WORKING PAPER
DIPARTIMENTO DI ECONOMIA PUBBLICA

Working Paper n.164

Elena Pisano e Simone Tedeschi

Micro Data Fusion of Italian Expenditures and Incomes Surveys

Roma, Aprile 2014

SAPIENZA
UNIVERSITÀ DI ROMA

Micro Data Fusion of Italian Expenditures and Incomes Surveys

Elena Pisano[i], **Simone Tedeschi**[ii]

[i] Bank of Italy, Tax Department. The opinions expressed here are those of the author and do not necessarily reflect the positions of the Institution.
[ii] Department of Economics and Law, Sapienza University of Rome, Italy.
Contact: e-mail: simone.tedeschi@uniroma1.it
The authors would like to thank an anonymous referee and the participants to the seminar held on the 29th of January 2014 at SOSE.

Abstract

The aim of this work is to match household consumption information from *Indagine sui Consumi delle Famiglie* (Household Budget Survey, HBS) by the Italian National Statistical Institute (ISTAT) with *Indagine sui Bilanci delle Famiglie Italiane* (Survey of Households' Income and Wealth, SHIW) by the Bank of Italy for the year 2010. The work offers a review of the main matching methodologies, coupled with a discussion of the underlying hypotheses (such as the CIA) which, in our case, are less demanding to assume given the presence consumption aggregates as common variables between the two surveys. Moreover, some tests measuring the validity of the matching procedure are presented in order to check the preservation of joint distributions.

The resulting sample is expected to allow better distributional and micro-econometric analyses on consumption income and wealth (e.g. Engel curves, consumption age/income profiles). Moreover, the very detailed integrated dataset would constitute a platform for an integrated microsimulation analysis of direct, indirect and wealth tax reforms which, so far, has not been feasible taking available sample surveys separately.

Our matching achieves a good preservation of the marginal distributions of all consumption aggregates from the donor survey. However, a thorough comparison of the original distributions suggests that the HBS is a convenient donor for the imputation of non-durable commodities only. Consumption aggregates closer to the concept of wealth (such as durables and the extraordinary expenditure for dwelling maintenance) or savings (such as mortgages and private pensions) prove to be better assessed by the longer - and more issue-specific - recall of the SHIW.

As secondary outcomes, the information derived from HBS on non-durables entails an increase in the dispersion and an upward adjustment of consumption profiles in the synthetic distribution relative to SHIW. This implies also a downsized average propensity to save for the household sector which gets closer to the National Accounts figures.

JEL classification: **C81, D12, D31**

Keywords: data fusion, propensity score, household consumption, income, wealth

Introduction

Propensity score matching (PSM, Rosenbaum and Rubin, 1983 and 1984) has become a standard approach to estimate causal treatment effects. Nevertheless, recently, researchers and the main national statistical institutes have been using this technique for integrating piece of information from different micro-data sources whenever different samples cannot be exactly matched by using identifiers such as social security numbers or fiscal codes (i.e. through a proper record linkage). Data fusion techniques aim at achieving a complete data file from different sources which do not contains the same units. In this sense, data fusion can be assimilated to a problem of missing-data imputation. Statistical peers - one (or more) assuming the role of donor(s) and the other the recipient, respectively - are usually found by means of PSM-nearest neighbor or hot deck procedures which serve to synthesize the multidimensional distance/similarity in a one-dimensional space.

The aim of this work is to impute household consumption information from the *Indagine sui Consumi delle Famiglie* (Household Budget Survey, HBS henceforth) by the Italian National Statistical Institute (ISTAT) to the *Indagine sui Bilanci delle Famiglie Italiane* (Survey of Households' Income and Wealth, SHIW) by the Bank of Italy using a matching technique. More specifically, the present work combines information from the Historical Database (integrated with information from the original cross sectional files) of SHIW 2010 with the wave 2010 of HBS.

Both surveys include information on household consumption but HBS is focused on this issue by specifically providing data on single household consumption goods and services bought or self-produced by Italian families. On the opposite, only SHIW contains incomes, together with several other information on wealth and socio-demographic characteristics.

Therefore, the problem involved in this data fusion is uncommon compared to the traditional case. In our case, indeed, the information on consumption to impute to SHIW is observed, though in a less disaggregated way, also in the SHIW file itself, thus allowing to use some aggregates of consumption expenditure in the vector of common variables. In addition, providing a thinner classification of consumption aggregates, we consider HBS to deliver a more accurate representation of the true distribution of some consumption aggregates (the ones homogenous to SHIW's aggregates i.e. food, other non-durables, and to a lesser extent, durables).

Previous experiments of data fusion in Italy are recent, starting from the early 2000s. A first attempt of integration of these two sources was already tackled in 1998 for the year 1991 (Rosati, 1998). It evaluated the feasibility of statistical matching in order to set up a new combined dataset, both at a macro-level, using aggregate information concerning family types, and at the individual micro-level. A further project was conducted by ISTAT (Cimino e Coli, 1998a, b, c). Starting from that experience, a joint ISTAT-Bank of Italy working group has produced an integrated dataset for years 1991, 1993, 1995, 1998 through the use of Bayesian networks and statistical matching techniques, to be used mainly for compiling Social Accounting Matrices (SAM). The resulting database was also suitable for *meso*-level economic analyses, yet not micro (Coli et al., 2006)[3].

Our work differs from the previous ones as it aims at providing an integrated synthetic micro-dataset to jointly analyze income, wealth and consumption distributions with a high degree of

[3] Analogous study for Italy, though on – partially - different data sources are Sisto (2006), which aim at assessing the feasibility of a data fusion between Multiscope Household Survey conducted by Istat in 2010 and SHIW; Montrone et al. (2011), which aim to build an integrated archive between HBS and Eu-Silc, particularly suitable for poverty analyses.

detail for both incomes-assets and consumption expenditure items. The resulting sample is expected to allow better distributional and micro-econometric analyses on consumption income and wealth (e.g. Engel curves, consumption age/income profiles). Moreover, the very detailed integrated dataset would constitute a platform for an integrated microsimulation analysis of direct, indirect and wealth tax reforms which, so far, has not been feasible taking available sample surveys separately.

Our task is twofold: on the one hand, we aim at fusing at best SHIW households with an equal number of HBS donor units so as to impute disaggregated expenditure items to the former sample. To do this we searched for those consumption aggregates which, properly recodified or treated, show a high level of comparability between the two sources. On the other hand, we aim at building up overall synthetic measures of household consumption which borrow the best information from the two files. This asks for discarding some imputed aggregates while retaining the original SHIW ones when the analysis suggests this latter source provides a more reliable picture of the true distributions.

The work is structured as follows: in Section 1 a brief review of the data imputation methodologies is reported. Section 2 discusses the main assumptions and specificities of our matching. In section 3 the preparatory tasks of the matching procedure are illustrated, coupled with a discussion of the criticalities on the durable reporting and some comparative empirical evidence from the two surveys. Finally, in section 4 a selection of main findings of the resulting distributions on the synthetic dataset is provided, together with a discussion of tests measuring the validity of the matching procedure. Section 5 concludes. Data issues with a description of the two surveys are reported in Appendix A.

1. Matching techniques and algorithms

1.1. Propensity score matching

Traditionally, propensity score methods (PSM) serve the purpose of analyzing causal effects of treatment (e.g. policies) from observational data. To analyze such data, an ordinary least square regression model using a dichotomous indicator of treatment among the explanatory variables is probably unsuitable, because the error term is likely to be correlated with some explanatory variable. In fact, when groups are not generated by mechanisms of randomized experiments and the researcher has no control on the treatment assignment, they would probably differ on their observed and unobserved characteristics. The propensity score, defined as the conditional probability of being treated given the observed characteristics, is then used in order to reduce selection (on observables) bias in the estimation of treatment effect, balancing the covariates between the two group (treated and control) and reproducing in this way a 'quasi-randomized' experiment.

PSM is used here to achieve the goal of a 'multidimensional imputation' in terms of a large missing data problem, rather than as an instrument to estimate policy treatment effects.

If we had to impute one variable only to the SHIW sample from the HBS[4] we might think about this problem in terms of imputation of a missing information through regression. Actually, we do not aim at achieving a full integration between the two datasets to obtain a sample which is

[4] Choosing the smaller file as recipient is common practice. In our case, the fused file is supposed to be employed for an integrated microsimulation analysis of direct and indirect tax system. Thus, the reference sample must allow to carry out the analysis also at the individual level. This is possible in SHIW while it is prevented in HBS.

the sum of both. Rather, given our aims, we conceive SHIW as the *recipient* sample and HBS as the *donor* of some missing information, thus creating a synthetic file from the two. The synthetic data set is thus just the completed SHIW file, while a significant amount of records as well as important sample information on common variables is discarded from HBS. On the opposite, whether the overall sample SHIW ∪ HBS was used for inference, the effect of *matching noise*[5] would be rather magnified.

As we need to impute several variables (*i.e.* the vector of consumption items), techniques based on the estimation of a distance function seem appropriate. Propensity score (PS) method is based on the definition of a distance function that evaluates the similarity among units of two samples and provides each unit of a sample with a "similar" unit from the other sample. Such a match is made in terms of a scalar summary of the multidimensional space representing each unit (family or individual). Hence, matching procedure depends essentially on two choices:

1. the choice of the distance measure to define "similar" units ;
2. the choice of the matching typology, i.e. a criterion to assess how many units match and how, according to the chosen distance.

PS of one unit (treated or non-treated i.e. belonging either to one or the other sample) is the probability of a unit being assigned to a particular treatment group given her characteristics before the treatment:

$$p_i = Pr[T = 1|z] = \frac{1}{1 + e^{-(\beta 0 + \beta 1 z1 + \cdots + \beta kzk)}}$$

Therefore, the matching procedure[6] first runs a logistic (or a probit) regression where the dependent variable (shiw) is equal to 1 if the observation comes from the recipient sample and zero otherwise conditional on the selected (instrumental) variable vector (***Z***). The *propensity score* is then the predicted probability (*p*) (or $log[p/(1-p)]$) resulting from this stage. In other terms, PS is a balancing score *b(**Z**)* defined as a function of the observed covariates ***Z*** such that the conditional distribution of ***Z*** given *b(**Z**)* is the same for "treated" (*i.e. shiw*=1) and control (*i.e. shiw*=0) units (D'Agostino, 1998).

At least two major alternative methods according to the distance function and the algorithm used can be mentioned:

1a) Nearest neighbor PS matching (NN).

This method consists of randomly ordering the treated and control units, then selecting the first treated unit and finding the control unit with the closest PS. Formally, treated unit *i* is matched to non-treated unit *j* such that:

$$d_{ij} = \left|p_i - p_j\right| = \min_{k \in \{D=0\}} \{\left|p_i - p_K\right|\}$$

This method can be slightly modified as follows:

1b) Nearest neighbor PS matching within caliper

For a pre-specified $\delta > 0$, treated unit *i* is matched to a non-treated unit *j* such that:

$$\delta > d_{ij} = \left|p_i - p_j\right| = \min_{k \in \{D=0\}} \{\left|p_i - p_K\right|\}$$

This method is the most simple and intuitive, and consists of matching every recipient units with the donor units which have the nearest PS within a fixed radius δ (*caliper*).

[5] Matching noise represents any discrepancy between the real data generating model and the underlying model of the synthetic complete data set (see D'Orazio et al. 2006).

[6] STATA codes PSMATCH2 matching algorithm, Leuven and Sianesi (2003).

2) Mahalanobis metric matching coupled with PS (M).

M is employed by randomly ordering units and then calculating a different (concept of) distance, *i.e.* the Mahalanobis, between the first recipient household and all donor units, such that:

$$d_{ij} = \left(u_i(p_i) - v_j(p_j)\right)^T \mathbf{A} \left(u_i(p_i) - v_j(p_j)\right)$$

where $\boldsymbol{u}_i$ is the $(k+1\times1)$ vector of k control covariates for the recipient plus – since we use Mahalanobis including the PS - an additional covariate that is the logit of the estimated propensity score of unit i (p_i). $\boldsymbol{v}_j$ is the $(k+1\times1)$ vector of k control covariates for the donor plus the logit of the estimated propensity score of unit j (p_j).

A is a symmetric positive definite matrix. In particular $\mathbf{A} = \mathbf{S}^{-1}$, where **S** is the unbiased estimator of the pooled within-sample covariance matrix of the matching variables from the full set of control units. This allows correlations between variables to be taken into account. The control (HBS) household j with the minimum distance d_{ij} is chosen as the match for the "treated" (SHIW) household i, and both units are removed from the pool. Such process is repeated until all SHIW households find a match. As the dimension of $\boldsymbol{Z}$ increases, then the average Mahalanobis distance between units increases; thus, this matching can be harder compared to a pure propensity score procedure. Actually, after Mahalanobis distance has been calculated, treated units can be matched to non-treated ones by using the concept of radius (caliper) or that of NN.

1.2. Methodological choices

The link function employed in our matching is based on a set of common characteristics ($\boldsymbol{Z}_i$) surveyed both in SHIW and HBS and properly recodified to make them the most homogeneous. This required a deep understanding of the sampling features of both sources and an accurate process of recodification of the main control variables. Moreover, procedures of imputation based on pseudo-random lottery (estimation, prediction and Monte Carlo techniques) have been applied in some particular cases. The choice of a proper common support of variables has represented a crucial task to accomplish since our matching problem slightly differs from the typical data fusion situation, while it presents some advantageous characteristics. In fact, the two surveys share consumption information, even though with a different degree of detail. This issue will be specifically addressed in section 2.

The unit of analysis for the matching process is the household; in particular, most of the variables in the common vector refer to the household head[7].

As a matching algorithm we alternatively use nearest neighbor within caliper and the propensity score coupled with a Mahalanobis metric for the $\boldsymbol{Z}_i$ variables. As we want to assign to each SHIW household a vector of consumption components, despite the significant difference in sample size (N_{Hbs}=22.246 and N_{Shiw}=7.951), we do not perform a one-to-one matching, letting HBS households being assigned to more than one SHIW record. Therefore, some "less similar" HBS units will be discarded by the matching procedure. This will force the algorithm to match all the recipient sample, even replicating donor units, if needed. However, the cost of dismissing a matching without replacement is that the extent of variation in conditioning covariates (**Z**) can be spuriously altered as a consequence of the matching algorithm. Yet, using a one-to-one matching without replacement we would not match the whole SHIW sample, unless enlarging too much the radius of acceptability (*caliper*). This entails losing the representativeness of the population in the

[7] This issue deserves a particular concern; it is tackled in the next section.

recipient sample and thus invalidating following statistical inference which is the meta-goal of this data fusion.

Moreover, in order to control for systematic differences between the two samples and obtain a more accurate matching we divide the joint dataset in 50 up to 100 strata (or cells) obtained by the combination of quintiles (deciles) of a homogeneous aggregate of household total consumption (TMC, see section 4) and 10 household typologies. Then we allow the matching among units conditional on being included in the same stratum only.

Finally, most of results presented in section 4 are obtained using Mahalanobis metrics, as it is preferred to nearest neighbor method due to a better performance in terms of both conditional variability of target variables and the joint distributions. A table summarizing the methodological choices adopted in this work is reported in Appendix C.

2. Main assumptions and specificities of our matching problem

The applied researcher is typically interested in the joint (or conditional) distribution of three (vectors of) variables $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ but often no database exists where such three variables are simultaneously observed. Sometimes, two distinct surveys are available, one containing $\boldsymbol{X}$ and $\boldsymbol{Z}$ and the other $\boldsymbol{Z}$ and $\boldsymbol{Y}$. In order to integrate the two datasets we have to suppose that information in $\boldsymbol{Z}$ are useful to jointly determine $\boldsymbol{X}$ and $\boldsymbol{Y}$. The fusion process is based on the assumption that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent conditional on $\boldsymbol{Z}$ even though they are unconditionally dependent (conditional independence assumption, CIA henceforth); however, they may be conditionally dependent in reality.

Formally the CIA can be expressed as $P(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}) = P(\boldsymbol{X}|\boldsymbol{Z})*P(\boldsymbol{Y}|\boldsymbol{Z})$. Under the CIA, one can prove that any inference based on the resulting dataset about the actually unobserved associations is valid. Departures from CIA will determine heavy bias in the estimates based on the integrated synthetic dataset. Unfortunately, this assumption is not testable, while one might want to test the matching quality and the sensitivity of results with respect to failure of common support condition (or unobserved heterogeneity). This latter issue is particularly relevant since the aim is to make inference about the relationships between variables that are not jointly surveyed starting from the resulting joint distribution.

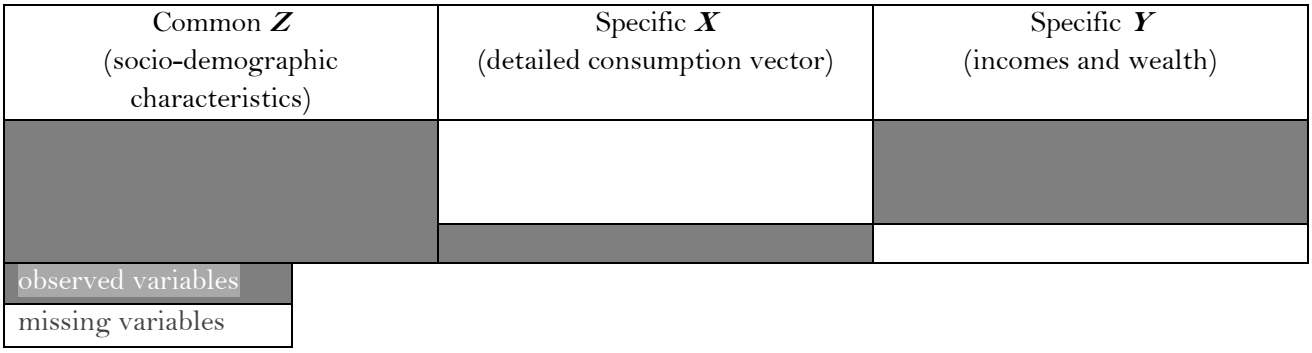
In the context of statistical matching, Rässler (2002) shows that the identification problem concerning the association of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is strictly related to the explanatory power of the common variables $\boldsymbol{Z}$ (in terms of $\boldsymbol{X}$ and $\boldsymbol{Y}$). The greater the latter, the smaller the range of admissible values of the unconditional association of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Rodgers (1984) shows that only a very high correlation (both between $\boldsymbol{Z}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$, $\boldsymbol{Y}$) narrows such range substantially. Thus, choosing a $\boldsymbol{Z}$ with a high explanatory power is crucial for the CIA to hold. An alternative approach pursues the unbiasness of the integration basing on auxiliary information (AI, Singh *et. al*, 1993) on the association between the two distributions which is assumed to closely describe the distribution not-jointly observed in the datasets to be fused. In both cases, the crucial underlying assumptions cannot be tested and the resulting empirical distribution is actually compatible with many unobserved distributions if those assumptions do not hold[8].

Our matching problem is common in the sense that we want to analyze a typical economic association, namely the joint distribution of consumers' expenditures and income/wealth, though at high level of details for consumption expenditure items. Therefore, at a first stance we could

[8] D'Orazio et al. (2004) describe a different approach to statistical matching explicitly dealing with the issue of uncertainty, implying the assessment of all the parameter values which are consistent with the available information.

include the vector of detailed consumption items ($\boldsymbol{C}=[\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_K]$, where $\boldsymbol{c}_k$ is a vector of households' consumption of commodity k) observed in the HBS into the vector $\boldsymbol{X}$, household incomes ($\boldsymbol{I}$) and wealth ($\boldsymbol{W}$) components observed in SHIW into the vector $\boldsymbol{Y}$, and the composite vector of socio-demographic household (and household head) characteristics (properly re-codified) in the common variables $\boldsymbol{Z}$. This would represent the typical data fusion problem analysed in depth in Rässler (2002) and depicted in Figure 2.1.

Figure 2.1: Typical data fusion



According to this scheme, the main problem is a quite weak (though statistically significant) explanatory power of $\boldsymbol{Z}$ in terms of $\boldsymbol{X}$ and $\boldsymbol{Y}$. In fact, though socio-demographic and educational information are useful to predict both consumption choices and income/wealth outcomes, regressions of consumption or income/wealth on $\boldsymbol{Z}$ explain only a very small share of variation in the dependent variables, leaving the remainder in the residual. In terms of the validity of the statistical matching and thus of the identification of $f(\boldsymbol{X},\boldsymbol{Y})$ this would imply a very wide range of admissible values for the unconditional association of $\boldsymbol{X}$ and $\boldsymbol{Y}$ and thus a great uncertainty in the results.

Still, a key feature of our case is that both surveys include information on consumption at the household level; however, while HBS is focused on this issue by specifically providing data on single household consumption goods and services, SHIW gathers information on consumption at a lower level of disaggregation. Since - as Vousten and de Herr (1989) demonstrate - there can be, *ceteris paribus*, a trade-off between the width and the accuracy about specific issues in a survey, we assume that the unconditional distribution of $\boldsymbol{C}$ is better represented in the HBS, although, as we shall discuss later, reliability of HBS is limited to some items only[9]. Nevertheless, we do not want to dismiss the consumption information contained in SHIW, though less detailed ($\boldsymbol{C}^a$ henceforth, where the superscript a stands for aggregates). Therefore, given the simultaneous availability of information on consumption, income and wealth, we take SHIW as a benchmark representation for the conditional distribution of $\boldsymbol{C}^a$ given $\boldsymbol{I}$, $\boldsymbol{W}$ and $\boldsymbol{Z}$. The circumstance that information on consumption expenditures is provided also in the recipient dataset determines a situation which is different (and more advantageous) compared to that represented in Figure 2.1. This situation is represented by Figure 2.2 where an overlapping of $\boldsymbol{X}$ and $\boldsymbol{Y}$ exists that does not flow directly into $\boldsymbol{Z}$. In fact, it has to be remarked that hundreds of consumption items surveyed with a recall period of one month or one quarter cannot be simply scaled up and thus be considered as coinciding or homogeneous to five or six consumption aggregates (such as food, durable, non-durable, etc..) with a recall period of one year. Nevertheless, whether conveniently treated, some information contained in $\boldsymbol{X}$ (i.e. $\boldsymbol{C}$) can be aggregated and used as common information flowing into $\boldsymbol{Z}$ so as to

[9] Namely, non-durable high frequency purchases.

recover a benchmark for the $\boldsymbol{X,Y}$ correlation structure. Hence, we can reproduce $\boldsymbol{C^a}$ in HBS as a vector of consumption commodities blocks' sums i.e. $\boldsymbol{C^a} = [\sum_{k=1}^{K_1} \boldsymbol{c_k}, \sum_{k=K_1+1}^{K_2} \boldsymbol{c_k}, \dots, \sum_{k=K_{A-1}+1}^{K} \boldsymbol{c_k}]$.

Figure 2.2: Our data fusion problem

Common $\boldsymbol{Z}$ (socio-demographic characteristics+ homogeneous consumption aggregates)	Specific $\boldsymbol{X}$ (detailed consumption vector)	Specific $\boldsymbol{Y}$ (incomes, wealth and consumption)

In a sense, if one is willing to assume that SHIW provides a good benchmark for the estimation of the joint distribution of consumption, income and wealth in the population, the CIA becomes a weaker assumption to maintain. This way represents an (internal) alternative to the exploitation of auxiliary information where AI on $f(\boldsymbol{X,Y,Z})$ is recovered from the recipient dataset rather than from a third data source. Furthermore, this assumption allows to check the matching quality at a superior level than the one usually testable (see section 4).

In practice, we include the common part of $\boldsymbol{C}$ and $\boldsymbol{C^a}$ in the $\boldsymbol{Z}$ control vector, but the whole information included in $\boldsymbol{C}$ within the HBS represents at the same time the target missing variable to be imputed in the SHIW sample.

In order to achieve this matching, a deep understanding and comparison of the sampling features of both sources is required. First, we carry out an accurate process of recodification of the highly detailed consumption items of HBS in terms of the medium-level-aggregation of SHIW; in addition the simulation of pseudo-random lottery procedures is performed to adjust key control variables where a significant difference in the sampling design and in the recall period make them poorly comparable (see section 3.2 on the treatment of durable goods).

3. Methodology

3.1. *Identification of matching unit*

In order to apply any statistical matching technique for imputing information on consumption, we first need to identify the proper matching unit. In this case, the reference unit is the household. However, the definition of household head is fairly different between the two datasets. In particular, while the reference person in SHIW is the one self-declared as responsible for the household economy, the reference person in HBS is the household identity record holder. This divergence accounts for a significant difference in terms of gender composition of household heads among the two sources (44.5% of female in SHIW vs 32.5% in the HBS, see Tab. 3.1).

To this end, we decided to perform a recodification of the reference person for matching aims.

As the registry sheet is more likely to be hold by men compared to the SHIW responsible of household economy, we assume that, if existing, husband/male partner is the household head in SHIW, except in case the female spouse/partner is the major earner among the partners. This implies a significant re-alignment of gender composition of household heads as well as a correction

for discrepancies in distribution of common, categorical variables (see section 3.3.1 and Appendix B).

Table 3.1: Share of male and female household heads in the two surveys

	HBS		SHIW_pre		SHIW_post	
Female	Freq.	Percent	Freq.	Percent	Freq.	Percent
0	16,698,963	67.5	13,387,725	55.5	16,693,904	69.2
1	8,056,407	32.5	10,722,158	44.5	7,415,979	30.8
Total	24,755,370	100	24,109,883	100	24,109,883	100

3.2. Durables issues and the amount of expenses for extraordinary maintenance of household dwellings

A specific adjustment required to make common variables homogeneous between the two surveys concerns expenditure on durable goods, included the expenses for extraordinary maintenance of household properties. HBS expenses are in fact mainly surveyed on monthly basis, but some durable items are recorded on the previous three months. As SHIW refers to the year, this difference requires a correction. For food and other non-durable expenditure with a monthly purchase frequency, the yearly amount is simply obtained by multiplying the corresponding values by 12[10]. For durables recorded in the previous month but with a purchase frequency lower than a month, we adopt some *ad hoc* hypotheses in order to account for the limited time of recording[11]. For durables surveyed in the last three months (mainly transport expenditure), a more demanding correction is implemented. First of all, HBS divides them by 3 in order to gather a 'monthly expenditure'. Hence, we first need to restore the whole value by multiplying by a factor of 3. We then need to impute probabilities to account for households not purchasing durables in the three months preceding the interview but likely to do it during the year. In other terms, recording only the last three months consumption of these items, HBS severely underestimates the share of individuals purchasing such goods over the year[12] (e.g. transport, Tab. 3.2).

We address this issue by estimating on SHIW a logit model of the probabilities of durables purchase in the year on covariates common to the two sources, and, on the basis of the latter, we impute the predicted probabilities to HBS households. The underlying assumption is that SHIW delivers a better representation of yearly durable purchases frequency. We apply this method to a subset of relevant durables only, the transport-related ones (cars, motorcycles, camping vans, etc…- "transp"). A Monte Carlo simulation is then run to select HBS households to be imputed such expenditure among those with no durables expenditure having the highest probabilities of doing such purchase according to the imputed score. We then calibrate the number of households with imputed purchases as the difference in the share of units with that kind of durables

[10] For the main non-durable aggregates, in this work we decided two overlook consumption seasonality issues.

[11] For instance, for clothes items we double the amount assuming this kind of expenditure is generally done twice a year (winter and summer). This heuristic solution can overestimate the amount for some households but, on aggregate terms, can compensate for households whose purchase has not been recorded since made during the remainder of the year (not in the last month). Other expenses, though surveyed in the last month, are unlikely to be done more than once a year, so original values are left. As a result, these latter can be underestimated. As these items account for little amounts, fortunately, comparison of the two sources on the "other nondurable" aggregate suggests a good fit.

[12] A further adjustment should account for the role of multiple purchases during the year. However, we can assume that for transport durables, these additional purchases are not frequent and should not affect significantly the overall amount.

expenditure between the two surveys (in order to obtain an overall share in HBS almost equal to the SHIW one).

Finally, to endow selected families with a given amount of durables, a propensity score matching procedure is applied within the HBS sample so as to provide them with a vector of durables of the "nearest" households in HBS itself (intra-sample matching). Results are presented in tables 3.2 and 3.3.

Table 3.2: Share of HBS households spending on transport durables (transp>0) before and after the imputation compared to SHIW

	HBS_pre	SHIW	HBS_post
transp>0	(%)	(%)	(%)
0	97.3	90.6	90.6
1	2.7	9.4	9.4
Total	100	100	100

Table 3.3: Distribution of "transp" in HBS before and after the imputation, compared to SHIW (over the whole sample and for positive values only)

HBS_pre					HBS_post					SHIW				
Transp					Transp					Transp				
	Perc.	Smallest				Perc.	Smallest				Perc.	Smallest		
1%	0	0			1%	0	0			1%	0	0		
5%	0	0			5%	0	0			5%	0	0		
10%	0	0	Obs	22,246	10%	0	0	Obs	22,246	10%	0	0	Obs	7,951
25%	0	0	Sum of Wgt.	2.49E+07	25%	0	0	Sum of Wgt.	2.49E+07	25%	0	0	Sum of Wgt.	2.41E+0
50%	0		Mean	211	50%	0		Mean	738	50%	0		Mean	1,116
		Largest	Std.Dev.	2,002			Largest	Std. Dev.	3,622			Largest	Std. Dev.	4,532
75%	0	47,342			75%	0	50,461			75%	0	58,000		
90%	0	47,360	Variance	4,00E+07	90%	0	50,461	Variance	1.31E+07	90%	-	60,000	Variance	2.05E+07
95%	0	50,461	Skewness	13.0	95%	3,300	57,009	Skewness	6.6	95%	10,000	70,000	Skewness	6.0
99%	9,500	57,009	Kurtosis	211.0	99%	18,500	57,009	Kurtosis	57.0	99%	23,000	100,000	Kurtosis	56.1
Transp>0					Transp>0					Transp>0				
	Perc.	Smallest				Perc.	Smallest				Perc.	Smallest		
1%	50.01	12			1%	50	12			1%	100	50		
5%	89	20			5%	100	12			5%	800	100		
10%	120	21	Obs	569	10%	160	12	Obs	2,022	10%	2,000	100	Obs	723
25%	320	29	Sum of Wgt.	6.73E+05	25%	500	20	Sum of Wgt.	2.38E+06	25%	5,000	100	Sum of Wgt.	2.26E+06
50%	4,000		Mean	7,812	50%	4,500		Mean	7,737	50%	10,000		Mean	11,895
		Largest	Std. Dev.	9,432			Largest	Std. Dev.	9,133			Largest	Std. Dev.	9,527
75%	12,800	47,342			75%	12,800	50,461			75%	16,000	58,000		
90%	20,000	47,360	Variance	8.90E+07	90%	19,000	50,461	Variance	8.34E+07	90%	24,000	60,000	Variance	9.08E+07
95%	26,000	50,461	Skewness	1.7	95%	25,000	57,009	Skewness	1.6	95%	28,000	70,000	Skewness	2.0
99%	40,000	57,009	Kurtosis	6.2	99%	40,000	57,009	Kurtosis	6.0	99%	50,000	100,000	Kurtosis	12.1

Unfortunately, the same procedure of imputation from SHIW cannot be applied for other durables aggregate (otherdur), including items such as furniture, furnishings, appliances etc..., as this latter aggregate is very dissimilar among the two sources: as HBS is much more detailed, the probabilities of purchasing at least one of the items included in such variable is considerably higher, and not comparable to the SHIW one (Tab. 3.4). Moreover, the distributions are significantly different (Tab. 3.5), as HBS shows bumps of small values due to the exhaustive list of goods surveyed compared to the more aggregate variable recorded in SHIW (which is likely to be mis-reported due to memory effect). However, the latter presents significantly higher average

values, which seem suggesting this source being more reliable with respect to these items, at least in the average amounts.

Table 3.4: Share of household with "otherdur" in the two survey

	HBS	SHIW
otherdur>2	(%)	(%)
0	5.6	65.7
1	94.4	34.3
Total	100	100

Table 3.5: Distribution of "otherdur" in HBS compared to SHIW (on the whole sample and among households with positive values only)

HBS					SHIW				
Otherdur					Otherdur				
	Perc.	Smallest				Perc.	Smallest		
1%	0	0			1%	0	0		
5%	0	0			5%	0	0		
10%	0	0	Obs	22,246	10%	0	0	Obs	7,951
25%	0	0	Sum of Wgt.	2.49E+07	25%	0	0	Sum of Wgt.	2.41E+07
50%	27		Mean	159	50%	0		Mean	649
		Largest	Std.Dev.	462			Largest	Std. Dev.	2,214
75%	130	20,000			75%	500	40,000		
90%	398	20,305	Variance	213,797	90%	1,780	40,000	Variance	4.90E+06
95%	707	21,550	Skewness	14.8	95%	3,000	40,000	Skewness	10.4
99%	1,929	24,109	Kurtosis	448.4	99%	10,000	50,000	Kurtosis	164.2
Otherdur>0					Otherdur>0				
	Perc.	Smallest				Perc.	Smallest		
1%	4	0.87			1%	60	25		
5%	12	1			5%	200	25		
10%	22	1	Obs	12,644	10%	250	25	Obs	2,696
25%	69	1	Sum of Wgt.	1.41E+07	25%	500	25	Sum of Wgt	8.28E+06
50%	111		Mean	281	50%	1,000		Mean	1,891
		Largest	Std.Dev.	587			Largest	Std. Dev.	3,453
75%	270	20,000			75%	2,000	40,000		
90%	648	20,305	Variance	3.44E+05	90%	4,000	40,000	Variance	1.19E+07
95%	1,029	21,550	Skewness	12.2	95%	7,000	40,000	Skewness	6.8
99%	2,500	24,109	Kurtosis	296.1	99%	15,000	50,000	Kurtosis	70.2

A similar procedure is carried out for the amount of expenses for extraordinary maintenance of household dwellings (extraord_maintain), which are characterized in HBS by a lower frequency and a lower average value of the declared purchase. Table 3.6 shows that despite the procedure of imputation, a significant difference in the two distribution holds. This issue induced us to further calibrate mean values after the fusion in order to match SHIW figures.

Table 3.6: Distribution of "extraord_maintain" in HBS before and after the imputation, compared to SHIW (on the whole sample and among households with positive values only)

HBS_pre					HBS_post					SHIW				
Mastrip					Mastrip					Mastrip				
	Perc.	Smallest				Perc.	Smallest				Perc.	Smallest		
1%	0	0			1%	0	0			1%	0	0		
5%	0	0			5%	0	0			5%	0	0		
10%	0	0	Obs	22,246	10%	0	0	Obs	22,246	10%	0	0	Obs	7,951
25%	0	0	Sum of Wgt.	2.49E+07	25%	0	0	Sum of Wgt.	2.49E+07	25%	0	0	Sum of Wgt.	2.41E+07
50%	0		Mean	136	50%	0		Mean	448.6	50%	0		Mean	1,006
		Largest	Std. Dev.	1,336			Largest	Std. Dev.	2,468.0			Largest	Std. Dev.	7,183
75%	0	52,048			75%	0	63,907			75%	0	130,000		
90%	0	63,907	Variance	1,784,429	90%	500	63,907	Variance	6.09E+06	90%	1,300	130,000	Variance	5.16E+07
95%	0	66,782	Skewness	26.7	95%	2,000	66,782	Skewness	12.7	95%	4,800	180,000	Skewness	21.6
99%	3,828	86,421	Kurtosis	1,149.9	99%	11,000	86,421	Kurtosis	242.0	99%	20,000	350,000	Kurtosis	714.6
mastrip>0					mastrip>0					mastrip>0				
	Perc.	Smallest				Perc.	Smallest				Perc.	Smallest		
1%	25.89	5.46			1%	39	5.46			1%	300	250		
5%	67	16			5%	84	15.66			5%	500	250		
10%	120	16	Obs	1,116	10%	150	16	Obs	3,252	10%	600	250	Obs	1,165
25%	300	16	Sum of Wgt.	1.24E+06	25%	350	16	Sum of Wgt.	3.57E+06	25%	1150	250	Sum of Wgt.	3.32E+06
50%	1,000		Mean	2,747	50%	1200		Mean	3,133	50%	3000		Mean	7,314
		Largest	Std. Dev.	5,368			Largest	Std. Dev.	5,842			Largest	Std. Dev.	18,140
75%	3,000	52,048			75%	3000	63,907			75%	6000	130,000		
90%	6,657	63,907	Variance	2.88E+07	90%	8,800	63,907	Variance	3.41E+07	90%	15,000	130,000	Variance	3.29E+08
95%	11,934	66,782	Skewness	6.7	95%	13,000	66,782	Skewness	5.2	95%	25,000	180,000	Skewness	8.7
99%	20,000	86,421	Kurtosis	74.5	99%	23,500	86,421	Kurtosis	43.0	99%	100,000	350,000	Kurtosis	115.0

3.3. Selection and recoding common variables

Finally, a significant effort has to be devoted in order to fill the control variables vector of common characteristics (***Z***) with the greater number of homogeneous socio-demographic and economic information, with a particular focus on consumption variables. This will serve to build a distance function to be minimized in order to match "similar" units from the two original samples.

We recoded variables surveyed in a different way but providing common information on the household or its HH to make them homogeneous.

3.3.1. Evidence on the common variable distributions in the two sources

This section summarizes the evidence on the comparison of the common variables among the two sources. Tables showing the distribution of socio-demographic features in SHIW and HBS are reported in Appendix B.

The recoding of the HH makes the distribution of many common variables rather compatible among the two sources.

Household age groups and region of residence (Tab. B.1, B.2) of the HH do not show significant differences in shares (discrepancies greater than 2% are found only for the age classes 35-39 and 40-44 years old – which are thin and close - and region Trentino Alto Adige). On the opposite, much more relevant redistribution among classes are found in the household types (Tab. B.4) and number of household components (Tab. B.3): in the latter, a dramatic excess of singles (4.7 pp) is recorded in HBS compared to SHIW, mainly to detriment of couples. This latter evidence is mirrored in the distribution by family types, where a shortage of lone persons with aged 35-64 and single-parents and a correspondent excess of couples without children with reference person aged 65 or more is found in SHIW relative to HBS. Analogously, the distribution by marital status (Tab. B.5) displays a significant redistribution between married and single.

Turning to the household head characteristics, a more comparable picture emerges: educational level (Tab. B.6) as well as occupational status (Tab. B.7) do not show substantial over or under-representation, except for a slight compensation between employed and not employed HH families (below 2 pp).

The distribution by branch of activity (Tab. B.8) displays a lower share in trade and catering services in SHIW and a positive discrepancy in private services to person, both just above the 2% threshold. A satisfying comparability is achieved looking at the distribution by work status, where only one difference greater than 2% is recorded in the blue-collar category (Tab. B.9), to the detriment of office workers and school teachers and not employed. This last evidence seems suggesting that the definition of HH still slightly under-represents women household-headed families in SHIW after the recodification. More significant discrepancy are instead observed in the distribution of house-related variables (Tab. B.10, B.11, B.12), probably owing also to differences in the accuracy of the surveys on this topics. In particular, a noteworthy positive difference can be observed in the number of second dwelling, where 91.8% of HBS households do not hold any second dwelling, compared to the 85% of SHIW. However, this last figure is likely to be severely underestimated even in SHIW (Cannari and Faiella, 2008). In addition, a substantial redistribution between occupiers and home owners is also observed.

Concerning continuous common variables, a series of figures comparing histogram distributions of consumption aggregates belonging to the two sources are reported below (Fig. 3.2). In particular, as mentioned above, as HBS contains a detailed list of expenditures, an accurate recodification process has been carried out in order to get aggregates comparable to SHIW, which provides a rougher breakdown of overall consumption into the following macro-variables: food, other non-durables and total durable consumption (means of transport and other durables). In addition, we report a figure for actual rents paid.

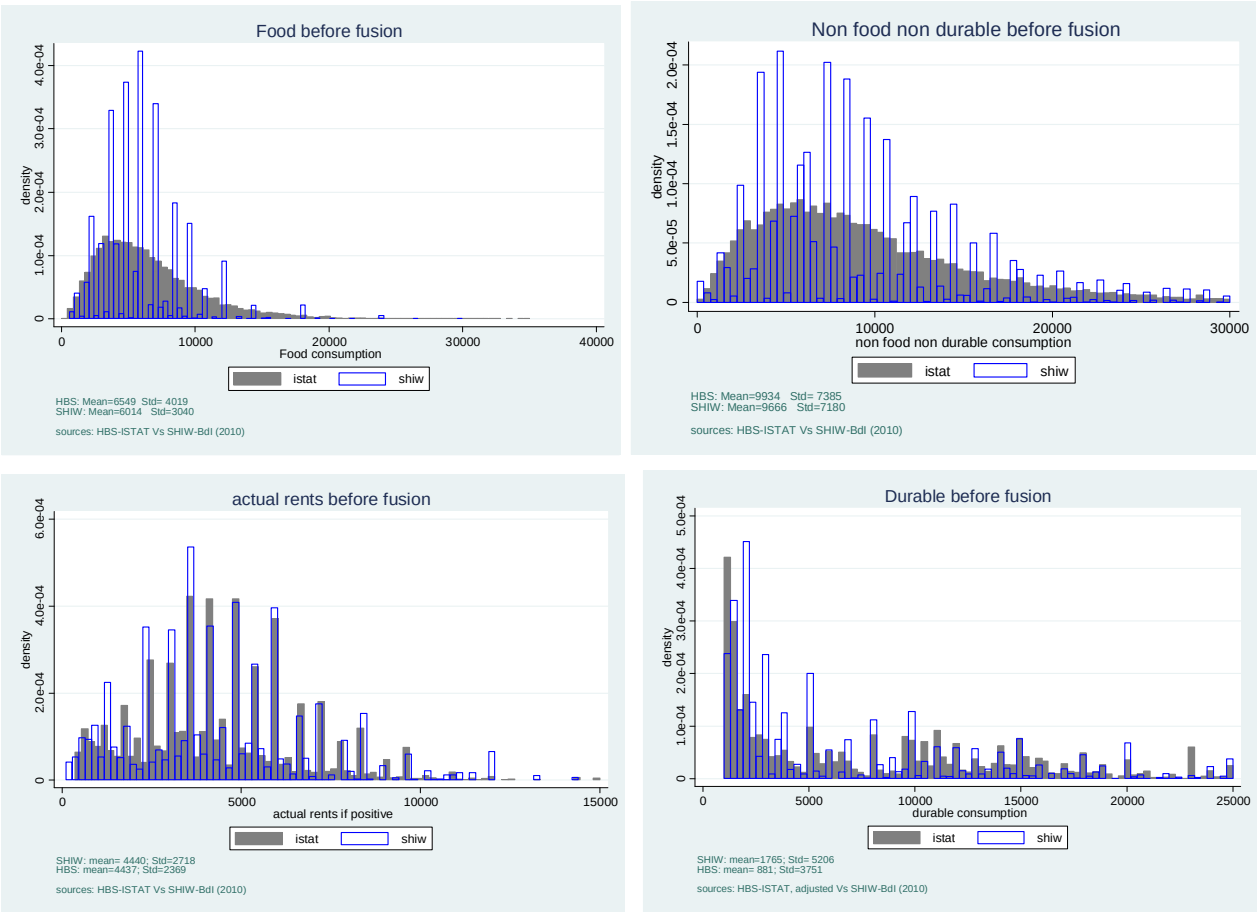
In general, SHIW distributions are much less smooth compared to HBS due to the lower accuracy of recording in the field of consumption relative to other information such as incomes and wealth. In fact, many modal values are thickened in correspondence of round amounts.

Therefore, we adopted a simple trick to deal with such heaping and rounding pattern and prevent bumps to work as attractor, which would cause a multi-modal synthetic distributions. More specifically, before the fusion process, we shocked at random with an almost null alteration of second and higher moments[13] those original SHIW sub-aggregates which display a clear heaping and rounding pattern (i.e. food and non-food non-durable expenditure).

Turning to specific distributions, while food and other non-durables (panel A) fit rather well both in terms of means and the higher statistical moments, durable consumption, as expected, is much more noisy and displays some inconsistencies. In particular, although the transport expenditure - as discussed in the previous section - appears more similar after adjustment at least in terms of mean and variance relative to the pre-adjustment situation, a substantial difference in total durable expenditure still emerges (881 vs 1,765 the averages, Panel D); this confirms the other durable consumption being definitely dissimilar between the two samples in terms of the overall distribution. Finally, actual rents paid (Panel C) display a definitely high degree of comparability, being a variable less subject to under-reporting or mis-reporting issues.

[13]The shock is an *iid* draw from a normal distribution with zero mean and 5 Euros standard deviation, about 0.13% of food consumption variability.

Figure 3.2: Main consumption aggregates distributions in the two surveys



Panel A. Food expenditure; Panel B: Non-food non-durable expenditure; Panel C: Actual rents; Panel D: Durable expenditure (trimmed).

In sum, the above evidence suggests that whilst HBS can be regarded as more reliable in providing the true non-durable purchases picture, SHIW appears to give a more credible representation of durables and other extraordinary expenditures at least for the average amounts. Indeed, it is not surprising that consumption aggregates closer to the concept of wealth (such as durables and the extraordinary expenditure for dwelling maintenance) or savings (such as mortgages and private pensions) prove to be better assessed by the longer - and more issue-specific - recall of the SHIW. This prompts us to consider SHIW as the benchmark for these aggregates (see also section 4.1). Thus, we believe a proper measure of total consumption cannot discard such original information coming from this source.

4. Matching results and validity tests

Following Rässler (2002 and 2004), four increasingly demanding levels of validity can be identified when dealing with the problem of statistical matching:

- First Level: Preserving Individual Values.
- Second Level: Preserving Joint Distributions.
- Third Level: Preserving Correlation Structures.
- Fourth Level: Preserving Marginal Distributions.

Since the true individual values are unknown, the only way the first level validity can be assessed is by means of a simulation study (Rässler, 2002). Thus, it is not possible to carry out tests on real data. The second and third levels too would require the knowledge of the ***(X,Y,Z)*** joint distribution or at least of its second moments.

Thus, we are able to test the validity of our data fusion at the fourth level; in addition, we assess different matching hypotheses and algorithms with respect to the joint ***(C, I, W)*** distribution observed in SHIW so as to account for the second and third levels of validity as well.

4.1. Lower level of validity: the marginal distributions

In order to check this level of validity, we first compare the unconditional distribution of several aggregates of consumption among the original HBS, SHIW and the resulting fused file (henceforth Synt). Results shown in this section are obtained with the PSM method with Mahalanobis distance and 100 cells of stratification.

Table 4.1 provides an overall picture in terms of first and second moments for the main common aggregates among the two original surveys and the synthetic one, considering the whole sample and to households with positive values only respectively. This table allows to recognize if differences in the average amount on the whole sample are due to discrepancies in frequency of households purchasing such good or to gaps in positive expenditure amount.

As it can be noticed, all synthetic univariate distributions reproduce rather faithfully the original HBS ones in terms of first moments; this result suggests the matching procedure successfully achieves the preservation of marginal distributions from the donor sample.

Food and non-food non-durable expenditures appear very well imputed in the fused dataset, being close in their first and second moments to the ones of the HBS, whose greater reliability for these items is due to the diary-based recording.

Table 4.1: Summary statistics on main consumption aggregates in the HBS, SHIW and fused files

	Mean		Std.Dev.	
Variable	all	>0	All	>0
food_istat	6,550	6,550	4,020	4,020
food_shiw	6,015	6,015	3,041	3,041
food_synt	6,263	6,263	3,271	3,271
other_nondurable_istat	9,591	9,591	8,034	8,034
other_nondurable_shiw	9,667	9,700	7,181	7,170
other_nondurable _synt	10,004	10,004	8,682	8,682
transport_istat	738	7,737	9,131	9,131
transport_shiw	1,116	11,895	4,532	9,521
transport_synt	729	8,420	3,497	8,740
otherdur_istat	159	281	587	587
otherdur_shiw	649	1,891	2,214	3,453
otherdur_synt	225	385	862	1,099
real_goods_istat	4	151	296	296
real_goods_shiw	83	1,408	662	2,352
real_goods_synt	11	314	138	652
health_ins_istat	14	373	413	413
health_ins_shiw	235	923	695	1,122

health_ins_synt	26	230	101	210
accident_ins_istat	25	237	226	226
accident_ins_shiw	235	923	695	1,122
accident_ins_synt	15	101	48	85
life_ins_istat	14	99	83	83
life_ins_shiw	188	1,636	932	2,278
life_ins_synt	15	101	48	85
ord_mantain_istat	66	631	1,234	1,234
ord_mantain_synt	66	616	424	1,158
actual_rents1_istat	764	4,437	2,370	2,370
actual_rents1_shiw	917	4,440	2,181	2,718
actual_rents1_synt	870	4,211	2,033	2,435
extraord_mantain_istat	449	3,133	5,841	5,841
extraord_mantain_shiw	1,006	7,314	7,182	18,132
extraord_mantain_synt	943	6,853	5,205	12,505
priv_pens_istat	2	143	129	129
priv_pens_shiw	278	1,891	1,125	2,359
priv_pens _synt	2	149	27	149
mortgage_istat	61	491	250	250
mortgage_shiw	818	7,469	3,049	5,929
mortgage_synt	56	490	177	243

Nevertheless, for durables, the preservation of the HBS distributions determined by the matching implies a dramatic difference from SHIW which, as discussed, we deem as providing the benchmark for estimating these components.

More specifically, despite our corrections in the original HBS file, the resulting expenditure on means of transports (transp) and the extraordinary expenditure for dwelling mantainance (extraord_maintain) in the fused dataset continue to be significantly undersized relative to SHIW in their amounts. As expected (even more so given the lack of adjustment), the same results apply to other durables (otherdur), where significant differences are recorded in the two original surveys both in frequency and in the average positive amount. On the opposite, as mentioned above, actual rents are rather comparable in terms of average positive amounts between the two surveys, despite some gaps in frequencies; thus, the resulting distribution lies in between. Finally, differences in the average positive amount for life insurances, mortgages and private pensions, both in frequencies and amounts mirror in very undersized synthetic distributions if compared with SHIW, which, again, seems to deliver a more realistic account of these items. Such gaps are mainly due to the lack of representativeness of this kind of expenditures in HBS owed to a monthly recording which, in absence of a proper adjustment, significantly underestimates the frequency of purchases over the year.

For the sake of distributional analysis, we focus on three measures of total household consumption:

- Total matching consumption (TMC)
- Total expenditure (TE)
- Overall consumption (OC)

The first (TMC) is the consumption aggregate used for the matching stratification. It includes all items pertaining food and non-food non-durable expenditures, durables, and real goods. It does not include other items such as the amount paid for health insurance policy, life insurance, private/supplementary pensions as well as for mortgage installment, actual and imputed rents. This measure of consumption is used for testing the matching validity; to this end, synthetic measure is compared to the original donor distribution. However, two additional larger definitions of consumption can be considered: total expenditure (TE), which includes the full yearly monetary expenditures; overall consumption (OC), which provides the broadest definition of yearly household consumption, collecting imputed consumption items as well (TE plus imputed rents - 1st and 2nd dwellings). Both these measures contain commodities characterized by significantly different original distributions not undergone to any adjustment before the matching.

To account for the fact that the quality of HBS on a set of commodities appears rather poor, we propose a further definition of consumption which retains durable items as well as health and life insurances, payments for mortgages, deposit to private pensions and real goods from the original SHIW dataset[14], while relying on synthetic values resulting from matching for non-durables expenditure only (henceforth Synt2). We regard this definition of consumption as a synthesis of the best from the two original distributions, implying higher means and variances in the final household consumption distributions (Tab. 4.2). In addition, we consider this latter as the more suitable for carrying out a correction of the overall household consumption age (or income) profile provided by the original SHIW sample (see section 4.2).

Appendix D gives account of the subcomponents included in each of the abovementioned measure of consumption and definitions as well as the chosen benchmark distribution.

Table 4.2: Summary statistics on total household consumption in the HBS, SHIW and fused files (Synt and Synt2)

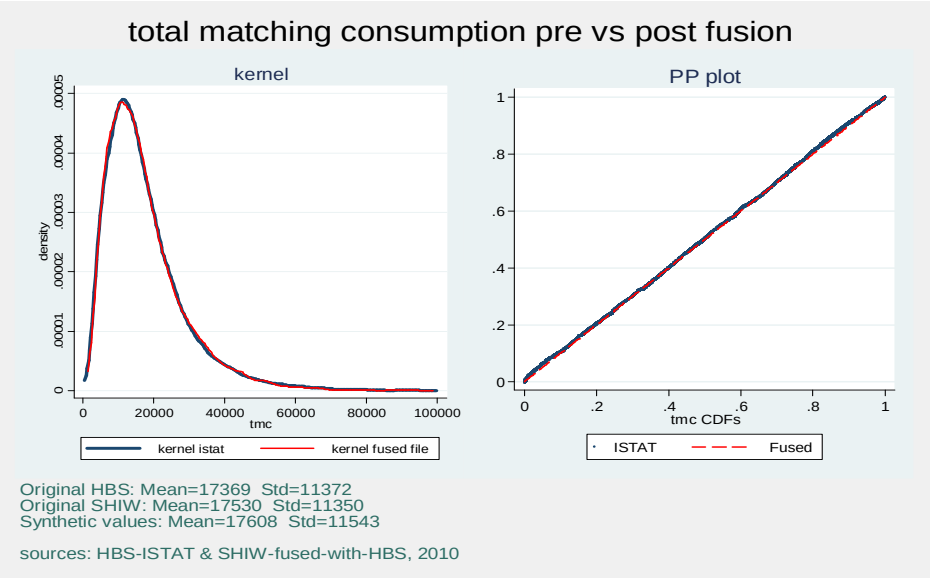
	mean	Std.Dev.
Variable	all	All
tmc_istat	17,359	11,382
tmc_shiw	17,530	11,350
tmc_synt	17,640	11,664
TE_istat	19,811	12,719
TE_shiw	21,031	15,999
TE_synt	20,650	14,013
TE_synt2	22,099	17,614
OC_istat	25,963	14,680
OC_shiw	28,366	22,168
OC_synt	27,240	17,980
OC_synt2	29,434	23,381

Moving from first and second moments to the whole distributional shape, we can notice that, consistently with the matching premises - yet despite the durables discrepancies - the unconditional distribution of TMC in the fused file overall fits very closely to the original HBS distribution. This result is due to the fact that the matching process selects donor units more

[14] However, for these aggregates as well, the imputed vector of goods and services can be employed for the internal (items) partition.

similar to those of the recipient; in addition, durable expenditures account for a small share of household expenditures.

Figure 4.1: Unconditional distributions of TMC



This graphical evidence is confirmed by standard inequality indices (Tab. 4.3). Indeed, the comparison displays a good preservation of the donor TMC inequality in the fused file. For instance, looking at the Gini, the gap between the two distribution is 0.001. Testing these differences between original HBS and the fused file in a framework of bootstrap inference (with 250 replications), TMC mean, Gini and General Entropy of the resulting file are not statistically different from that of HBS at the 1% level.

Table 4.3: Inequality indexes for TMC

Donor, original HBS sample					Fused file				
Percentile ratios					**Percentile ratios**				
p90/10	p90/50	p10/50	p75/25		p90/10	p90/50	p10/50	p75/25	
5.24	2.21	0.41	2.37		5.28	2.27	0.41	2.36	
Generalized Entropy indices GE(a), where a = income difference sensitivity parameter, and Gini coefficient					**Generalized Entropy indices GE(a), where a = income difference sensitivity parameter, and Gini coefficient**				
GE(-1)	GE(0)	GE(-1)	GE(-2)	**Gini**	GE(-1)	GE(0)	GE(-1)	GE(-2)	**Gini**
0.26	0.21	0.20	0.24	0.336	0.25	0.20	0.20	0.23	0.337

Considering TE, the structural differences between the two original surveys - with reference to additional components included in such definition of expenditure and despite the adjustments and imputations to make the two sources more comparable -, makes the synthetic representation a mixture of the two original samples, both in mean and distribution (Fig. 4.4).

However, the resulting distribution, on the one hand, and the original SHIW, on the other side, remain roughly comparable.

In terms of overall dispersion, the resulting Gini for OC in the fused dataset (Fig. 4.5) is slightly lower than the original SHIW file due to the equalizing impact of house-related components (imputed rents) which are less dispersed in the donor file than in the recipient.

On the opposite, the Synt2 version provides the higher mean and variability owing to non-durable consumption borrowed by HBS through the matching. Looking at the kernel and the PP-plot it is possible to notice that Synt and - even more - Synt2 distributions show lower peaks around the mode and thinner tails (i.e. are more platykurtic) compared to the original SHIW ones.

Figure 4.4: TE distribution in SHIW and synthetic file

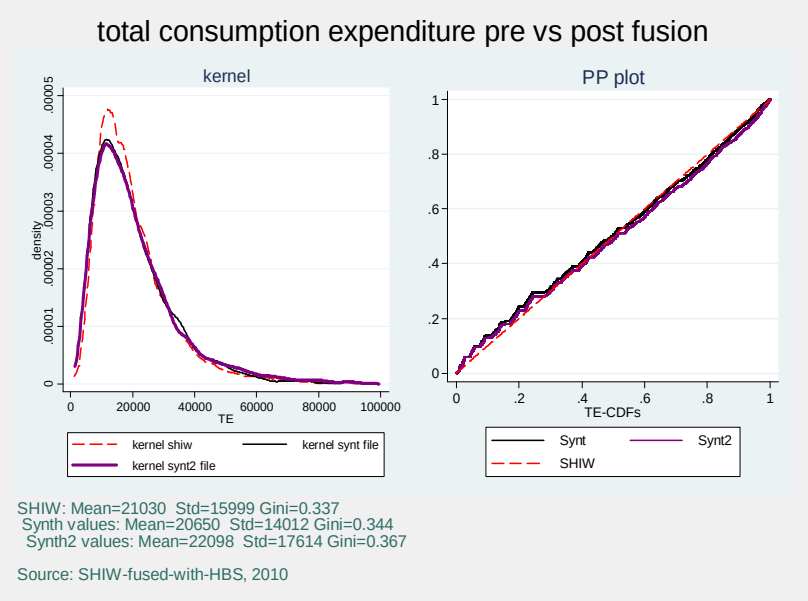
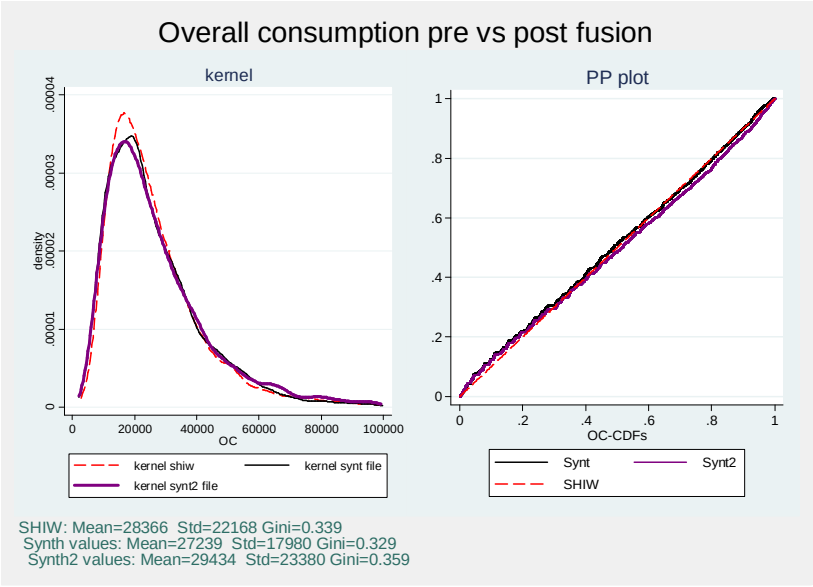


Figure 4.5: OC distribution in SHIW and synthetic file



4.2. Higher levels of validity: correlation and joint distribution

In order to test the validity at the second and third level, we should be able to observe the joint distribution $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, or at least their correlation structure. In fact, the joint distribution is preserved when, considering all variables, $\tilde{f}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = f(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})$ or, a bit less requiring, $\widetilde{cov}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = cov(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})$ where "~" indicates the synthetic parameters. Rässler and Fleisher

(1998) show that the fusion covariance $\widetilde{cov}(\boldsymbol{X},\boldsymbol{Y})$ equals the true $cov(\boldsymbol{X},\boldsymbol{Y})$ if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are, on average, uncorrelated conditional on $\boldsymbol{Z}$ (i.e. if $E\{cov(X,Y|Z)\}=0$).

As previously discussed, if we are willing to assume the $(\boldsymbol{C^t, I, W})$ distribution observed in SHIW as a valid term of reference to infer the population one, we can also make some test at these levels.

In particular, relying on this assumption, we retrieve information on the joint distributions of $(\boldsymbol{C^t, I, W, (Z)})$ by estimating a consumption equation[15] where the dependent variable is the (log of)TMC measured, respectively, on the original SHIW variable and the synthetic one. The explanatory variables vector (income, wealth variables and socio-demographic characteristics) comes from the recipient SHIW file and is the same in the two estimates. Then we can check the (*t*-test) differences in the estimated coefficient vector in order to assess the magnitude of the correlation structure dissimilarity. Individually, the differences in coefficients are pretty small and non-significantly different from zero at conventional level of significance (Tab. 4.4).

In order to draw general conclusions, we assess the overall vector dissimilarity within a standard Hausman (χ^2) test. Despite differences are negligible whether individually taken, the test reject the null hypothesis that the difference in coefficients is not systematic. However, this is not surprising since this test is rather demanding and the standard errors for the synthetic model are further inflated by the matching process.

Yet, such estimates can be employed to obtain some metrics for assessing different matching hypotheses (different algorithms, control covariates, stratification…) in terms of distance from the original SHIW conditional distribution. We then build two alternative indicators: the first is the absolute sum of estimated coefficient differences (ASEC) and the second is the absolute mean of predicted values differences (AMPV). In practice, we evaluate the level of preservation of the joint distribution in the fused dataset in terms of departure from the estimated $f(\boldsymbol{C}|\boldsymbol{I},\boldsymbol{W},\boldsymbol{Z})$ in SHIW, subject to the constrain that all the recipient units are matched with at least one donor unit.

Table 4.5 compares the outcomes provided by the Mahalanobis distance with 100 cells stratification, Mahalanobis with 50 cells and the nearest neighbor caliper coupled with 50 cells.

The first solution proves to be superior since provides the lower level for both metrics subject to the constraint of matching all the SHIW sample[16].

Table 4.4: Hausman test on TMC for analyzing the degree of preservation of joint distributions from SHIW to the synthetic dataset

dep: ln{TMC}	Synthetic	SHIW	Difference	SE
ln{disposable income}	0.176	0.179	-0.003	6.96E-05
ln{real wealth}	0.012	0.01	0.001	2.20E-05
ln{finacial wealth}	0.024	0.023	0.002	1.25E-05
ln{financial debt}	0.01	0.01	-0.001	1.03E-05
ln{actual rent}	0.065	0.062	0.002	1.23E-04
ln{imputed rent}	0.073	0.064	0.008	8.87E-05
number of earners	0.123	0.109	0.014	7.20E-05
Age	0.009	0.007	0.003	1.90E-05
Age^2	0	0	0	1.64E-07
HH woman	-0.042	-0.049	0.007	1.13E-04

[15] We use OLS estimator with standard error robust to heteroskedasticity.

[16] Nearest neighbor distance with 50 cells does not allow the full match of all recipient units, even increasing indefinitely the caliper.

Number of components	0.082	0.073	0.009	4.80E-05
Marital status (omitted: Never married)				
Married	-0.153	-0.114	-0.039	1.51E-04
Separeted, divorced	-0.132	-0.107	-0.025	1.73E-04
Widowed	-0.135	-0.086	-0.049	1.59E-04
Education (omitted: None)				
Primary	0.024	0.009	0.015	2.18E-04
Lower-secondary	0.137	0.086	0.051	2.33E-04
Upper-secondary	0.213	0.156	0.056	2.46E-04
Tertiary	0.281	0.225	0.057	2.75E-04
Postgraduate	0.213	0.206	0.007	4.47E-04
Occupational status (omitted: in work)				
First-time job seeker	-0.098	-0.034	-0.064	1.18E-03
Housewife	-0.049	-0.04	-0.009	9.66E-04
Rentier	-0.004	0.017	-0.021	9.07E-04
Pensioner	-0.069	-0.019	-0.05	9.41E-04
Unemployed	0.089	0.036	0.053	1.10E-03
Branch of activity (omitted: Agriculture)				
Manifacturing	-0.138	-0.065	-0.073	2.88E-04
Building and construction	0.023	0.04	-0.017	1.68E-04
Retail trade, lodging and catering	-0.019	0.007	-0.026	2.10E-04
Transport and communication	-0.013	0.024	-0.037	1.90E-04
Credit and insurance	-0.035	0	-0.035	2.60E-04
Real estate, renting, professional and business activity	0.199	0.174	0.025	3.02E-04
Properties				
Dummy second dwellings	0.014	0.017	-0.003	1.12E-04
Tenant	0.174	0.126	0.048	1.20E-03
With usufruct, use without charge	0.016	-0.001	0.017	1.64E-04
Work status (omitted: blue-collar, freelance)				
Office worker or school teacher	0.107	0.087	0.02	2.01E-04
Junior manager/cadre	0.176	0.129	0.046	3.12E-04
Manager, senior official self-employed	0.286	0.34	-0.054	3.75E-04
Member of the arts or professions	0.243	0.214	0.029	3.07E-04
Sole proprietor	0.134	0.139	-0.005	3.92E-04
Not employed	-0.002	-0.013	0.011	9.31E-04

HAUSMAN TEST

Ho: difference in coefficients not systematic: chi2(38) = 99770.01; Prob>chi2 = 0.0000

Table 4.5: Statistics on f(TMC|I,W,Z)

TMC	*Mahalanobis 100 cells*	*Mahalanobis 50 cells*	*Nearest Neighbor Caliper 50 cells*
ASEC	0.993	1.210	1.515
AMPV	0.046	0.067	0.055
N_{Synt}	7951	7951	7861

The last figures display a secondary outcome of the analysis, which is useful to compare our results with the existing literature on this topic on Italian data. In particular, we estimate household total expenditure (head)age and income profiles, comparing the original SHIW figures and the synthetic measure which employs imputed non-durables only from the matching (Synt2).

Our results - though based on different hypotheses and methodologies (e.g. mass imputation through PSM, instead of regression based imputation) as well as being referred to a more recent wave - are, on the whole, qualitatively comparable to those from Battistin et al. (2003) and, partially, to Cifaldi and Neri (2013). However, we obtain a (positive) less pronounced correction on total consumption relative to the previous studies. Such differences are at least in part due to different hypotheses of annualization of HBS non-durable expenditures[17].

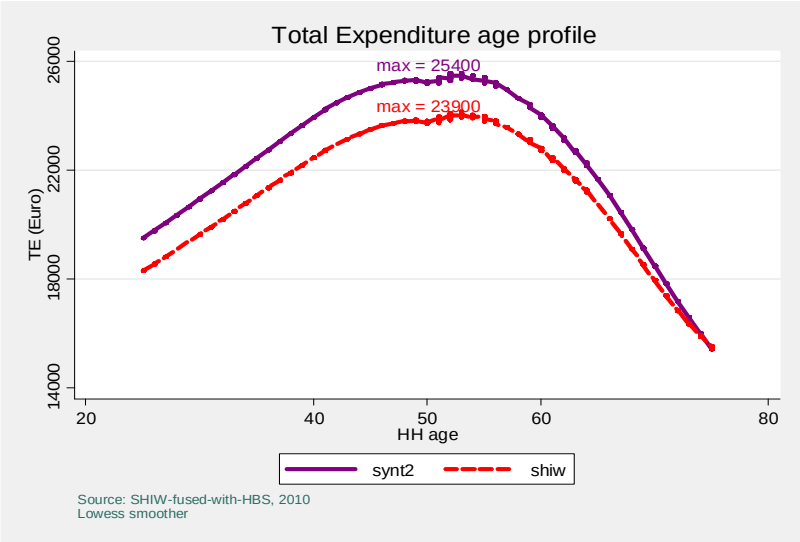
Our correction adjusts the age profile (Fig. 4.6) upward with a peak around 50 of about 6 percentage points.

In order to analyze the shape of the total expenditure adjustment on SHIW as provided by the synthetic values, Table 4.7 shows an OLS regression of the relative difference, at the household level, between TE expressed by the Synt2 version and the original SHIW value (thus capturing the correction on the non-durables imputations only). The explanatory variables are income and socio-demographic characteristics. The estimated coefficients show *inter alia* that the correction is increasing along disposable income deciles with a maximum in the ninth one.

At the aggregate level, such adjustment implies a (monetary) propensity to save for the household sector which is downsized from 20% to about 16%, thus closer to the macro estimates provided by National Accounts which, according to ISTAT, in 2010 was around 12%[18].

The expenditure income profile (Figure 4.7, upper panel) shows a synthetic distribution characterized by a higher propensity to consume for middle and upper-middle classes and thus a greater curvature. In the lower panel of the same figure it is possible to observe that this result implies a higher asymptotic tendency for the synthetic propensity to consume compared to SHIW figure.

Figure 4.6: Total expenditure age profile



[17] It has to be remarked that none of these studies, including ours, can rely on standardized procedures free from *ad hoc* hypotheses. In our work, the adjustments adopted to annualize consumption can determine an underestimation of the purchase frequencies over the year; however, previous works, comparing the average monthly HBS expenditure with the SHIW one can result in an upward bias since they treat goods with frequencies and amount variable over the year (such as clothing) as constantly purchased every month.

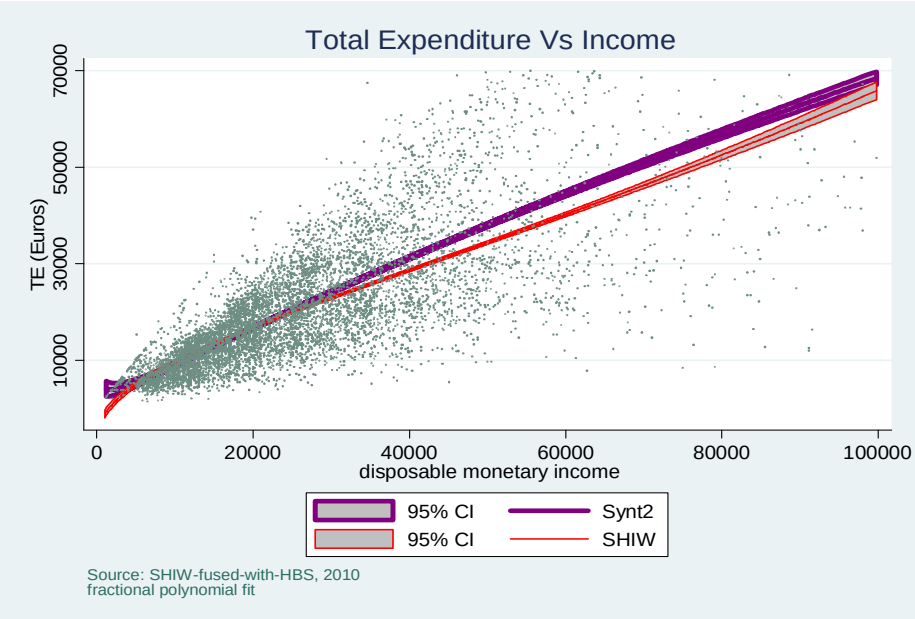
[18] The discrepancy between micro and macro estimates in terms of aggregate economic figures is a well-known problem in literature.

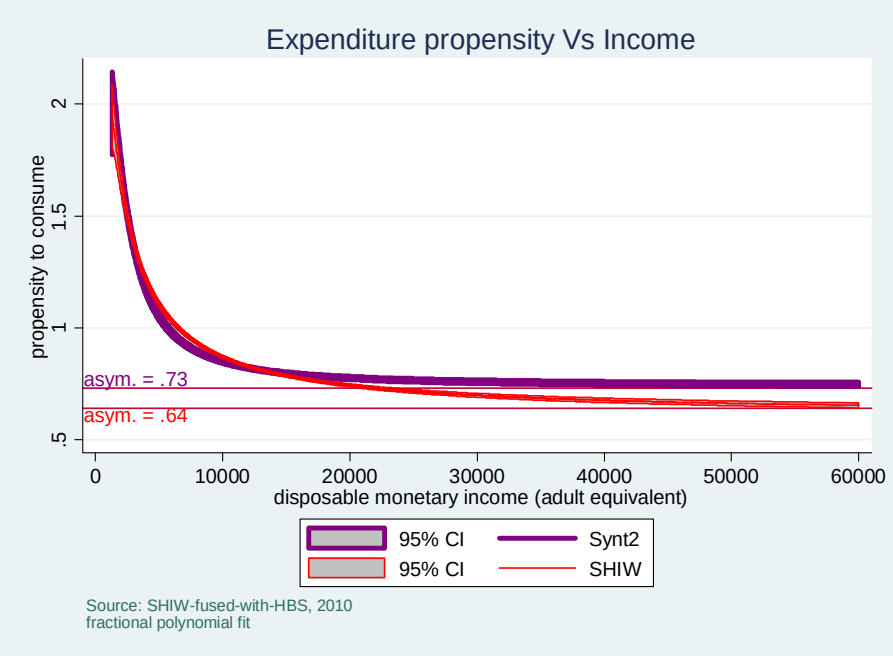
Table 4.7: TE adjustment due to non-durable corrections (Synt2)

Dependent var.: ▫TE*	Coef.	Std. Err.	P>t
Age	0.004	0.002	0.006
age2	0	0	0.003
Income dec3	0.033	0.013	0.014
Income dec4	0.039	0.012	0.001
Income dec5	0.064	0.012	0
Income dec6	0.079	0.013	0
Income dec7	0.08	0.012	0
Income dec8	0.102	0.013	0
Income dec9	0.127	0.017	0
Income dec10	0.075	0.017	0
log{debt}	0.001	0.001	0.067
Manifacturing	-0.041	0.018	0.021
Building and constr.	-0.025	0.011	0.021
Middle-school	0.02	0.01	0.05
High school	0.029	0.01	0.003
Degree	0.041	0.015	0.006
Tenant	0.029	0.014	0.037
Unemployed	-0.038	0.013	0.003
office work, teacher	-0.02	0.01	0.042
Single	-0.02	0.009	0.027
Real estate, professional and business activities	0.078	0.046	0.089
Widow	-0.024	0.011	0.024
_Intercept	-0.163	0.039	0

$\blacksquare TE = (TE_{synt2} - TE_{shiw}) / TE_{shiw}$

Figure 4.7: Expenditure and consumption propensity income profiles





Upper panel: Expenditure income profile; Lower panel: Expenditure propensity.

Table 4.7: Aggregate propensity to consume/save

	Aggregate Consumption propensity	Aggregate propensity to save
SHIW	0.797	**0.203**
Synt2	0.836	**0.162**

5. Conclusions

This work aims at providing a reliable dataset characterized by high details for income sources and consumption items vector for distributional and microeconometrics purposes. Moreover, it aims to deliver a data source for integrated direct and indirect tax-benefit microsimulation models for Italy.

Such goals are pursued by imputing household consumption information from the *Indagine sui Consumi delle Famiglie* (Household Budget Survey, HBS) by the Italian National Statistical Institute (ISTAT) to the *Indagine sui Bilanci delle Famiglie Italiane* (Survey of Households' Income and Wealth, SHIW) by the Bank of Italy, for the year 2010, using matching techniques based on the propensity score method.

We deal with a particular matching problem, compared to the traditional case. In our case, indeed, the information on consumption to impute is observed, though in a less disaggregated way, also in the recipient file itself (SHIW), thus allowing to include some aggregate consumption expenditure in the common variables control vector used for the matching.

The study offers a careful analysis of the quality of information in the two surveys and reveals that, despite the focus on consumption and the high degree of details of the HBS, some consumption items are more reliably recorded in SHIW. Thus, as a first step, we aimed at fusing at best recipient and donor units so as to impute disaggregated expenditure items to the former

sample; on the other hand, we aim at creating a dataset which borrows the best information from the two files. This asks for discarding some imputed aggregates while retaining the original SHIW ones when the analysis suggests this latter source provides a more reliable picture of the true distributions. However, the imputed vector of goods can be used for the internal partition of SHIW more reliable aggregates.

In sum, our matching achieves a good preservation of the marginal distributions of all consumption aggregates from the donor. However, a thorough comparison of the original distributions suggests that the HBS is a convenient donor for the imputation of non-durable commodities only.

Durable expenditures, in particular those related to the purchase of means of transports and to the extraordinary maintenance of household properties, are dramatically undersized compared to SHIW despite our correction on frequencies, while other durables are too heterogeneous between the sources to be corrected. Therefore, overall, durable consumption - which, however, represents a small share of total household expenditure - seems better assessed by the recall method of this latter survey.

For other items related to the concepts of savings and wealth as well (e.g. mortgages and private pensions etc...) SHIW seems to provide a more reliable picture both in terms of frequencies and amounts.

Considering higher levels of validity for the data fusion, our metrics suggest a satisfying preservation of consumption, income and wealth correlations observed in the recipient sample.

As secondary implications, the information derived from HBS on non-durables entails an increase in the dispersion and an upward adjustment of consumption profiles in the synthetic distribution relative to SHIW. This implies also a downsized average propensity to save for the household sector which gets closer to the National Accounts figures.

References

Battistin, E., Miniaci, R., Weber, G. (2003), "What Do We Learn from Recall Consumption Data?", Journal of Human Resources, 38(2), 354-385.

Brandolini, A. (1999), "The distribution of personal income in post-war Italy: source description, data quality, and the time pattern of income inequality", Temi di discussione, n. 350, Bank of Italy, Rome.

Cannari L., Faiella I. (2008), "House prices and housing wealth in Italy, in: Household Wealth in Italy", Banca d'Italia (ed.), Rome, 91-110.

Cifaldi, G., Neri, A. (2013), "Asking income and consumption questions in the same survey: what are the risks?," Temi di discussione n.908, Bank of Italy, Rome.

Cimino E., Coli A. (1998a), "La Sam come schema per l'integrazione tra conti economici e informazioni di natura sociale. Un esercizio per il 1990" Iscona-Istat, Roma, 30th October 1998.

Cimino E., Coli A. (1998b), "The compilation of a social accounting matrix for Italy", 25th General Conference of The International Association for Research in Income and Wealth. Cambridge, UK 23th- 29th August 1998.

Cimino E., Coli A. (1998c), "Schema di integrazione dei conti nazionali nella SAM, con dati socioeconomici", Rapporto finale per il progetto CNR "Misure e parametri per la politica economica e sociale".

Coli A., Colombini, S. , Di Zio, M., D'Orazio, M., Faiella, I., Siciliani, I., Sacco, G., Scanu, M., Tartamella, F. (2006), "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", Istat, collana Documenti n. 12;

D'Agostino, R.B. (1998), "Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group", Stat Med.,17, 2265–2281.

D'Orazio, M., Di Zio, M., Scanu, M. (2004), "Statistical matching and the likelihood principle: uncertainty and logical constraints", Technical Report Contributi 2004/1, Istituto Nazionale di Statistica, Roma.

D'Orazio, M., Di Zio, M. e Scanu, M. (2006), *Statistical Matching: Theory and Practice*. Chichester, England, and Hoboken, NJ: Wiley.

Leuven, E. and Sianesi, B. (2003), "PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing", http://ideas.repec.org/c/boc/bocode/s432001.html.

Montrone, S., Perchinunno, P., De Blasi, L., L'Abbate, S. (2011), "Le tecniche di integrazione di dati per lo studio della povertà", Annali del Dipartimento di Scienze Statistiche "Carlo Cecchi", Università degli Studi di Bari Aldo Moro - Vol. X: 51-70.

Rässler, S., Fleischer, K. (1998) "Aspects Concerning Data Fusion Techniques", ZUMA Nachrichten Spezial, 4, 317-333.

Rässler, S. (2002) *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*. New York: Springer.

Rässler, S. (2004), "Data Fusion: Identification Problems, Validity, and Multiple Imputation", Austrian Journal of Statistics, 33(1,2), 153-171

Rodger, W.L. (1984), "An Evaluation of Statistical Matching", Journal of Business and Econometric Statistics, 2, 91-102.

Rosati, N. (1998), "Matching statistico tra dati Istat sui consumi e dati Bankitalia sui redditi per il 1995", Dipartimento di scienze economiche "M. Fanno", Università degli studi di Padova.

Rosenbaum, P.R., Rubin, D.B. (1983), "The Central Role of the Propensity Score in observational Studies for Causal Effects", Biometrika, 70(1), 41-55.

Rosenbaum, P.R., Rubin, D.B. (1984), "Reducing Bias in Observational Studies using Subclassi□cation on the Propensity Score", Journal of the American Statistical Association, 79, 516-524.

Sianesi, B. (2001) "Implementing propensity score matching estimators with Stata", available at http://fmwww.bc.edu/RePEc/usug2001/psmatch.pdf

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, 19, 59-79.

Sisto A. (2006), "Propensity score matching: un'applicazione per la creazione di un database integrato ISTAT-Banca d'Italia", Working Paper n.63, Dipartimento di Politiche Pubbliche e Scelte Collettive – POLIS, Università del Piemonte Orientale.

Vousten, R., de Heer, W. (1998), "Reducing non-response: the POLS fieldwork design", Netherlands official statistics, 13, 16-19.

Appendix A. Data issues

A.1 SHIW

The Bank of Italy's Survey of Households Income and Wealth (SHIW) is considered the official source for distributional analysis.

The survey collects information on economic situation - income and wealth (since 1987), savings and consumption behaviour (since 1980) - and social features of a sample of families in the period 1966-2010. Sample size varies from 3000 families in the 1966 to about 8000 since 1986. In 2010, the base year for the analysis, sample size amounts to 19,836 individuals and 7,951 households.

Since 1989, a panel section composed of households already interviewed in the previous wave is provided for. The panel size was 15% of the sample in the 1989 but increased over time to reach the 45% in the 1995. Moreover, since 1995 people leaving a family included in the panel and creating a new family were included too (Brandolini, 1999). In 2010, 4,621 out of nearly 8000 are panel.

The sampling scheme is organized in two-stages: firstly, primary sampling units (municipalities) are split into 51 strata defined by regions and population size. Municipalities are drawn according to this stratification; in a second step households are randomly selected within the stratum.

The Historical Archive used for the analysis collects waves since 1977 (no micro-data are available for earlier years) and provides files containing income and wealth and consumption adjusted according to homogeneous definitions (excluding variables which were not collected in a systematic way) both at household and individual level; weights aligning socio-demographic distributions with ISTAT population statistics and labour force survey (post-stratification) are also provided for (Brandolini, 1999).

The survey unit is the household, i.e. "group of individuals linked by ties of blood, marriage or affection, sharing the same dwelling and pooling all or part of their incomes" (Brandolini, 1999); however, as most information are gathered at individual level, analyses on personal variables are allowed as well.

Most of SHIW incomes are net of taxes and social security contributions, hence it does not provide any information on tax and redistribution.

The survey contains information on disposable income from several sources such as wages, pensions, self-employment/business income (including family firms, unincorporated companies shareholders returns) and social or private transfers, in addition to imputed rents for owner occupied dwellings, actual rents and capital incomes (interest, dividends and capital gains).

The high level of details on personal income sources allows larger or thinner definitions aggregating single income sources to be specified according to the aim of the analysis.

A lower degree of detail is reserved to consumption, which is recorded by means of macro aggregates such as food, other non-durables, and durables (valuables, transports, electrical appliances). In addition, a general question on the monthly expenditure on all items (excluding main durables, rents, mortgage installment, insurance payments separately recorded) is offered.

Finally, special sections are devoted to real and financial wealth. In particular, they provide details on main dwelling and other properties owned by households, together with several figures on real and financial liabilities.

A.2 HBS

The Household Budget Survey (HBS) by Italian National Institute of Statistics (ISTAT) collects a rich set of information on both socio-demographic characteristics and detailed information on consumption behaviour of a cross-section of Italian households for a very disaggregated set of commodities (durables and non-durables) such as food, dwelling, furniture, clothing, health, transport, communication items, recreational goods, education, holidays, etc. Up to 1996 the survey included 77 categories of items, while since 1997 goods are grouped in 273 classes. In fact, in 1997 both the survey design and the procedure for acquisition and validation of results have undergone a deep process of revision in order to align definitions and methodology to the recent European precepts and to improve quality of data.

The sampling scheme is organized in two-stages:

1) firstly, municipalities are selected among two groups according to the size of population; chief towns of provinces are fully included and selected to take part to the survey every month, while the remaining are grouped in strata according to some economic and geographic characteristics and are extracted every 3 months;

2) in a second step households are randomly selected within the stratum from the registry office records.

As a result, the survey unit is the legal family recorded by the registry office.

Sample size is around 28,000 households from 480 municipalities and weights allowing for a re-calibration of population in each stratum and for the distribution by household size within region are also provided for.

Data are recorded by means of two complementary methods: a) a diary where the household keeps track of expenditures made (Libretto degli Acquisti) and of quantities of internally produced goods consumed in the previous 7 days (Taccuino degli Autoconsumi); b) a proper interview for the remaining purchases done in the previous month and for durables bought in the previous 3 months. It has to be remarked that expenditure is provided on a monthly basis, so commodities recorded on a wider recording period are made monthly in the survey by dividing the amount for the number of months they are recorded for (durables are divided by a factor of 3). This feature has required some delicate adjustments both on amounts and frequency (see section 4.2) in order to work on an yearly basis.

Given the high degree of detail, the survey represents the official source for the construction of cost-of-living indices and the production of poverty (absolute and relative) consumption-based statistics in Italy.

Since 1979 a purely indicative question concerning household monthly income (by range) has been introduced in the questionnaire (not reported in the survey); however, unfortunately, the reliability of such information is rather limited due to a high under-reporting which undermines the estimations.

Appendix B. Comparison of variables common to the two datasets

Table B.1: Household head age group distribution SHIW vs HBS

HH Age class	HBS	SHIW	SHIW-HBS
3=15-17	0.01		
4=18-24	0.55	0.81	0.26
5=25-29	2.19	2.36	0.17
6=30-34	5.53	5.3	-0.23
7=35-39	8.9	6.75	-2.15
8=40-44	10.27	12.34	2.07
9=45-49	10.48	10.16	-0.32
10=50-54	9.77	9.66	-0.11
11=55-59	9.1	8.15	-0.95
12=60-64	9.08	9.96	0.88
13=65-69	7.76	8.21	0.45
14=70-74	8.54	9.53	0.99
15=75 and over	17.83	16.79	-1.04
Total	100	100	

Table B.2: Distribution by region of residence

Region	HBS	SHIW	SHIW-HBS
1+2=Piemonte e Valle d'Aosta	8.25	9.75	1.5
3=Lombardia	17.09	15.1	-1.99
4=Trentino Alto Adige	1.71	4.86	3.15
5=Veneto	8.06	6.49	-1.57
6=Friuli Venezia Giulia	2.23	2.02	-0.21
7=Liguria	3.15	3.81	0.66
8=Emilia Romagna	7.8	6.4	-1.4
9=Toscana	6.43	6.86	0.43
10=Umbria	1.5	1.48	-0.02
11=Marche	2.56	2.6	0.04
12=Lazio	9.32	8.94	-0.38
13=Abruzzo	2.16	1.59	-0.57
14=Molise	0.52	1.1	0.58
15=Campania	8.38	7.52	-0.86
16=Puglia	6.13	6.12	-0.01
17=Basilicata	0.92	2.73	1.81
18=Calabria	3.1	2.99	-0.11
19=Sicilia	7.95	6.53	-1.42
20=Sardegna	2.73	3.1	0.37
Total	100	100	

Table B.3: Distribution of number of family members

n. components	HBS	SHIW	SHIW-HBS
1	30.3	25.56	-4.74
2	27.36	30.79	3.43
3	20.29	19.51	-0.78
4	16.79	18.01	1.22
5	4.09	4.59	0.5
6	0.87	1.42	0.55
7	0.21	0.04	-0.17
8	0.08	0.07	-0.01
9	0.02		
10	0.01		
12	0	0.01	0.01
Total	100	100	

Table B.4: Distribution by household type

Household typology	HBS	SHIW	SHIW-HBS
1= Lone person with aged 35 or less	3.18	2.77	-0.41
2= Lone person with aged 35-64	12.08	9.3	-2.78
3= Lone person with aged 65 or more	15.05	13.49	-1.56
4= Couple without children with reference person aged 35 or less	1.44	1.71	0.27
5= Couple without children with reference person aged 35-64	7.87	7.74	-0.13
6= Couple without children with reference person aged 65 or more	10.95	14.29	3.34
7= Couple with 1 child	16.61	18.13	1.52
8= Couple with two children	15.28	16.45	1.17
9= Couple with three of more children	3.68	4.71	1.03
10= Single-parent	8.18	4.62	-3.56
11= Other typologies	5.68	6.79	1.11
Total	100	100	

Table B.5: HH marital status distribution

HH Marital status	HBS	SHIW	SHIW-HBS
1 = married	58.22	61.93	3.71
2 = single	17.04	14.08	-2.96
3 = separated/divorced or widower/widow	24.74	24	-0.74
Total	100	100	

Table B.6: Distribution of educational level of HH

HH Educational level	HBS	SHIW	SHIW-HBS
1 = none	4.32	4.11	-0.21
2 = elementary school	22.66	21.77	-0.89
3 = middle school	34.87	36.74	1.87
4 = high school	26.3	25.78	-0.52
5 = bachelor's degree	10.76	10.39	-0.37
6 = post-graduate	1.1	1.2	0.1

qualification.			
Total	100	100	

Table B.7: Distribution of HH by occupational status

HH Occupational status	HBS	SHIW	SHIW-HBS
0 = employed	52,08	53,39	1,31
1 = first-job seeker	0,32	0,28	-0,04
2 = homemaker or pensioner	42,41	42,69	0,28
3 = unemployed	2,72	2,99	0,27
4 = student	0,41	0,45	0,04
5 = other not employed (including well-off)	2,06	0,19	-1,87
Total	100	100	

Table B.8: Distribution of HH by branch of activity

HH Branch of activity	HBS	SHIW	SHIW-HBS
1 = agriculture	2.4	2.32	-0.08
2 = manufacturing	9.48	10.37	0.89
3 = building and construction	6	5.15	-0.85
4 = wholesale and retail trade, lodging and catering services	10.72	8.63	-2.09
5 = transport and communication	3.48	2.67	-0.81
6 = services of credit and insurance institutions	1.46	1.98	0.52
7 = real estate and renting services, other professional and business activities	4.42	2.97	-1.45
8 = domestic services and other private services to persons	5.45	7.54	2.09
9 = general government, defence, education, health and other public services	11.21	11.46	0.25
10 = extra-territorial	0.2	0.11	-0.09
11=not employed	45.2	46.8	1.6
Total	100	100	

Table B.9: Distribution of HH by work status

HH Work status	HBS	SHIW	SHIW-HBS
1 = blue-collar worker or similar	17.59	19.82	2.23
2 = office worker or school teacher	17.42	16.27	-1.15
3 = junior manager/cadre	3.36	2.72	-0.64
4 = manager, senior official	2.48	1.68	-0.8
5 = member of the arts or professions	2.65	2.99	0.34
6 = sole proprietor	2.25	1.3	-0.95
7 = freelance	6.04	6.03	-0.01
8 = owner or member of a family business	0.17	1.67	1.5
9 = active shareholder/partner	0.13	0.9	0.77
10 = not employed.	47.92	46.61	-1.31
Total	100	100	

Table B.10: Distribution by resident status

Resident status	HBS	SHIW	SHIW-HBS
1 = home owner or with the right of redemption	73.63	68.82	-4.81
2 = tenant	17.23	20.66	3.43
3 = with right of usufruct, use without charge	9.15	10.52	1.37
Total	100	100	

Table B.11: Second dwellings

Second dwellings	HBS	SHIW	SHIW-HBS
0	91.85	85.04	-6.81
1	7.35	11.71	4.36
2	0.67	2.27	1.6
3	0.11	0.63	0.52
4	0.02	0.24	0.22
5		0.06	0.06
6		0.01	0.01
8		0.02	0.02
9		0.01	0.01
Total	100	100	

Table B.18: Location of the main dwelling

Location of the main dwelling	HBS	SHIW	SHIW-HBS
1= city	80.41	86.42	6.01
2=small town, village	13.41	6.71	-6.7
3= hamlet, detached houses, farm area	6.18	6.87	0.69
Total	100	100	

C. Main matching hypotheses

Donor file	HBS 2010 (Istat) Sample size: 22,246 households
Recipient file	SHIW 2010 (Bank of Italy) Sample size: 7,951 households 19,836 individuals
Matching method	Propensity score (PS)
Matching algorithms	1) Nearest neighbor (NN) 2) Mahalanobis metric (M)
Stratification	50, 80 or 100 cells
Strata variables	1) Total Matching Consumption (TMC*) quantiles (5, 8 or 10) times 2) Household typology (10 modalities)
Common vars.	o **categorical**: age class, region, gender, n.components, marital status, education, occupational status, branch of activity, work status o **continuous**: food consumption, real consumption, vehicles, extraordinary exps, imputed, actual rents (1° dwell. only)

D. Final synthetic definitions of household consumption

	Description	Synt	Synt2	Benchmark
Total matching consumption (TMC)	Food, non-food-non-durable expenditures; durables, real goods	All items resulting from statistical matching	NO	HBS
Total expenditure (TE)	Non-durables (food and other non-durables); durables; real goods; health and life insurances; payments for mortgages; deposit to private pensions.	All items resulting from statistical matching	Synthetic values resulting from matching for non-durables expenditure only	SHIW
Overall consumption (OC)	TE + 1st and 2nd dwellings imputed rents			SHIW

Pubblicato in proprio
Dipartimento di Economia Pubblica
Facoltà di Economia
Università degli Studi di Roma “La Sapienza”
Via del Castro Laurenziano 9 – 00161 Roma

ISSN 1974-2940

Dipartimento di Economia Pubblica
Università degli studi di Roma "La Sapienza"
Via del Castro Laurenziano 9 – 00161 Roma