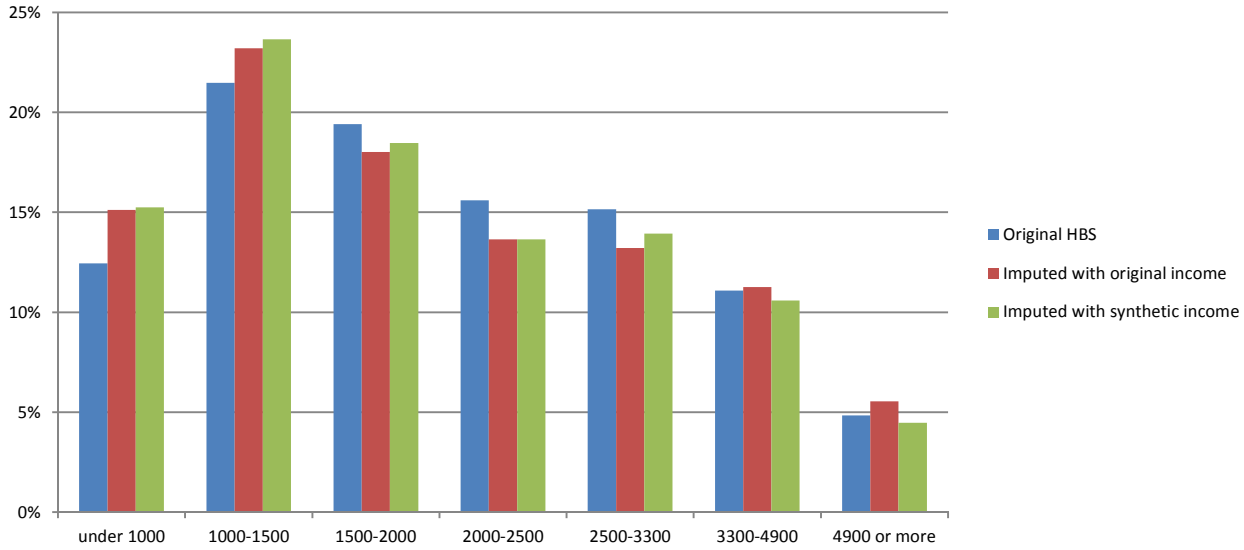
macroareas (North-East, North-West, Centre, South, Islands) and, (ii) the number of durable goods owned by the family. The choice of this subset of variables is further confirmed by the analysis related to exploration of uncertainty; in fact, [9] show that this approach represents also a powerful method for selecting the matching variables.

2.2. Selection of the matching methods

A first step in our matching procedure consisted in applying random hot deck under CI assumption, with SILC playing the role of recipient (data set A), while HBS is the donor (B) (computation performed in R using the package "StatMatch" [6]). The donation classes are defined by crossing macro-areas with the number of durable goods; in practice for a SILC household in a given donor class it is imputed the consumption category observed on a randomly chosen donor household in the same HBS donor class. In a second phase, SM uncertainty approach was investigated by calculating the Fréchet bounds for the contingency table crossing income and consumption categories (see [4]). The results confirmed that the uncertainty associated with the estimation of the contingency table $Y \times Z$ is high; this is well shown by the high estimate of the average widths of the bounds for $Y \times Z$ cells' probabilities (see Table A in Annex of [4]). In practice, given the chosen matching variables, CI of income and consumption cannot be considered a valid assumption. Usually, CI can be maintained if one of the matching variable is a close proxy of one of the target variables (income and consumption). In our application HBS, which represents the source of reliable consumption variables (Z), provides also some rough and incomplete income information (Italian HBS income data are not usually available to external users). Therefore the possibility of using such information in the matching process is crucial. However the completeness and reliability of HBS income data (few questions on net monthly income at household level) are not comparable with those of SILC, moreover the marginal distributions are not coherent. Instead of using this information in its original version, we decided to create in HBS a derived categorical income variable being comparable with the SILC one by exploiting all the HBS information concerning the use of household income (e.g. consumption and saving). This new synthetic HBS income variable has been used in SM as one of the matching variables. As stated before, given that this income related variable is highly correlated with one of the target variables (household income), the CI assumption becomes a more realistic hypothesis not far from truth. In practice we performed again the random hot deck procedure using geographical macro-area and number of owned durable goods to create the donation classes, then the categorized income variable (derived in HBS and truly observed in SILC) was used to further restrict the subset of potential donors. In particular for each SILC household, the potential donors in the HBS household were those living in the same macro area, having the same number of durable goods, and belonging to the same or adjacent - lower or upper - class of income. As shown in figure 3.2, imputing monthly consumption classes in SILC using the new HBS derived income variable resulted in an improvement in the marginal distribution of consumption highest classes imputed in

SILC, if compared to the results of the same matching applied using the HBS original income variables (not derived).

Figure 3.2 - Comparison of consumption classes imputed with original HBS income and with synthetic income. Year 2013 (values in percentage and in Euros)



2.3. Matching data from complex sample surveys

After investigating the random hot deck imputations, we also applied the Renssen's weights calibration approach [11]. The first step consisted in the harmonization of the marginal distributions of the matching variables (e.g. macro areas, number of durable goods, class of income). In particular the weights in both the surveys were calibrated to yield the same estimate of the marginal distribution of the matching variables. The target distribution was obtained as a compromise (weighted average) between the marginal distributions estimated on the two separate data sources before the harmonization. After the harmonization, the joint distribution of income and consumption, i.e. two-way contingency table $Y \times Z$, is achieved by estimating the "synthetic two-way table". As shown by Renssen, the estimator is approximately unbiased if the CI holds. In addition, the Renssen's approach was also applied at micro level to impute consumption classes in SILC. The procedure allows estimating for each unit in HBS, the probabilities of belonging to each of the consumption classes [3]. Then the estimated probabilities were used to derive a predicted consumption category for each recipient household. Two procedures were investigated: (i) imputation of the category with the highest estimated probability and (ii) random selection of the category to impute with probabilities proportional to estimated probabilities (consumption categories with highest estimated probability have higher chances of being imputed. See [7]). Figure 3.3 shows that, as expected, method (ii) provides a marginal distribution of the imputed consumption closer to the reference one (estimated from the donor HBS) than the one provided with method (i) for all the

consumption classes. Furthermore, the imputation of the category with the highest estimated probability is not able to provide a reliable estimate of the highest class of consumption. The application of Renssen's procedure for imputing at micro level a value of expected consumption rather than the expected class of consumption was also investigated. It is a mixed procedure consisting in (a) imputing predicted values of the monthly consumption (Z) in both SILC (recipient) and HBS (donor), then (b) for each recipient unit it is imputed the observed value of Z on the closest donor in HBS, whereas the distance is computed by considering just the Z predicted values in both the data sources. Preliminary results concerning the marginal distribution of the imputed consumption are satisfactory, as shown in Figure 3.4.

Figure 3.3 - Comparison of consumption classes estimated in SILC after imputation based on the Renssen's procedure. Year 2013 (values in percentage and in Euros)

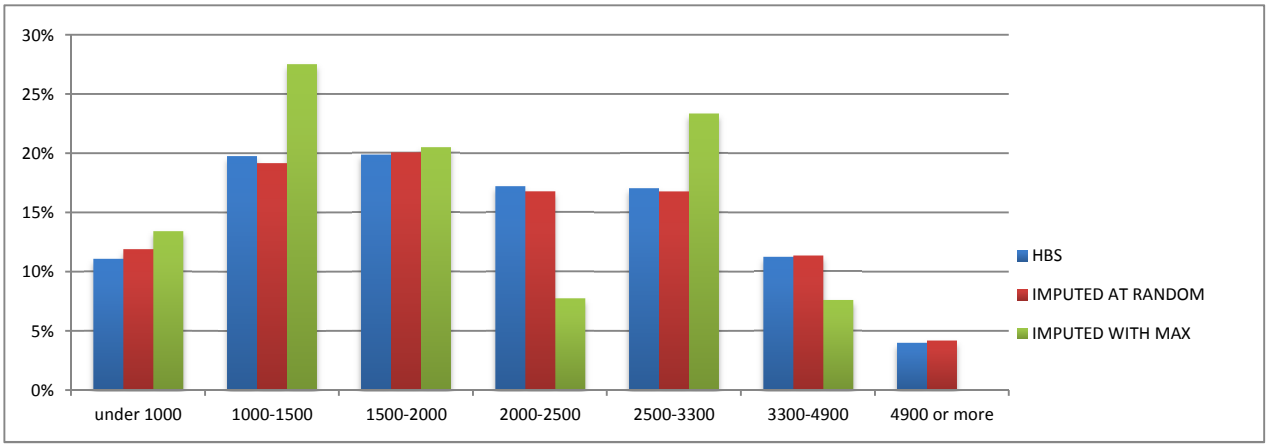
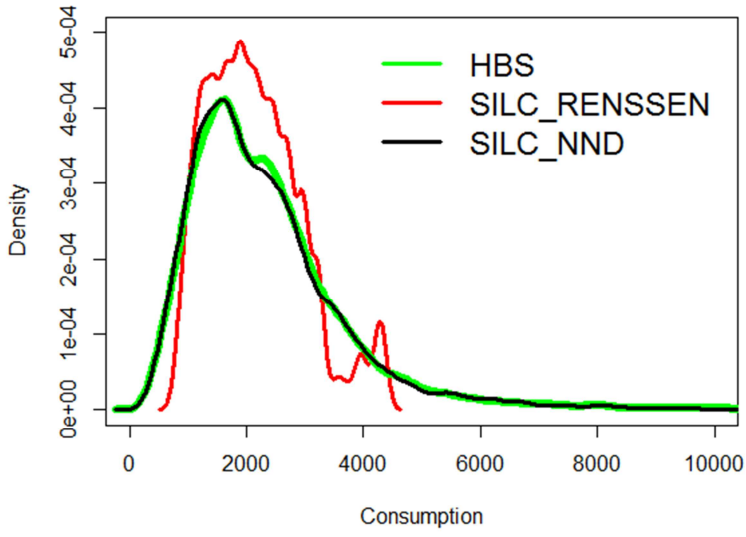


Figure 3.4 - Comparison of original HBS and imputed consumption by Renssen's procedures, applied to categorical variables (SILC_RENSSEN) and to continuous variables (SILC_NND). Year 2013



4. Conclusions

The production of official statistics on the joint distribution of income, consumption and wealth requires a sound methodological basis and a deep evaluation of the pre-requisites in terms of consistency of the surveys and the common variables necessary for the successful application of the statistical matching techniques. The current process of modernization of social surveys at European level is going towards a better integration and coordination of surveys which is expected to facilitate the matching process. In this contest an ex-ante harmonization of common variables, statistical units and concepts in SILC and HBS could effectively enhance the application of matching techniques and could simplify the estimation of parameters or indicators on the joint distribution of variables of interest. Istat has carried out a deep harmonization process of the social surveys that basically goes beyond the core social variables. All the common variables of HBS and SILC have been harmonized and reconciled with expected positive effects on the future SM applications. After a testing phase, the Italian HBS has switched to a new consumption expenditures survey in 2014, with a first data release in 2015. The renewed HBS has been redesigned to harmonize as much as possible the common variables with SILC. Particularly attention has been paid to demographic variables, household composition, family relationship with the reference person, level of education, ILO labour status. Furthermore the information on dwelling facilities have been extended in order to get closer to those provided by SILC (so as to allow also the estimation of the imputed rent by a regression method as applied in EU-SILC). As pointed out before, in order to improve the matching results the use of available information and/or proxy variables for one of the two target concepts are essential. In our exercise the presence in HBS of some rough income information was crucial in improving the matching results; the few valuable questions about the use of the household income (e.g. consumption and savings) has allowed us to reconstruct HBS income classes and compare them with those of EU-SILC. It is clear that in the future, improving the quality of information on household monthly income collected in HBS as well as the collecting data on savings can provide precious information for data integration purposes. The current revision of the EU-SILC legal basis represents another chance for enhancing the micro integration of social surveys and for more general integration purposes. The future availability of information on consumption and wealth in the SILC module 2017 could improve markedly the quality of the matching results. For instance, at present, few consumption variables on housing costs (utilities, rents, mortgage interests, regular maintenance and repairs) are collected in SILC questionnaire and it is known that housing costs would represent shared variables with high predictive power for matching purposes. As far as matching methods are concerned, the Renssen's procedure performed very well in matching at both the macro and micro levels. It permits to account for survey design and survey weights. Moreover, it maintains coherence between distributions estimated from different sources. However, when dealing with categorical variables,

imputation at micro level should be further investigated to avoid the well-known drawbacks (e.g. negative estimated cell probabilities). These drawbacks can be overcome if the target variable is not categorized, our preliminary results in this direction seem promising. Finally approaches based on the analysis of uncertainty can be further exploited in order to use as much as possible all the information available in the data at hand. For this reason further experiments should be done by considering the recent developments in [2] that explicit study properties of uncertainty intervals in statistical matching of complex sample surveys, and in [1] that study uncertainty for continuous variables, as income and expenditure, using Gibbs sampling.

Bibliography

[1] Ahfock D., Pyneb S., Lee S.X., McLachlan G.J. (2016) Partial identification in the statistical matching problem. *Computational Statistics and Data Analysis*, 104, 79-90.

[2] Conti P. L., D. Marella, M. Scanu (2015): Statistical matching analysis for complex survey data with applications, *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2015.1112803.

[3] Donatiello G., D'Orazio M., Frattarola D., Rizzi A., Scanu M., Spaziani M. (2016), The role of the conditional independence assumption in statistically matching income and consumption, in "*Statistical Journal of the IAOS*, vol. Preprint, no. Preprint, pp. 1-9, 2016, 22 April, DOI: 10.3233/SJI-161000: http://content.iospress.com/articles/statistical-journal-of-the-iaos/sji1000.

[4] Donatiello G., M. D'Orazio, D. Frattarola, A. Rizzi, M. Scanu, M. Spaziani (2014), Statistical Matching of Income and Consumption Expenditures. *International Journal of Economic Sciences*, III (3): 50-65.

[5] Donatiello G., D. Frattarola, A. Rizzi, M. Spaziani, Statistical Matching of EU-SILC and HBS: Some Critical Issues (2014), *Proceedings of the Workshop on Best Practices for EU-SILC Revision*, Banco de Portugal, Lisbon,15 October: http://ine.pt/scripts/eu-silc2014/day1.html.

[6] D'Orazio M. (2016), *StatMatch: Statistical Matching (aka data fusion). R package version 1.2.4*., http://CRAN.R-project.org/package=StatMatch.

[7] D'Orazio M. (2012), Statistical matching when dealing with data from complex survey sampling, in *Report of WP1*. *State of the art on statistical methodologies for data integration*, ESSnet project on Data Integration: 33–37, http://www.essnet-portal.eu/sites/default/files/131/ESSnetDI_WP1_v1.32.pdf.

[8] D'Orazio M., M. Di Zio, M. Scanu (2012), Statistical Matching of Data from Complex Sample Surveys. *Proceedings of the European Conference on Quality in Official Statistics - Q2012*, 29 May–1 June, Athens, Greece.

[9] D'Orazio M. M. Di Zio, M. Scanu (2015), The use of uncertainty to choose the matching variables in statistical matching. *New Techniques And Technologies For Statistics (NTTS) Conference 2015*, Brussels 10-12 March.

[10] D'Orazio M., M. Di Zio, M. Scanu (2006), *Statistical Matching: Theory and Practice*. John Wiley & Sons, Chichester.

[11] Renssen R.H. (1998), Use of Statistical Matching Techniques in Calibration Estimation. *Survey Methodology*, 24: 171-183.

[12] 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*, 19: 59-79.

Methodological issues submitted to the attention of the Committee

Question 1. The use of the proxy income as a matching variable, would promote the imputation of income to HBS, because income as observed on EU-SILC would be nearly independent to any variable observed on HBS given the proxy income and the other matching variables. Is this use appropriate?

Question 2. In the actual application, on the contrary, expenditures are imputed on EU-SILC given the proxy income. This would imply that only the bivariate distribution on income and expenditures given the matching variables can be the object of inference. In this case, the overall EU-SILC files with imputed expenditures cannot be disseminated as a file valid for any kind of inferences.

Question 3. It is currently under investigation the use of a proxy variable of some of the household expenditures components in order to include among the matching variables also this additional term, whose correlation with the actual expenditures should improve the overall statistically matched file. In this case, the overall EU-SILC files with imputed expenditures or imputed wealth can be disseminated as a file valid for any kind of inferences

Question 4. Economists look forward a file where the analysis can be extended to a third variable: wealth. In Europe, wealth is usually determined in an additional sample (Survey on Household Income and Wealth, SHIW), that in Italy is conducted by the Bank of Italy. In order to have such a file, proxy variables of respectively income and expenditures could be used in order to impute SHIW, again an imputed file whose justification relies in the approximation of the actual relationship between wealth and respectively income and expenditures, given their matching variables, with the conditional independence assumption. Is the SHIW file imputed with income and expenditures approximately valid for the joint analysis of the trivariate distribution of households wealth, income and expenditures?

Question 5. Uncertainty is due to the lack of joint information on the pair of variables of interest and correspond to the width of the set of models that are compatible with the available data. This kind of approach does not characterize only statistical matching problems: for instance, National Statistical Institutes usually treat missing data under the validity of the untestable MAR (Missing at Random) model. Can uncertainty be a quality measure of the imputed data set under the MAR assumption? Should research on uncertainty in areas different from statistical matching useful for National Statistical purposes?