Statistical Matching of Income and Consumption Expenditures

Gabriella Donatiello, Marcello D'Orazio, Doriana Frattarola, Antony Rizzi, Mauro Scanu, Mattia Spaziani[1]

ABSTRACT[2]

The purpose of this paper is to evaluate the possibility of applying statistical matching on two different data sources to create an integrated database with detailed information on households income and consumption expenditures in Italy. The data to integrate are those of EU-SILC (EU Statistics on Income and Living Condition) 2012, with income reference year 2011, and the HBS (Household Budget Survey) 2011. This paper explores which are the matching approaches more suitable with the final objective and provides insights concerning some important steps of the integration process. In order to avoid the statistical matching under the *conditional independence assumption* (CIA) it is evaluated the usage of the available auxiliary information (household monthly income) and the main results are also presented.

Keywords: Statistical matching, Survey data integration, Income, Consumption

JEL Classification: C15, C14

Authors
Gabriella Donatiello, Italian National Institute of Statistics (ISTAT), Socio-economic Statistics Directorate, Viale Oceano Pacifico, n. 171, Rome 00144, Italy. Email: donatiel@istat.it.
Marcello D'Orazio, Italian National Institute of Statistics (ISTAT), Structural Economic Statistics on Enterprises and Institutions, International Trade and Consumer Prices Directorate, Viale Oceano Pacifico, n. 171, Rome 00144, Italy. Email: madorazi@istat.it.
Doriana Frattarola, Italian National Institute of Statistics (ISTAT), Socio-economic Statistics Directorate, Viale Oceano Pacifico, n. 171, Rome 00144, Italy. Email: frattarola@istat.it.
Antony Rizzi, Italian National Institute of Statistics (ISTAT), Socio-economic Statistics Directorate, Viale Oceano Pacifico, n. 171, Rome 00144, Italy. Email: anrizzi@istat.it.
Mauro Scanu, Italian National Institute of Statistics (ISTAT), Development of Information Systems and Corporate Products, Information Management and Quality Assessment Directorate, Via Cesare Balbo, n. 16, Rome 00184, Italy. Email: scanu@istat.it.
Mattia Spaziani, Italian National Institute of Statistics (ISTAT), Socio-economic Statistics Directorate, Viale Oceano Pacifico, n. 171, Rome 00144, Italy. Email: mspaziani@istat.it.

[1] The views expressed in this paper are solely those of the authors and do not involve the responsibility of ISTAT.
[2] This work is a revised version of the paper presented during the 9th International Academic Conference, which was organized by IISES and held in April 13-16, 2014 in Istanbul, Turkey.

1. Introduction

In recent years, there has been increasing interest in using appropriate instruments to measure household living conditions. Actually defining material living condition needs to consider the level of consumption as well as the economic resources in terms of income and wealth that enable household consumption of goods and services. Collecting information on the joint distribution of income, consumption and wealth at the micro level poses several difficulties for National Statistical Institutes. In particular, setting up a new survey is unfeasible because of budget constraints as well as a significant reporting burden on respondents given the high amount of data to be collected in a single survey. As a result a better exploitation of existing data sources becomes extremely important and statistical matching techniques could represent a valid alternative for producing statistics on the distribution of variables not jointly collected in a single survey.

This paper focuses on the application of statistical matching techniques on two different sample surveys for providing joint information on household income and consumption expenditures in Italy at the micro level. The data to integrate are those of EU-SILC (EU Statistics on Income and Living Condition) 2012, with income reference year 2011, and the HBS (Household Budget Survey) 2011. These surveys are both conducted by ISTAT. In this paper, the matching approaches more suitable with our final goal and the most important issues of the integration process are presented. It is worth noting that in our case it is not possible to perform statistical matching under the *conditional independence assumption* (CIA), i.e. independence between income and consumption given some common information in both the data sources. To avoid the CIA it is evaluated the usage of the available auxiliary information (e.g. household monthly income). In alternative, the statistical matching approach based on the exploration of the uncertainty due to the absence of joint information on households expenditures and income is considered. In order to improve the quality of the matching procedure the advantage in having a more efficient *ex-ante* data collection system as well as a better harmonization of common variables of SILC and HBS and other important social surveys is also discussed.

2. Statistical matching for providing information on household income, consumption and wealth

The growing interest for multidimensional analysis of poverty and social exclusion has shifted the attention towards more appropriate instruments of analysis and the availability of integrated statistics on households income, consumption and wealth. Studies on households living standards have traditionally focussed on the economic dimension of well-being, using either data on income or consumption expenditures. However the income or consumption single-handedly cannot fully explain the households material conditions (OECD, 2013). It is well known that low levels of income do not necessarily imply low levels of consumption as households could preserve consumption by adjusting savings or receiving cash support from relatives. Theoretical arguments also seem to favour consumption as a better measure of living standards since the consumption expenditures well reflect households long-run resources rather than current income (Meyer and Sullivan, 2011; Brewer and O'Dea, 2012). Even though the consumption of goods and services is considered a key indicator of living standards, the actual and future household consumption possibilities are mainly determined by income and wealth.

In this context, the availability of coherent and reliable data on the distribution of all the households economic resources could significantly enhance the multidimensional analysis of poverty and vulnerability. The production of integrated statistics on income, consumption and wealth could help to identify the effect of policy actions in particular on households in need and/or with different characteristics. The measurement of both income and expenditures levels could allow comparison of

consumption patterns and economic behaviors at different points in the income distribution and could also support analysis on the redistributive impact of fiscal measures.

For all these reasons there is a general consensus on the need for distributional measures of well-being as a joint function of income, consumption and wealth, but there is not yet a common framework for their joint collection and analysis. The production of integrated statistics on income, consumption and wealth in household surveys is currently among the priorities of the National Statistical Institutes (NSIs). However, the current financial constraints that do not allow the setting up of new surveys and the aim of containing the statistical burden on respondents set limits to a rapidly developing of a system of integrated micro data sets in social surveys. As a consequence a better exploitation of existing data sources turns out to be an up-to-date challenge for NSIs. The use of administrative archives for statistical purposes is a well-established practice and the combination of survey and administrative sources is considered a primary tool for obtaining relevant data on income or wealth. From this point of view, the data matching techniques are also considered as a valid alternative for producing statistics on variables not jointly collected in a single survey.

In fact statistical matching (otherwise known as data fusion, data merging or synthetic matching) usually aim to achieve a micro data file from different sources that have a set of variables in common but do not contain the same units or the same identifier. Statistical matching procedures are essentially model-based techniques that are able to get timely results with reduction of costs and response burden. Nevertheless there are several methodological issues involved in the matching process that need to be taken into account in particular for assessing the quality of the final results. At European level, as no single data source provides joint information on all the relevant variables, joint statistics on income, consumption and wealth would be mainly based on SILC, HBS, and Household Finance and Consumption Survey from European Central Bank. Eurostat has strongly encouraged Member states to develop data integration methodologies in social statistics and the dissemination of best practices. The most important aim of this strategy is the provision of a multidimensional measurement of poverty in order to complement the current European key indicators. At present two main integration techniques were identified by Eurostat: an *ex-post* data matching based on the available social surveys and an *ex-ante* collection of information on wealth/consumption in the SILC survey.

It should be noted that for applying data matching techniques strong prerequisites such as coherence of data sources and of the common variables are essential (Eurostat 2013). Survey data must also be defined consistently and collected comparably, with a better harmonization of common variables across SILC, HBS and other social surveys, not limited only to the core social variables (mainly socio-demographic variables). For this purpose an *ex-ante* collection of information has the primary goal to collect new variables in the SILC questionnaire/module, in order to have new shared survey questions on consumption/wealth which can act as "hooks" for matching purposes. The availability of new variables with high predictive power can certainly improve the quality of the results and of the whole matching process.

3. Introduction to statistical matching

Statistical matching (hereafter denoted as SM) or *data fusion* techniques have been proposed to integrate data from two surveys referred to the same target population with the objective of investigating the relationship between variables not jointly observed in a single data source. In the basic SM framework, the surveys to integrate, denoted as *A* and *B*, share a set of variables *X*, while the variable *Y* is observed only in *A*, and the variable *Z* is observed just in *B*. The final objective of SM is to explore the relationship between *Y* and *Z*. To achieve such a goal, it may be not necessary to integrate

the initial data sources at micro level if objective of the inference consists in estimating one or more parameters (correlation coefficient between *Y* and *Z*; regression coefficients, contingency table $Y \times Z$). Integration at micro level is necessary when the SM final goal consists in providing a *fused* or *synthetic* data set which contains all the variables (*X*,*Y*,*Z*). This data set can be obtained by limiting attention to a given data set (say *A*) and imputing in it the missing variables (*Z* in this case). In alternative, the synthetic data set can be derived by concatenating the two data sources and then filling in the missing variables.

Many of the techniques proposed for SM at micro level are based on methods developed for the imputation of missing values: parametric (e.g. regression imputation), nonparametric (hot deck imputation) or mixed methods (e.g. methods based on predictive mean matching).

An important issue in SM application is related to the underlying assumptions. In fact, most of the techniques proposed in literature assume (i) conditional independence (CI) of the target variables given the common variables (i.e. *Y* and *Z* are independent once conditioning on *X* variables) and, (ii) the observations in the available samples are independent and identically distributed (i.i.d.) (i.e. the sample is a simple random sample).

Conditional independence is a very limiting assumption that rarely holds in practice. It can be avoided if some auxiliary information concerning the relationship between *Y* and *Z* is available (estimates of a correlation coefficient, additional data sources observing jointly the target variables) or by approaching SM in terms of *uncertainty*.

The i.i.d. assumption on observations is also difficult to be maintained when matching data from complex sample surveys involving two or more stages of selection of the sample units. In such a case it would be better to apply SM methods that account explicitly for the sampling design and the unit weighs. The SM methods that explicitly take into account the sampling design and the corresponding sampling weights are essentially two: (a) Renssen's approach based on calibrations of the weights (Renssen 1998), and, (b) Rubin's *file concatenation* (Rubin 1986) (for major details see D'Orazio *et al.* 2010 and 2012). The Renssen's approach seems more flexible and suitable when the variables under study (*X*, *Y* and *Z*) are categorical, a typical situation in sample surveys on households. Such approach is promising in matching HBS with EU-SILC, but further investigations are needed in order to derive valuable results.

3.1 Hot deck procedures for statistical matching

Hot deck procedures are frequently encountered in SM applications when the objective is the creation of the synthetic data set. The reason relies in their simplicity and the possibility of using software developed for imputing missing values. In fact they consist in filling in the missing variable in the data set chosen as the *recipient* by using the other data set as the *donor*. In practice, if *A* is the recipient, then it will be imputed with the values of *Z* selected from the other data set *B*, which plays the role of donor. The donation is typically based on the variables, *X*, available in both the data sets.

Commonly encountered hot deck procedures for SM are: *random hot deck*, *nearest neighbor hot deck* and *rank hot deck* (see Section 2.4 in D'Orazio *et al.*, 2006; Singh *et al.*, 1993).

In both nearest neighbor and rank hot deck, the donor selected in file *B* is the closest with respect to the recipient record in *A*. In the first case (nearest neighbor), distance is computed on a proper subset of the *X* variables, called the *matching variables*. The second case (rank hot deck) is appropriate only when *X* is a univariate continuous variable but, before computing distance, the values of *X* are substituted by the corresponding percentage points of the empirical cumulative distribution function. In both cases donor selection can be limited to a given subset of the available donors, typically those having the same characteristics (gender, geographic area, etc.) of the recipient unit. Moreover, the selection of the donor

can be constrained in order to avoid a donor unit to be selected more than once (see. D'Orazio *et al.*, 2006, pp. 42-43). The constrained approach requires a higher computational effort, but it ensures a better preservation of the marginal distribution of the variable being imputed. The main problem with the distance based methods, consists in selecting the variable involved in the computation on the distance (just one for rank hot deck), whereas choosing too many matching variables may result in poor matching results.

In random hot deck, the donors are chosen at random, but this choice is usually carried out within opportune subsets of donors: those sharing the same characteristics of the recipient records, e.g. in terms of geographical area, gender, typology, etc. This method is particularly suited when dealing with categorical *X* variables; in such case random hot deck consists in estimating the conditional distribution of *Z* given *X* and then drawing an observation from it.

It is worth noting that random hot deck can still be applied when dealing with a single continuous *X* variable. In fact, as suggested by D'Orazio (2013), it is possible to select at random units in the subset of the closest records of the recipient, according to criteria specified by the researcher concerning the given *X* variable, for instance $|x_{rec} - x_{don}| \leq d_0$ where the distance threshold d_0 is fixed a priori by the researcher (other possible criteria are suggested in the help pages of the package StatMatch for the R environment; D'Orazio 2013).

Finally, in random hotdeck the donors can be selected with probability proportional to weights associated to the donors (*weighted random hotdeck*) as highlighted by Andridge and Little (2009) in the context of imputing missing values in a survey. D'Orazio *et al.* (2012) explored such use of the weights in the context of SM by comparing several naive procedures, the first results show that rank and random hot deck tend to perform quite well in terms of preservation in the synthetic data set of the marginal distribution of the imputed variable *Z*.

3.2 Uncertainty approach to statistical matching

If knowledge is restricted to the conditional distributions of *Y* given *X* and *Z* given *X*, the statistical association between *Y* and *Z* given *X* is actually inestimable. The available sources of information can only derive intervals of compatibility of association parameters of $YZ|X$ with what has been actually observed: $Y|X$ and $Z|X$. If the variables are categorical, the association parameters to consider are the probability distribution of the contingency table $YZ|X$ (D'Orazio et al, 2006): the probability intervals that declare how uncertain is the association $YZ|X$ are defined by the Fréchet bounds, i.e.

$$max(0; \theta_{y.|x} + \theta_{.z|x} - 1) \leq \theta_{yz|x} \leq \min(\theta_{y.|x};\ \theta_{.z|x})$$

where θ stands for probability and the dot represents marginalization with respect to one of the variables. If these intervals are sufficiently narrow, it is possible to use one of the distributions in the class given by the Fréchet bounds as a representative: this distribution can be used for imputation purposes if a complete data set on *X, Y, Z* is sought. When dealing with numerical variables as *Y* and *Z*, one possibility is to extend the use of the Fréchet bounds to empirical cumulative distribution functions, as in Conti et al (2012). This last paper tackles also the problem to define unique measures of uncertainty for the whole distribution, as well as to consider the case of structural zeros. In this context, this kind of constraint can turn out useful in order to detect what happens to subsets of population corresponding to the vast majority of the population, once a negligible number of statistical units with

rare Y and Z associations are deleted. This allows to see how uncertainty reduces imposing that some areas of the real plane become impossible for $YZ|X$.

4. Survey data integration of income and consumption expenditures

The most important aim of our exercise is to enhance the EU-SILC data on income and social exclusion with consumption information derived from HBS survey, through the statistical matching techniques more appropriate with our final goal. We used HBS 2011 as a donor data set and we imputed consumption expenditures classes in SILC 2012 (with income reference year 2011) in order to obtain a synthetic micro data set. It should be noted that the integration process mainly consists of several steps which can be summarized as follows: (i) preliminary analysis of the data sets; (ii) reconciliation of the data sets through the harmonization of definition and classification; (iii) selection of the matching variables; (iv) selection of the matching methods more suitable with the final objective; (v) quality assessment of the results. In the following paragraphs the most important aspects of the matching procedure are concisely highlighted.

4.1 Analysis of the data sets

In Italy, both the HBS and EU-SILC surveys are carried out by ISTAT and cover the same population (private households). These two major social surveys are based on a two-stage simple random sampling design. The primary sampling units (PSU) are the municipalities and the second stage units (SSU) are the households. Inside each administrative region, the PSU are stratified according to their demographic size and, in order to guarantee self-weighting design in each region, the total of residents in each stratum is approximately constant. The evaluation of frequency distributions (weighted and non-weighted) of the variables in both datasets proved that keeping the respective weights of the two surveys is rather suitable (Table 4.1), although additional analyses for dealing with the treatment and harmonization of survey weights are also considered.

Table 4.1 Comparison between HBS and EU-SILC by sample size and population size

	Household level		Individual level	
	HBS	**EU-SILC**	**HBS**	**EU-SILC**
Sample size	23,158	19,578	57,613	47,365
Population size	25,165,002	25,429,176	60,286,784	60,797,109

4.2 Harmonization of the datasets

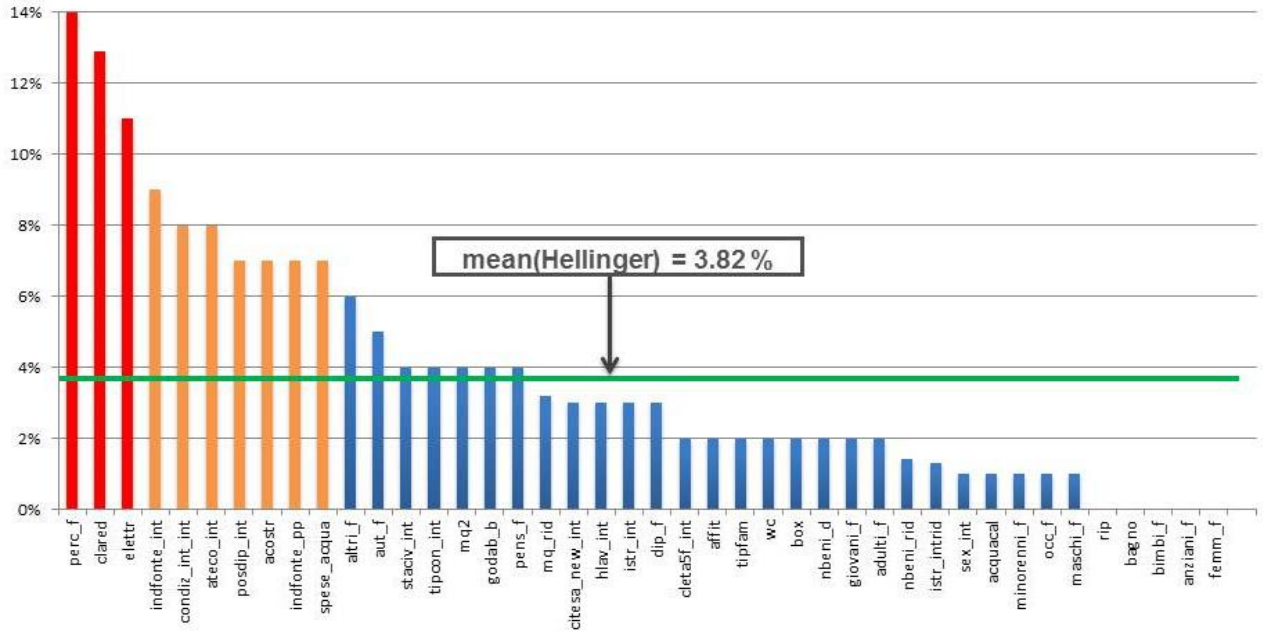
In order to apply a statistical matching procedure it is necessary to select a set of common variables that have to be comparable. At the outset the common variables need to be harmonized across the two datasets by comparing the definitions in the two surveys and afterwards by harmonizing the definitions and classifications in such a way as to make them homogeneous (D'Orazio *et al* 2006). It is known that the existence of common variables in different sources can help to improve the final estimations and relax the underlying assumptions inherent the matching process. The SILC and HBS surveys show a large number of common variables whose quality and coherence are in general quite good. However the harmonizing 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. The analyses are done at household level and in some cases the variables are aggregated from the individual level. The selected common variables are shown in table 4.2.
As a measure of coherence of the common variables, the Hellinger Distance (HD) has been used for analyzing the similarity/dissimilarity of the variables distributions across the two data sets. Figure 4.1 shows the HD of the common variables and the monthly household income, number of earners and costs for electricity finally present the highest values of HD. The dissimilarity of electricity expenditures is primarily due to differences in the way the questions are collected in the two surveys, while the large discrepancy of income variable and number of earners is quite foreseeable. In HBS the latter variables have not the same quality and level of detail as in SILC and for that reason these variables are not included in the next phase of selection of matching variables. Marginal distributions which have HD distance below 5% (the chosen arbitrary threshold) are considered coherent. Values of HD greater than 5% and lower than 10% refer to the followings: main income source, main activity, professional status, NACE code, year of dwelling construction and water expenditures. Nonetheless these variables have been included in the next step of the analysis as the relative frequencies are in the 0-5% range.

Table 4.2 Selected common variables HBS - EU-SILC

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

Figure 4.1 - Hellinger distance of the common variables



4.3 Selection of the matching variables

The phase of selection of matching variables is one of the most important as it could influence the results of the matching exercise, as a consequence the modelling techniques and multivariate analysis are frequently applied in this step. In our case the method used for choosing the matching variables from the set of common variables is, according to D'Orazio *et al* (2006), a compromise between intersection and union of those variables that are statistically significant in explaining variations in both income and consumption and which can behave as good predictors. As a result we have selected as response variables the monthly household income in SILC and the monthly household consumption expenditure in HBS. In order to compress the deviation from the mean both variables are expressed in logarithm. The matching variables can be chosen by applying different methods (parametric and non-parametric, regression, etc.) Then the selection is finally made by looking at variables that are more statistically significant in jointly explaining the response variable. In our case these are: (i) the macroareas (North-East, North-West, Centre, South, Islands) and, (ii) the number of durable goods owned by the family.

4.4 Matching between HBS and EU-SILC

A first step in our procedure consisted in the application of random hot deck under CIA, using the R package StatMatch (D'Orazio, 2013). Then the exploration of SM uncertainty was also applied. The random hot deck is performed by specifying the donation classes, formed by crossing the joint distribution of number of durable goods and macroareas and for each record in a given donation class, and a donor is selected at random. In the end an actual observed value for classes of consumption is imputed in to EU-SILC. The main result of this procedure is presented in Table 4.3.

Table 4.3 Income classes by classes of consumption under CIA (in euros and in percentage values)

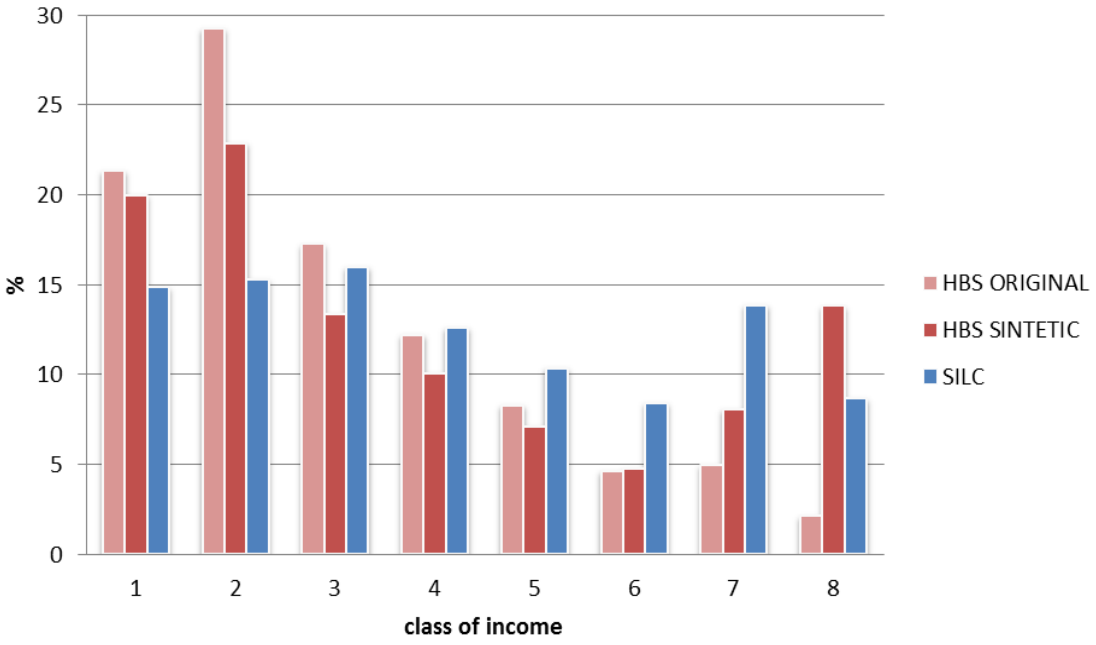
	Consumption								
Income	**under 1000**	**1000-1500**	**1500-2000**	**2000-2600**	**2600-3100**	**3100-3600**	**3600-5200**	**5200 or more**	**Total**
under 1000	**4.45**	4.32	3.01	1.87	1.07	0.66	0.75	0.39	16.52
1000-1500	3.26	**4.02**	3.30	2.25	1.37	0.91	1.09	0.58	16.79
1500-2000	2.12	3.37	**3.29**	2.52	1.64	1.13	1.46	0.81	16.34
2000-2600	1.11	2.25	2.52	**2.11**	1.47	1.04	1.42	0.83	12.76
2600-3100	0.57	1.45	1.89	1.73	**1.29**	0.93	1.32	0.80	9.97
3100-3600	0.41	1.07	1.47	1.39	1.07	**0.79**	1.14	0.70	8.04
3600-5200	0.43	1.39	2.18	2.23	1.79	1.34	**2.02**	1.28	12.66
5200 or more	0.16	0.59	1.07	1.21	1.03	0.79	1.25	**0.82**	6.93
Total	12.51	18.46	18.74	15.30	10.73	7.60	10.45	6.21	100

The CIA cannot be verified from the matched datasets and it is clearly an unsatisfactory model for expenditures and income whatever the conditioning variables are. This conclusion is confirmed by the uncertainty analysis carried out by calculating the Fréchet bounds for the contingency table between the variables of interest given the two common variables being considered. For this goal, a first step is to harmonize the joint distribution (Renssen, 1998) of the matching variables (macroareas and number of durable goods). The harmonized weights are then applied to the computation of Fréchet bounds: the lower and upper bounds for the contingency table on the categorized income and consumption classes are shown in Annex Table A.

Any distributions whose cell frequencies are inside the extrema are compatible with the available information and can be a valid inference. For instance, the frequency of the cell corresponding to the first classes of expenditures and income should lay between 0.92% and 11.69% (Annex Table A). The distribution under the CIA (Table 4.3) is in between the extrema (Annex Table A). In general the average width of the uncertainty bounds is 7.8%, that seems too wide given the phenomena being studied.

Some previous works on SM techniques applied to social surveys (Coli *et al* 2005) have highlighted the importance of using the household income variable as auxiliary information in order to relax the CIA and improve the estimation of the correlation structure. Nonetheless the income variable is clearly collected with different levels of detail in HBS and SILC: in SILC the household income is a sum of each type of income collected at individual level from family members with more than 16 years old. In HBS income is observed at household level from a multi-response variable that is included in an *ad hoc* section about income and savings. In order to reduce the large income discrepancy in the two surveys, the additional information from the HBS income section has been used in order to estimate a new income variable which to some extent has decreased the household income's underestimation in HBS (Figure 4.2).

Figure 4.2 Comparison of HBS and EU-SILC income classes



After that the new HBS variable has been used among the selected matching variables to perform the SM procedures. In our case the auxiliary information can be represented by the approximation of the actual income/expenditure relationship. In particular, the auxiliary information concerning the reconstructed income is used in the random hot deck procedure in further restricting the subset of potential donors. In this application, the subset of potential donors are the ones living in the same macroarea and having the same number of durable goods whose distance from the recipient is in the neighbourhood of size 1: this means that the donors could be chosen in the same or in the upper/lower class of income.

Table 4.4 shows the joint distribution of variables of interest obtained under this approach. It is worth noting the result of this constraint in the hot deck procedure: the number of households with consumption greater than income decreases versus that under CIA assumption (from 40,07% to 34,33%). Moreover, the sum of frequencies lying on the diagonal is greater than the same under CIA (25,14% RnD vs 18,79% CIA). Looking at row frequencies (Annex Table B and Table C) it would appear that unlikely assignments between classes of consumption and income is limited. More specifically, there is a significant decrease of those frequencies corresponding to classes of consumption that differ more than three from the respective class of income.

Looking more closely at the impact of the imputations on other variables not selected as matching variables, it should be noted that the household typology presents a similar distribution to the original in HBS (Table 4.5). We believe that this is a very promising starting point for the distributional analysis on the propensity to consume by main socio-demographic variables such as household type.

Table 4.4 Income classes by classes of consumption - Random hot deck (in euros and in percentage values)

	Consumption								
Income	**under 1000**	**1000-1500**	**1500-2000**	**2000-2600**	**2600-3100**	**3100-3600**	**3600-5200**	**5200 or more**	**Total**
under 1000	**5.64**	4.63	2.89	1.43	0.92	0.43	0.46	0.20	16.59
1000-1500	3.66	**5.01**	3.45	1.98	1.06	0.63	0.67	0.33	16.79
1500-2000	1.54	4.47	**4.37**	2.40	1.21	0.79	0.93	0.56	16.29
2000-2600	0.43	2.61	3.48	**2.59**	1.39	0.76	0.95	0.51	12.70
2600-3100	0.16	1.06	2.29	2.48	**1.92**	0.79	0.86	0.39	9.94
3100-3600	0.05	0.40	1.40	1.86	1.79	**0.99**	1.03	0.51	8.04
3600-5200	0.05	0.34	1.21	1.77	2.00	2.09	**3.08**	2.16	12.70
5200 or more	0.01	0.10	0.38	0.69	0.97	1.12	2.13	**1.55**	6.95
Total	11.53	18.63	19.47	15.19	11.27	7.59	10.11	6.21	100

Table 4.5 Comparison of HBS and Imputed consumption classes by household typology - Random hot deck (in euro and in percentage values)

		Consumption	**under 1000**	**1000-1500**	**1500-2000**	**2000-2600**	**2600-3100**	**3100-3600**	**3600-5200**	**5200 or more**	**Total**
Household typology	**Single member under 35**	**Hbs**	5.4	5	4.7	2.7	2.2	2.2	1.8	1.7	**3.5**
		Imputed	6.3	7.1	5.3	5.1	3.3	3.8	3	2.7	**5.1**
	Single member 35-64	**Hbs**	18.3	18.7	16.6	14.9	11.3	8.9	8	9	**14.3**
		Imputed	17.5	15.1	13.4	10.3	8.9	7	9.8	7.2	**12.2**
	Single member 65 and over	**Hbs**	45.3	25.8	15	8.8	5.4	6.7	4.1	3.8	**16**
		Imputed	40.6	22.7	12.6	7.9	6.4	4.6	3.8	2.8	**15.2**
	Couple with r.p. (a) under 35	**Hbs**	0.7	1.6	1.9	1.9	1.2	2.3	2.2	0.9	**1.6**
		Imputed	0.4	1.4	2.1	2.8	1.9	3.4	3.2	1.3	**2**
	Couple with r.p. 35-64	**Hbs**	2.3	4.9	7.4	8.5	9.4	9.7	7.6	5.6	**6.8**
		Imputed	3.1	4.2	6.3	6.3	6.6	7.5	8.7	9.6	**6**
	Couple with r.p. 65 and over	**Hbs**	9.7	11.1	11.1	10.5	9.6	8.8	7	7.2	**9.8**
		Imputed	8.2	10.5	9.7	7.4	7.8	6.7	6.5	4.6	**8.3**
	Couple with 1 child	**Hbs**	4.8	10.7	14.1	16.8	21.5	20	24.8	22.5	**15.8**
		Imputed	9.2	11.7	16.7	22.5	23.1	21.1	22.1	29.4	**17.8**
	Couple with 2 children	**Hbs**	2	7.1	11.6	17.5	21.2	22.5	25.5	26.5	**15**
		Imputed	5.3	9.6	15.5	18	22	24.5	24.5	22.7	**15.8**
	Couple with 3 or more children	**Hbs**	0.9	1.4	3	3.4	4.1	4.9	7	6.8	**3.5**
		Imputed	1	3.3	4.3	3.8	2.7	5.9	4.1	4.7	**3.5**
	Single parent	**Hbs**	6.7	9	10	10.2	7.8	8.9	7.7	8.5	**8.8**
		Imputed	5.1	9.5	8.1	8.6	9	9.6	6.1	5.6	**7.9**
	Other typology	**Hbs**	3.8	4.8	4.6	4.8	6.2	5.2	4.5	7.5	**5**
		Imputed	3.1	4.9	6	7.2	8.3	5.8	8.4	9.4	**6.2**

(a) Reference person

4.5 Evaluating results of statistical matching

It is well known that the quality evaluation of the matching results is a very complex step as it involves some critical issues in measuring the accuracy and reliability of the final estimates. A first basic quality check consists of the comparison of the marginal and joint distributions for imputed variables between the donor and the recipient data set. Figure 4.3, shows the HD of the joint distribution of the imputed and observed consumption classes in HBS by some variables of interests for the analysis. It is important to notice that household typology obtained by random hot deck preserves an adequate HD value of 5%. The distributions of class of consumption imputed and observed, as shown in Figure 4.4, are very similar with an HD of 0.6%.

Figure 4.3 Comparison of the joint distribution of imputed and observed consumption classes by matching variables and by household type (HD value)

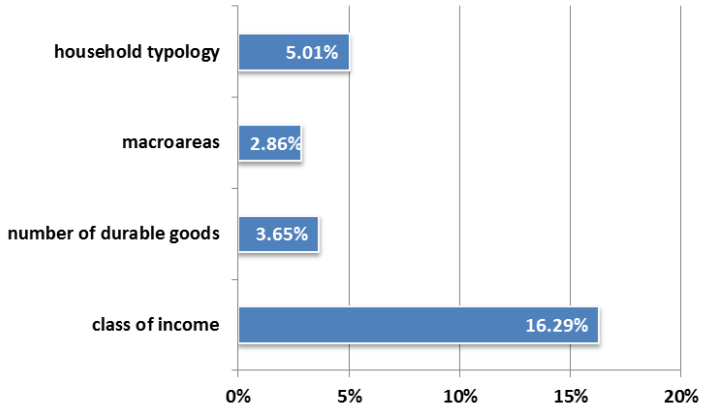
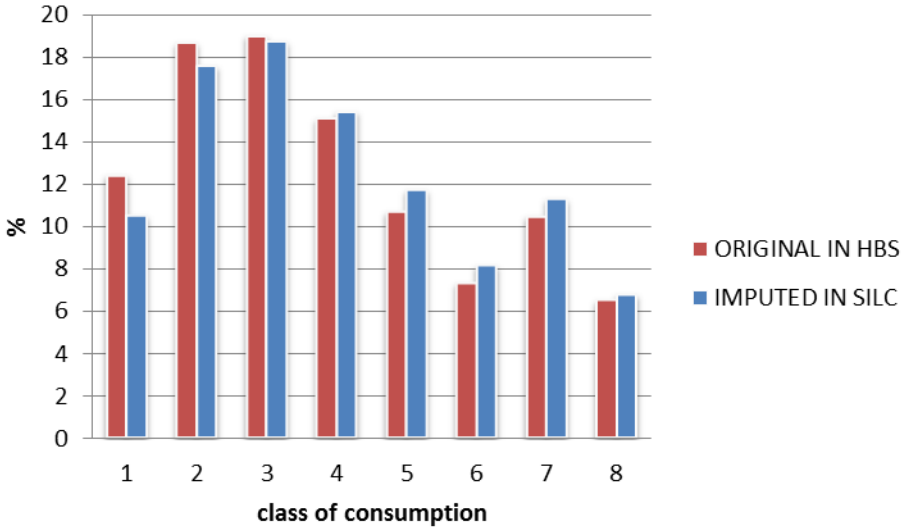


Table 4.6 Final correlation between class of income and class of consumption in observed data (HBS) and in imputed data

Pearson		Spearman	
Hbs	**Imputed**	**Hbs**	**Imputed**
0.5865	0.5840	0.5900	0.5874

Figure 4.4 Comparison of HBS observed consumption classes and EU-SILC imputed consumption classes



5. Some concluding remarks

The production of integrated statistics on household income, consumption and wealth mostly answers to the growing demand to provide data for measuring households economic well-being at the micro level. The National Statistical Institutes generally collect data through the household surveys, frequently using the linking with administrative sources and statistical matching techniques are increasingly exploited as an additional tool for combining data from different sources. However in order to improve the accuracy and consistency of integrated data sets and the general quality of matching results many data requirement are needed. The coherence of data sources and of the common variables which can be used for data matching purposes certainly plays an important role. A better harmonization across SILC, HBS and other important social surveys, not limited only to the core social variables (socio-demographics variables), seems no longer deferrable. At European level, an *ex-ante* collection of new variables in SILC questionnaire/module also has the important aim to address analytical needs based on the joint distribution of income, consumption and wealth.

Our exercises on integration of income and consumption expenditures in Italy are in fact a work in progress. Nonetheless the preliminary results are quite satisfactory. The random hot deck method seems to perform quite well, whereas an exercise based on Renssen's approach has shown some of the well-known drawbacks and needs further examination. In the same way a more in-depth analysis of the uncertainty and how to reduce it, for example by introducing a kind of constraint such as structural zero, will certainly improve our final estimates of the integrated data set.

References

Andridge R.R., Little R.J.A. (2009) "The Use of Sample Weights in Hot Deck Imputation". *Journal of Official Statistics*, No. 25, pp. 21-36.

Brewer, M., & O'Dea, C. (2012), *Measuring living standards with income and consumption: evidence from the UK*. Retrieved June 23, 2012, from Institute for Social & Economic Research (ISER) - University of Essex: https://www.iser.essex.ac.uk/publications/working-papers/iser/2012-05.pdf.

Coli, A., F. Tartamella, G. Sacco, I. Faiella, M. Scanu, M. D'Orazio, Di Zio, M., Siciliani, I., Colombini, S. and Masi, A. (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*", Technical Report, Working Group ISTAT- Bank of Italy, Rome.

Conti, P.L., Marella, D., Scanu, M. (2012), "Uncertainty analysis in statistical matching". *Journal of Official Statistics*, No. 28, pp. 69-88.

D'Orazio, M. (2013), *StatMatch: Statistical Matching (aka data fusion). R package version 1.2.1*. http://CRAN.R-project.org/package=StatMatch.

D'Orazio, M., Di Zio, M., Scanu, M. (2006), *Statistical Matching: Theory and Practice*. John Wiley & Sons, Chichester, ISBN: 0-470-02353-8.

D'Orazio, M., Di Zio, M., Scanu, M. (2010), "*Old and new approaches in statistical matching when samples are drawn with complex survey designs"*. (Sessione specializzata "Matching techniques, censuses and administrative data") Atti della 45ma riunione scientifica della Società Italiana di Statistica, Padova 16-18 giugno 2010.

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

Eurostat (2013), *Statistical matching: a model based approach for data integration*. Methodologies and Working Paper. Luxembourg: Publications Office of the European Union, 2013.

Meyer, B., & Sullivan, J. (2011), "Further Results on Measuring the Well-Being of the Poor Using Income and Consumption". *Canadian Journal of Economics*, Vol 44, No 1, pp. 52-87.

OECD (2013), *OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth*. Paris OECD Publishing.

Renssen, R. H. (1998), "Use of Statistical Matching Techniques in Calibration Estimation". *Survey Methodology*, No 24, pp. 171-183.

Rubin, D.B. (1986), "Statistical matching with adjusted weights and multiple imputations". *Journal of Business and Economic Statistics*, No 4, pp. 87-94.

Singh A.C., Mantel H., Kinack M., Rowe G. (1993) "Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption". *Survey Methodology*, No. 19, pp. 59-79.

ANNEX

Table A - Uncertainty bounds

Income classes	Consumption classes	Low.cx	CIA	Up.cx
1	**1**	0,0092	0,0445	0,1169
2	**1**	0,0000	0,0327	0,0963
3	**1**	0,0000	0,0212	0,0731
4	**1**	0,0000	0,0111	0,0505
5	**1**	0,0000	0,0057	0,0361
6	**1**	0,0000	0,0041	0,0291
7	**1**	0,0000	0,0043	0,0280
8	**1**	0,0000	0,0016	0,0155
1	**2**	0,0000	0,0432	0,1404
2	**2**	0,0000	0,0402	0,1588
3	**2**	0,0000	0,0337	0,1377
4	**2**	0,0000	0,0225	0,0999
5	**2**	0,0000	0,0145	0,0722
6	**2**	0,0000	0,0107	0,0579
7	**2**	0,0000	0,0139	0,0635
8	**2**	0,0000	0,0059	0,0370
1	**3**	0,0000	0,0301	0,1089
2	**3**	0,0000	0,0330	0,1357
3	**3**	0,0000	0,0329	0,1565
4	**3**	0,0000	0,0252	0,1255
5	**3**	0,0000	0,0189	0,0953
6	**3**	0,0000	0,0147	0,0764
7	**3**	0,0000	0,0218	0,0936
8	**3**	0,0000	0,0107	0,0519
1	**4**	0,0000	0,0187	0,0816
2	**4**	0,0000	0,0225	0,1042
3	**4**	0,0000	0,0252	0,1270
4	**4**	0,0000	0,0210	0,1237
5	**4**	0,0000	0,0173	0,0996
6	**4**	0,0000	0,0139	0,0804
7	**4**	0,0000	0,0223	0,1058
8	**4**	0,0000	0,0121	0,0619
1	**5**	0,0000	0,0107	0,0582
2	**5**	0,0000	0,0137	0,0734
3	**5**	0,0000	0,0164	0,0899
4	**5**	0,0000	0,0147	0,0948
5	**5**	0,0000	0,0129	0,0957
6	**5**	0,0000	0,0107	0,0800
7	**5**	0,0000	0,0179	0,0976
8	**5**	0,0000	0,0103	0,0602
1	**6**	0,0000	0,0066	0,0452
2	**6**	0,0000	0,0091	0,0576
3	**6**	0,0000	0,0113	0,0674
4	**6**	0,0000	0,0104	0,0723
5	**6**	0,0000	0,0093	0,0748
6	**6**	0,0000	0,0079	0,0733
7	**6**	0,0000	0,0134	0,0750
8	**6**	0,0000	0,0079	0,0549
1	**7**	0,0000	0,0075	0,0452
2	**7**	0,0000	0,0109	0,0592
3	**7**	0,0000	0,0146	0,0748
4	**7**	0,0000	0,0142	0,0800
5	**7**	0,0000	0,0133	0,0833
6	**7**	0,0000	0,0114	0,0745
7	**7**	0,0000	0,0202	0,1024

8	**7**	0,0000	0,0125	0,0683
1	**8**	0,0000	0,0039	0,0308
2	**8**	0,0000	0,0058	0,0390
3	**8**	0,0000	0,0081	0,0489
4	**8**	0,0000	0,0083	0,0540
5	**8**	0,0000	0,0080	0,0561
6	**8**	0,0000	0,0070	0,0559
7	**8**	0,0000	0,0128	0,0621
8	**8**	0,0000	0,0082	0,0545

Table B - Row frequencies of income classes by class of consumption (a) – CIA (in euros and in percentage values)

	Consumption (C)										
Income (Y)	**under 1000**	**1000-1500**	**1500-2000**	**2000-2600**	**2600-3100**	**3100-3600**	**3600-5200**	**5200 or more**	**Total**	**C > Y (sum)**	**Y > C (sum)**
under 1000	27.0	26.0	18.0	**11.0**	**6.0**	**4.0**	**5.0**	**2.0**	100.0	**29.0**	**-**
1000-1500	19.0	24.0	20.0	13.0	**8.0**	**5.0**	**6.0**	**3.0**	100.0	**24.0**	**-**
1500-2000	13.0	21.0	20.0	15.0	10.0	**7.0**	**9.0**	**5.0**	100.0	**21.0**	**-**
2000-2600	**9.0**	18.0	20.0	17.0	12.0	8.0	**11.0**	**7.0**	100.0	**18.0**	**9.0**
2600-3100	**6.0**	**15.0**	19.0	17.0	13.0	9.0	13.0	**8.0**	100.0	**8.0**	**20.0**
3100-3600	**5.0**	**13.0**	**18.0**	17.0	13.0	10.0	14.0	9.0	100.0	**-**	**37.0**
3600-5200	**3.0**	**11.0**	**17.0**	**18.0**	14.0	11.0	16.0	10.0	100.0	**-**	**49.0**
5200 or more	**2.0**	**8.0**	**15.0**	**18.0**	**15.0**	11.0	18.0	12.0	100.0	**-**	**59.0**

(a) Due to rounding totals in the table may not correspond to the sum of the respective components. Percentage compositions are automatically rounded, for this reason, the sum of the respective components may not be equal to 100%.

Table C - Row frequencies of income classes by class of consumption (a) - Random hot deck (in euros and in percentage values)

	Consumption (C)										
Income (Y)	**under 1000**	**1000-1500**	**1500-2000**	**2000-2600**	**2600-3100**	**3100-3600**	**3600-5200**	**5200 or more**	**Total**	**C > Y (sum)**	**Y > C (sum)**
under 1000	34.0	28.0	17.0	**9.0**	**6.0**	**3.0**	**3.0**	**1.0**	100.0	**21.0**	**-**
1000-1500	22.0	30.0	21.0	12.0	**6.0**	**4.0**	**4.0**	**2.0**	100.0	**16.0**	**-**
1500-2000	9.0	27.0	27.0	15.0	7.0	**5.0**	**6.0**	**3.0**	100.0	**14.0**	**-**
2000-2600	**3.0**	21.0	27.0	20.0	11.0	6.0	**7.0**	**4.0**	100.0	**11.0**	**3.0**
2600-3100	**2.0**	**11.0**	23.0	25.0	19.0	8.0	9.0	**4.0**	100.0	**4.0**	**12.0**
3100-3600	**1.0**	**5.0**	**17.0**	23.0	22.0	12.0	13.0	6.0	100.0	**-**	**23.0**
3600-5200	**0.0**	**3.0**	**10.0**	**14.0**	16.0	16.0	24.0	17.0	100.0	**-**	**27.0**
5200 or more	**0.0**	**1.0**	**5.0**	**10.0**	**14.0**	16.0	31.0	22.0	100.0	**-**	**31.0**

(a) Due to rounding totals in the table may not correspond to the sum of the respective components. Percentage compositions are automatically rounded, for this reason, the sum of the respective components may not be equal to 100%.