Data Fusion Between Bank of Italy-SHIW and ISTAT-HBS

Simone Tedeschi and Elena Pisano

Department of Economics and Law, Sapienza University of Rome, Bank of Italy, Tax Department

October 2013

Data Fusion Between Bank of Italy-SHIW and ISTAT-HBS

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

[i] Department of Economics and Law, Sapienza University of Rome, Italy.
e-mail: simone.tedeschi@uniroma1.it

[ii] Bank of Italy, Tax Department. The opinions expressed here are those of the author and do not necessarily reflect the positions of the Institution.

Introduction

Propensity score matching (PSM, Rosenbaum and Rubin, 1983) has become a quite 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 (i.e. to increase the dimensions of a distribution of characteristics) 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 - are usually found by means of PSM-nearest neighbour or hot deck procedures which serve to synthesize their multidimensional distance/similarity along some common variables in a one-dimensional space.

This approach is not immune from drawbacks since it assumes the conditional independence of the variable not jointly observed given common variables (conditional independence assumption, CIA henceforth). 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.

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 matching technique[3].

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.

[3] A project aiming at integrating these two sources was already tackled by a joint ISTAT-Bank of Italy working group in 1998. It evaluated the feasibility of statistical matching of ISTAT survey on Consumption and Bank of Italy survey on Income and Wealth, 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. As far as we know, this project was not carried on in the following years, especially with regard to the micro level, after the ISTAT survey revision.

In particular, we combine information from the Historical Database (integrated with information from the original cross sectional files) of SHIW 2010 with the wave 2010 of HBS.

This work aims at providing an integrated synthetic dataset in order to jointly analyze income, wealth and consumption distributions with a high degree of detail for both incomes-assets and consumption expenditure items. The resulting sample is expected to allow better multidimensional-distributional analyses on consumption income and wealth. The following spillover will be an integrated microsimulation analysis of direct, indirect and wealth tax reforms which, so far, has not been feasible taking available sample surveys separately.

The link function is based on a set of common characteristics (*Z_i)* surveyed both in SHIW (*Shiw*) and HBS *(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 coupled with pseudo-random lottery procedures of imputation (estimation, prediction and Monte Carlo techniques) in some cases. The choice of a proper common support of variables has represented a crucial task to accomplish since, as we will show at length, the matching problem we deal with slightly differs from the typical data fusion situation, while it presents some advantageous characteristics. In fact, the two surveys share the total consumption information, even though with a different degree of detail.

The unit of analysis for the matching process is the household; in particular, most of the variables in the common vector refers to the household head[4].

As a matching algorithm we use propensity score coupled with a *Mahalanobis* metric for the $\boldsymbol{Z}_i$ variables. We perform a matching with replacement, assigning to each SHIW household a value of the main consumption components derived from the “nearest” HBS household in terms of the common characteristics. Therefore, some “less similar” HBS units will be discarded by the matching procedure.

We match one or more units of HBS dataset to one unit of SHIW according to some assumptions, which allow to preserve - in the synthetic sample - the main marginal and conditional distributions observed in SHIW as well as the covariance structures, and, simultaneously, reproducing closely unconditional distributions (and their covariance structure) of all the main consumers' items observed in HBS dataset.

The work is structured as follows: in the first section a description of the two survey and a description of the main assumptions and specificities of our matching problem is offered. Secondly, a brief review of the data imputation methodologies is reported. In the third section the preparatory tasks of the matching procedure are illustrated, coupled with a discussion of the criticalities on the

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

durable reporting and some empirical evidence from the two surveys. Finally, in the fifth and final paragraph, 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.

1. Data issues

1.1 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.

1.2 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.

2. Main assumptions and specificities of our matching problem

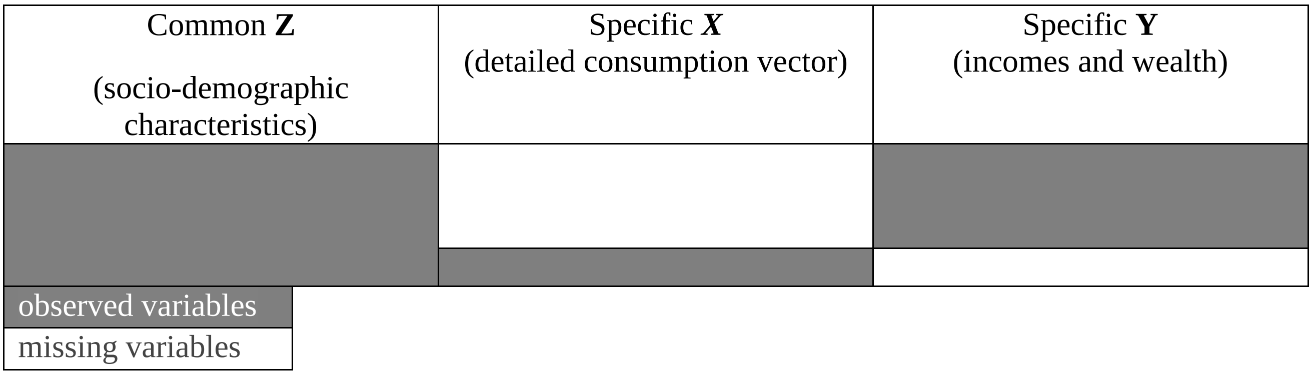
The typical situation a statistician or applied micro-economist faces is that of being interested in the joint (or conditional) distribution of three (vectors of) variables $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ but no database exists where such three variables are simultaneously observed. 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; 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.

Rässler (2002) lists different situations which can be tackled by statistical matching. In particular, she analyzes the situation referred as ‘data fusion’ - i.e. the case in which there are groups of variable that are never jointly observed (say $\boldsymbol{X}$ and $\boldsymbol{Y}$) in a sample. In this case, the dataset resulting by the matching is aimed at making the analysis of the unobserved relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ feasible. She shows how 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.

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 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 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 situation for data fusion



According to this scheme, the main problem we would deal with 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 rest 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.

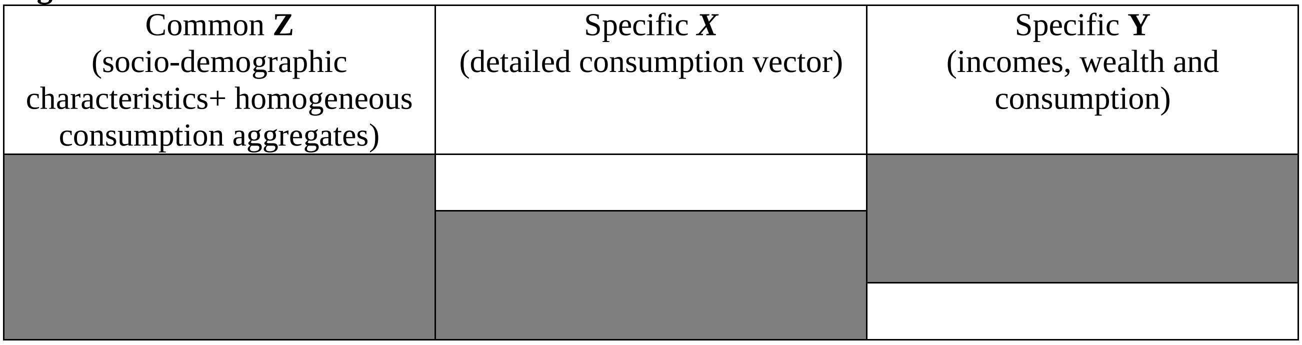
Still, a key feature of the problem we take on is that both surveys include information on household consumption; however, while HBS is focused on this issue by specifically providing data on single household consumption goods and services, SHIW gathers information on consumption at a lower level of 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 HBS[5]. 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).

In particular, the joint available observation of $\boldsymbol{C^a}$, $\boldsymbol{I}$ and $\boldsymbol{W}$ and thus their correlation structure. That is, as a second assumption, 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 (SHIW) determines a situation which is different (and more advantageous) compared to that represented in Figure 2.1. This situation can be represented by Figure 2.2 where an overlapping of $\boldsymbol{X}$ and $\boldsymbol{Y}$ exists that does not flow directly into $\boldsymbol{Z}$. Indeed, it has to be remarked that hundreds of consumption items surveyed with a recall period of one month or one quarter cannot be simply recoded 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 and flow into $\boldsymbol{Z}$ so as to recover a benchmark for the $\boldsymbol{X}$,$\boldsymbol{Y}$ correlation structure. Hence we can reproduce $\boldsymbol{C^a}$ in HBS as a vector of consumption commodities blocks' sums i.e. $\boldsymbol{C^a} = [\sum_{\boldsymbol{k=1}}^{\boldsymbol{K_1}} \boldsymbol{c_k}, \sum_{\boldsymbol{k=K_1+1}}^{\boldsymbol{K_2}} \boldsymbol{c_k}, \dots, \sum_{\boldsymbol{k=K_A+1}}^{\boldsymbol{K}} \boldsymbol{c_k}]$.

Figure 2.2: Our situation for data fusion


Common **Z**
(socio-demographic characteristics+ homogeneous consumption aggregates)
Specific ***X***
(detailed consumption vector)
Specific **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 (AI, see Singh et al., 1993) where AI on *f(**X**,**Y**,**Z**)* is recovered from the recipient dataset rather than from a third data source.

[5] Although, as we shall see later, limited to non-durable high frequently bought items.

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 HBS represents at the same time the target missing variable to be integrated 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 rather poorly comparable (see section 4.2 on the treatment of durable goods).

3. Matching techniques and algorithms

Following Rubin (1977) underlying mechanism that generates missing data can be considered either ignorable or non-ignorable. One unique situation in which missingness completely at random (MCAR) may be reasonably expected to hold is when missing data are induced by the design. In this sense data fusion can be conceived as an imputation problem or, following Rässler (2002), a mass imputation.

As previously discussed, the identification problem of the association among variables not-jointly-observed related to the data fusion will strictly depend on the explanatory power of the common variable ($\boldsymbol{Z}$). A strong explanatory power of $\boldsymbol{Z}$ makes the CIA easier to hold. An alternative approach draws the unbiasness of the integration on the availability of 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 fuse. 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.

On the grounds of this fact, 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.

In the next subsection, we discuss the use of propensity score for the sake of data fusion and describe how we address our specific matching problem.

3.1 Propensity score

Traditionally, propensity score methods (PSM) serve the purpose of analyzing causal effects of treatment (such as policies) from observational data. To analyze such data, an ordinary least square regression model using a dichotomous indicator of treatment is probably unsuitable, because the error term is likely to be correlated with 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) 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 in our context in order to achieve the goal of a ‘multidimensional imputation’ in terms of a large missing data problem, rather than an instrument to estimate policy treatment effects.

If we had to impute one variable only to the SHIW sample from HBS we might think about this problem in terms of imputation of a missing information through regression techniques. Actually, we do not aim at achieving a full integration between the two dataset to obtain a sample which is the sum of both neither. Rather, given our future 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[6]. The synthetic data set is thus just the completed SHIW file, while a significant amount of records as well as important sample information on $\mathbf{Z}$ is discarded from HBS. On the opposite, whether the overall sample SHIW ∪ HBS was used for inference, the effect of *matching noise*[7] would be rather magnified.

As discussed in previous section, we have to deal with a particular matching problem, compared to the traditional case. In our case, indeed, the information on consumption we want to impute to SHIW is observed, though in a less disaggregated way, also in the SHIW file itself, thus allowing us to use some aggregates consumption expenditure in the common $\boldsymbol{Z}$ vector. In addition, providing a thinner classification of consumption aggregates, we conceive HBS to deliver a more accurate representation of the true distribution of some consumption aggregates (the ones homogenous to SHIW's $\boldsymbol{C^a}$, i.e. food, other non durables, and to a lesser extent, durables).

Therefore, we aim at preserving in the fused file the marginal distributions of the target variables from the donor sample, i.e. $\tilde{f}(\boldsymbol{X}) \cong f(\boldsymbol{X})$ as well as $\tilde{f}(\boldsymbol{X}, \boldsymbol{Z}) \cong f(\boldsymbol{X}, \boldsymbol{Z})$, where $\sim$ indicates values and parameters produced by the fusion procedure. At the same time, we attempt to minimize the

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

[7] 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).

difference between the joint distribution $f(\boldsymbol{X^a}, \boldsymbol{Y}, \boldsymbol{Z})$ observed in SHIW and the resulting joint synthetic distribution$\tilde{f}(\boldsymbol{X^a}, \boldsymbol{Y}, \boldsymbol{Z})$, where $\boldsymbol{X}^{a}$=$\boldsymbol{C}^{a}$.

The resulting file will be representative of the same population of the *recipient* file (and will use its weights) but it will be enriched with new information coming from *donor* units which are one-by-one similar to those of the *recipient* file.

As we need to impute several variables (*i.e.* 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) is the probability of a unit (belonging to HBS) being assigned to a particular treatment group (SHIW) given her characteristics before the treatment, that is:

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

Therefore, the STATA code used[8] first runs a logistic (or a probit) regression wherethe dependent variable (Shiw) is equal to 1 if the observation comes from the recipient sample and zero otherwise conditional on on the selected (instrumental) variables (***Z***). the *propensity score* is then the predicted probability (p) or log[p/(1 − p)] resulting from this stage. 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 equal to *Hbs*=1) (D’Agostino, 1998).

We then use two alternative definitions of distance :

1a) Nearest neighbor 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) Caliper matching

[8] PSMATCH2 matching algorithm, Leuven and Sianesi (2003).

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

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

Actually, the procedure uses a combination of a) and b) that is *nearest neighbor within caliper*. This method is the most simple and intuitive, and consists of matching every recipient (SHIW) units with the donor (HBS) units which have the nearest PS within a fixed radius (*caliper*).

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 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 HBS 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 ***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.

As we want to assign to each SHIW household a vector of consumption components, despite the significant difference in sample size (being 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. This fact will force the algorithm to match all the recipient sample, even replicating donor units, if needed. However, the cost of dismissing a matching with 'no 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 with 'no replacement' option we would not match the whole SHIW sample, unless enlarging too much the radius of acceptability (*caliper*). This fact entails losing the representativeness of the population in the 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 quintiles (deciles) of household total consumption used for matching (TMC, see section 5) 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 5 are related to with *Mahalanobis* metrics, as it is preferred to nearest neighbour method due to a better performance in terms of both conditional variability of target variables and the joint distribution.

4. Methodology

4.1 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, as mentioned in the data issue section, 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 ISTAT 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. 4.1 and Fig. 4.1).

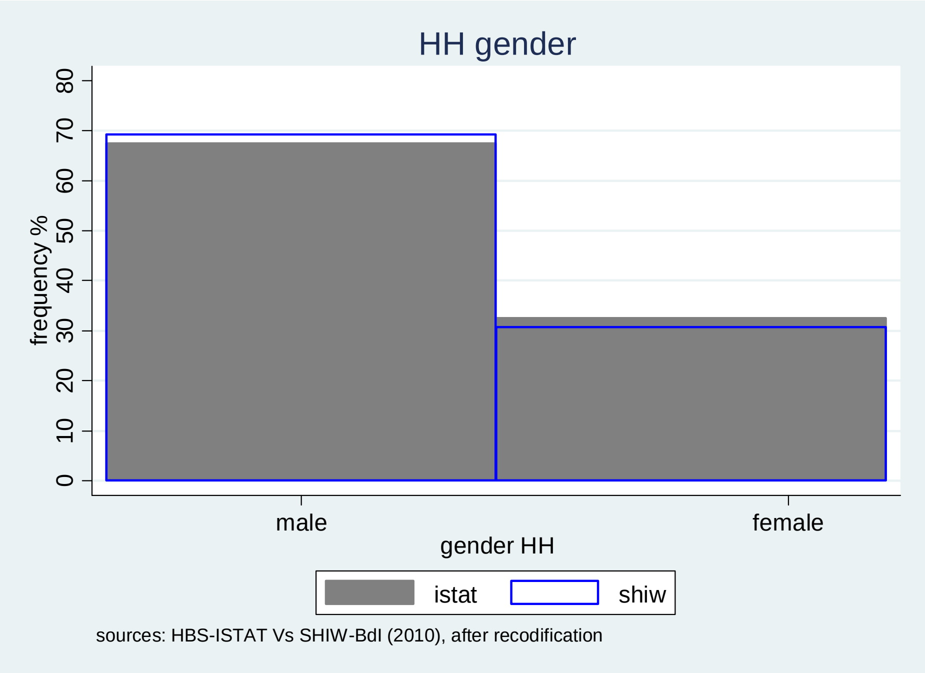
To this end, we decided to perform a recodification of the reference person for matching aims.

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).

Table 4.1: Share of male and female household heads in the two surveys

	HBS			SHIW_pre			SHIW_post		
Female	Freq.	Percent	Cum.	Freq.	Percent	Cum.	Freq.	Percent	Cum.
0	16,698,963	67.46	67.46	13,387,725	55.53	55.53	16693904	69.24	69.24
1	8,056,407	32.54	100	10,722,158	44.47	100	7,415,979.00	30.76	100
Total	24,755,370	100		24,109,883	100		24,109,883.00	100	

Figure 4.1: Share of male and female household heads in the two surveys, after recoding



sources: HBS-ISTAT Vs SHIW-BdI (2010), after recodification

4.2 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 (*consal*) and other non durable (*condiv*) expenditure with a monthly purchase frequency, the yearly amount is simply obtained by multiplying by 12[9]. 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[10]. For durables surveyed in the last three months (mainly transport expenditure), a more demanding correction is implemented. First of all, ISTAT 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

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

[10] 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 rest 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.

three months consumption of these items, HBS severely underestimates the share of individuals purchasing such goods over the year[11] (e.g. transport, Tab. 4.2).

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, imputing the predicted probabilities to HBS households. Implicitly, we are assuming that SHIW deliver a better representation of yearly durable purchases frequency and its determinants. We apply this method to a subset of relevant durables only, the transport-related ones (cars, motorcycles, camping vans, etc… included in aggregate CDUR1). A Monte Carlo simulation is then run to select HBS households effectively doing the purchase among those with no durables expenditure having the highest probabilities of doing it according to the imputed score. We then calibrate the number of households with such purchases as the difference in the share of units with that kind of durables expenditure between the two survey (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 4.2 (right panel) and 4.3.

Table 4.2: share of HBS households spending on transport durables (cdur1>0) before and after the imputation compared to SHIW

	HBS_pre			SHIW			HBS_post		
Cdurpos=(cdur>0)	Freq	Percent	Cum.	Freq.	Percent	Cum.	Freq.	Percent	Cum.
0	24,224,717	97.3	97.3	21,847,636	90.6	90.6	22,549,422	90.6	90.6
1	673,456	**2.7**	100	2,262,247	**9.4**	100	2,348,751	**9.4**	100
Total	24,898,173	100		24,109,883	100		24,898,173	100	

Table 4.3: distribution of CDUR1 in HBS after the imputation, compared to SHIW (over the whole sample and for positive values only)

cdur1									
HBS					SHIW				
	Percentiles	Smallest				Percentiles	Smallest		
1%	0	0			1%	0	0		
5%	0	0			5%	0	0		
10%	0	0	Obs	22246	10%	0	0	Obs	7951
25%	0	0	Sum of Wgt.	22246	25%	0	0	Sum of Wgt.	7951
50%	0		Mean	722.8601	50%	0		Mean	1049.03
		Largest	Std. Dev.	3674.715			Largest	Std. Dev.	4444.469
75%	0	50460.57			75%	0	58000		
90%	0	50460.57	Variance	1.35e+07	90%	0	60000	Variance	1.98E+07
95%	3000	57009.21	Skewness	7.140308	95%	9400	70000	Skewness	6.778935

[11] 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.

99%	18000	57009.21	Kurtosis	67.54449	99%	22000	100000	Kurtosis	75.77432
				cdur1>0					
	Percentiles	Smallest				Percentiles	Smallest		
1%	39.99	12			1%	120	50		
5%	99.99001	12			5%	500	100		
10%	168.99	12	Obs	2022	10%	1800	100	Obs	723
25%	500.01	12	Sum of Wgt.	2022	25%	5000	100	Sum of Wgt.	723
50%	5000.01		Mean	7952.891	50%	10000		Mean	11536.42
		Largest	Std. Dev.	9544.856			Largest	Std. Dev.	9815.833
75%	12600	50460.57			75%	15000	58000		
90%	18999.99	50460.57	Variance	9.11e+07	90%	22400	60000	Variance	9.64E+07
95%	24999.99	57009.21	Skewness	1.78224	95%	28000	70000	Skewness	2.515958
99%	42000	57009.21	Kurtosis	7.165017	99%	50000	100000	Kurtosis	16.18676

Unfortunately, the same procedure of imputation from SHIW cannot be applied for durables contained in the CDUR2 variable, which includes other 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 SHIW one (Tab. 4.4). Moreover, the distributions are significantly different (Tab. 4.5), 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 then likely to be mis-reported due to memory effect).

Table 4.4: share of household with cdur2 in the two survey

	HBS			SHIW		
cdurpos_2	Freq.	Percent	Cum.	Freq.	Percent	Cum.
0	1,402,353	5.63	5.63	15,830,854	65.66	65.66
1	23,495,820	94.37	100	8,279,029	34.34	100
Total	24,898,173	100		24,109,883	100	

Table 4.5: distribution of CDUR2 in the two sources (on the whole sample and among households with positive values only)

				cdur2					
		HBS					SHIW		
	Percentiles	Smallest				Percentiles	Smallest		
1%	0	0			1%	0	0		
5%	0	0			5%	0	0		
10%	40	0	Obs	22246	10%	0	0	Obs	7951
25%	133.99	0	Sum of Wgt.	22246	25%	0	0	Sum of Wgt.	7951
50%	392.465		Mean	732.7815	50%	0		Mean	613.3396
		Largest	Std. Dev.	1143.266			Largest	Std. Dev.	2174.748
75%	901.62	22057.52			75%	500	40000		
90%	1721.29	24603.99	Variance	1.31E+06	90%	1500	40000	Variance	4.73E+06
95%	2513.35	24610.94	Skewness	6.376698	95%	2900	40000	Skewness	10.44901
99%	5116.67	25215.66	Kurtosis	81.47168	99%	8300	50000	Kurtosis	155.3065
				cdur2>0					
	Percentiles	Smallest				Percentiles	Smallest		

1%	0	0			1%	0	0		
5%	51.67	0			5%	0	0		
10%	105.21	0	Obs	2022	10%	0	0	Obs	723
25%	283.33	0	Sum of Wgt.	2022	25%	0	0	Sum of Wgt.	723
50%	675		Mean	1049.988	50%	0		Mean	1248.333
		Largest	Std. Dev.	1216.078			Largest	Std. Dev.	3828.459
75%	1373.32	9923.33			75%	1000	30000		
90%	2423.8	10150.13	Variance	1.48E+06	90%	2500	40000	Variance	1.47E+07
95%	3237.54	10778.15	Skewness	3.131282	95%	5000	40000	Skewness	7.032888
99%	5812.12	12920.7	Kurtosis	19.00826	99%	25000	40000	Kurtosis	60.83293

A similar procedure is carried out for the amount of expenses for extraordinary maintenance of household dwellings (MASTRIP), which are characterized in HBS by a lower frequency and a lower average value of the declared purchase. Table 4.6 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.

The general idea is that durable expenses – i.e. low frequency and usually high level -, being closer to the stock variables pertaining household wealth, might be more reliable in the SHIW so we consider worthwhile to account for that in the analyses that will be based on the synthetic dataset.

The Appendix shows the figures (histograms) comparing the sample distributions of the above-mentioned variables, after the correction (Fig. A8 and A9).

Table 4.6: Mastrip in HBS after the imputation, compared to SHIW

Mastrip									
HBS					SHIW				
Percentiles		Smallest			Percentiles		Smallest		
1%	0	0			1%	0	0		
5%	0	0			5%	0	0		
10%	0	0	Obs	22246	10%	0	0	Obs	7951
25%	0	0	Sum of Wgt.	22246	25%	0	0	Sum of Wgt.	7951
50%	0		Mean	488.6637	50%	0		Mean	1010.746
		Largest	Std. Dev.	2606.871			Largest	Std. Dev.	6725.154
75%	0	63907.08			75%	0	130000		
90%	549.99	63907.08	Variance	6795778	90%	1600	130000	Variance	4.52e+07
95%	2346.99	66781.66	Skewness	12.36181	95%	5000	180000	Skewness	25.77713
99%	12000	86421.3	Kurtosis	234.884	99%	20000	350000	Kurtosis	1059.867

4.3 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 and with a particular focus on consumption variables. This will serve to build a *distance function* to be minimized in orderto match “similar” units from the two original samples.

To build the ***Z*** vector we first append the two subsample obtained by keeping the household heads (HH) only. We recoded variables surveyed in a different way but providing common information on the household or its HH to make them homogeneous.

The subset of controls we obtained after the recoding procedure is the following:

a) ireg_m: region of residence of the family according to ISTAT codification
1+2=Piemonte e Valle d’Aosta
3=Lombardia
4=Trentino Alto Adige
5=Veneto
6=Friuli Venezia Giulia
7=Liguria
8=Emilia Romagna
9=Toscana
10=Umbria
11=Marche
12=Lazio
13=Abruzzo
14=Molise
15=Campania
16=Puglia
17=Basilicata
18=Calabria
19=Sicilia
20=Sardegna

b) ncomp: number of household component

c) tipfam: household typology:
1= Lone person with aged 35 or less
2= Lone person with aged 35-64
3= Lone person with aged 65 or more
4= Couple without children with reference person aged 35 or less
5= Couple without children with reference person aged 35-64
6= Couple without children with reference person aged 65 or more
7= Couple with 1 child
8= Couple with two children
9= Couple with three of more children
10= Single-parent
11= Other typologies

d) female: dummy=1 if female HH

e) eta15_hh: HH age classes:
1=0-5
2=6-14
3=15-17
4=18-24
5=25-29
6=30-34
7=35-39
8=40-44
9=45-49
10=50-54

11=55-59
12=60-64
13=65-69
14=70-74
15=75 and over

f) staciv_hh: HH marital status:
1 = married
2 = single
3 = separated/divorced or widower/widow.

g) studio_hh: HH educational level:
1 = none
2 = elementary school
3 = middle school
4 = high school
5 = bachelor's degree
6 = post-graduate qualification.

h) condprof_hh: HH occupational status:
0 = employed
1 = first-job seeker
2 = homemaker or pensioner
3 = unemployed
4 = student
5 = other not employed (including well-off)

i) qualp_hh: HH main employment, work status:
employee:
1 = blue-collar worker or similar
2 = office worker or school teacher
3 = junior manager/cadre
4 = manager, senior official
self-employed:
5 = member of the arts or professions
6 = sole proprietor
7 = freelance
8 = owner or member of a family business
9 = active shareholder/partner
10 = not employed.

j) sett_hh: HH main employment, branch of activity:
1 = agriculture
2 = manufacturing
3 = building and construction
4 = wholesale and retail trade, lodging and catering services
5 = transport and communication
6 = services of credit and insurance institutions
7 = real estate and renting services, other professional and business activities
8 = domestic services and other private services to persons
9 = general government, defence, education, health and other public services
10 = extra-territorial

k) consal: yearly amount of expenditure on food

l) condiv: yearly amount of expenditure on other non-durables

m) creali: yearly amount of real goods bought (jewellery, old and gold coins, works of art, antiques, including antique furniture)

n) cdur1: amount of expenditure on means of transport (see section 3.2)
o) cdur2: amount of expenditure on other durables (furniture, furnishings, appliances, etc.)
p) affimp_p1 (e p2): annual rent potentially receivable (imputed rent) on first (and second) houses.
q) affpag: annual rent actually paid.

In addition several other variables are available in both survey and have been recoded (not used as controls):
r) sec_case: dummy property second houses
s) godab_m: resident status of main dwelling
1 = home owner or with the right of redemption
2 = tenant
3 = with right of usufruct, use without charge
t) mastrip1 and 2 = extraordinary maintenance of principal residence and second houses (see section 4.2)
v) ubic: location of the main dwelling
1= city
2= small town, village
3= hamlet, detached houses, farm area
w) ancostr: year of construction of the building of main dwelling
x) anpos: year the household has become owner of the main dwelling
y) godabit: type of property right
1= owned by the household
2= rented or sublet
3= under redemption agreement
4= occupied in usufruct
5=occupied free of charge, i.e.loaned by friends or relatives or given in exchange for services, such as caretaking, cleaning and so on
z) ratamutuo: mortgage installment
aa) assvita: life insurance policy
ab) pensint: personal retirement plan or supplementary pension fund payments
ac) assdanni: private health and accident insurance payments

4.4 Evidence on the common variable distribution in the two sources

The following tables show the distribution of ***Z*** in SHIW and HBS. Corresponding figures are reported in the Appendix.

As it can be noticed, the recoding of the HH makes the distribution of many of the common variables rather compatible among the two sources.

Household age groups and region of residence (Tab. 4.7, 4.8) 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 types of households (Tab. 4.10) and number of household components (Tab. 4.9): in the former, a dramatic excess of singles (4.6 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 (class 2) and single-parents (class 10) 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. 4.11) displays a significant redistribution between married and single.

Turning to the household head characteristics, a more comparable picture emerges: educational level (Tab. 4.12), as well as occupational status (Tab. 4.13) 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. 4.14) displays a lower share in trade and catering services (4) in SHIW and a positive discrepancy in private services to person (8), 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. 4.15), to the detriment of office workers and school teachers and non employed. This last evidence seems suggesting the definition of HH still slightly under-representing women household-headed families in SHIW after the recodification. More significant discrepancy are instead observed in the distribution of house-related variables (Tab. 4.16, 4.17, 4.18), 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 et al., 2007). In addition, a substantial redistribution between occupiers and home owners is also observed.

Table 4.7: Household head age group distribution SHIW vs HBS

	HBS			SHIW			
eta15_hh	Freq.	Percent	Cum.	Freq.	Percent	Cum.	SHIW-HBS
15-17	2,362	0.01	0.01				
18-24	135,314	0.55	0.56	196,308	0.81	0.81	0.26
25-29	542,576	2.19	2.75	568,674	2.36	3.17	0.17
30-34	1,368,267	5.53	8.28	1,277,428	5.3	8.47	-0.23
35-39	2,201,803	8.9	17.18	1,627,100	6.75	15.22	-2.15
40-44	2,539,969	10.27	27.45	2,973,971	12.34	27.56	2.07
45-49	2,591,974	10.48	37.93	2,448,799	10.16	37.72	-0.32
50-54	2,416,636	9.77	47.7	2,328,081	9.66	47.38	-0.11
55-59	2,249,510	9.1	56.8	1,965,062	8.15	55.53	-0.95
60-64	2,245,216	9.08	65.88	2,400,896	9.96	65.49	0.88
65-69	1,918,863	7.76	73.64	1,978,281	8.21	73.7	0.45
70-74	2,111,430	8.54	82.17	2,297,002	9.53	83.23	0.99
75 and over	4,409,333	17.83	100	4,048,281	16.79	100	-1.04
Total	24,733,253	100		24,109,883	100		

Table 4.8: Distribution by region

HBS	SHIW

ireg_m	Freq.	Percent	Cum.	Freq.	Percent	Cum.	SHIW-HBS
Piemonte e Valle d’Aosta	2,055,188	8.25	8.25	2,350,238	9.75	9.75	1.5
Lombardia	4,256,002	17.09	25.35	3,641,659	15.1	24.85	-1.99
Trentino Alto Adige	426,349	1.71	27.06	1,172,852	4.86	29.71	3.15
Veneto	2,006,927	8.06	35.12	1,564,242	6.49	36.2	-1.57
Friuli Venezia Giulia	554,992	2.23	37.35	486,544	2.02	38.22	-0.21
Liguria	785,134	3.15	40.5	918,491	3.81	42.03	0.66
Emilia Romagna	1,942,278	7.8	48.3	1,543,983	6.4	48.43	-1.4
Toscana	1,601,365	6.43	54.74	1,652,897	6.86	55.29	0.43
Umbria	373,810	1.5	56.24	357,363	1.48	56.77	-0.02
Marche	637,554	2.56	58.8	627,950	2.6	59.37	0.04
Lazio	2,319,587	9.32	68.11	2,155,934	8.94	68.31	-0.38
Abruzzo	537,453	2.16	70.27	384,293	1.59	69.9	-0.57
Molise	128,238	0.52	70.79	264,893	1.1	71	0.58
Campania	2,087,166	8.38	79.17	1,813,538	7.52	78.52	-0.86
Puglia	1,527,210	6.13	85.3	1,474,556	6.12	84.64	-0.01
Basilicata	227,980	0.92	86.22	657,048	2.73	87.37	1.81
Calabria	771,087	3.1	89.32	722,086	2.99	90.36	-0.11
Sicilia	1,979,915	7.95	97.27	1,573,279	6.53	96.89	-1.42
Sardegna	679,938	2.73	100	748,037	3.1	100	0.37
Total	24,898,173	100		24,109,883	100		

Table 4.9: Distribution of number of family members

	HBS			SHIW			
ncomp	Freq.	Percent	Cum.	Freq.	Percent	Cum.	SHIW-HBS
1	7,544,326	30.3	30.3	6,162,832	25.56	25.56	-4.74
2	6,811,328	27.36	57.66	7,422,728	30.79	56.35	3.43
3	5,050,667	20.29	77.94	4,703,779	19.51	75.86	-0.78
4	4,179,627	16.79	94.73	4,341,185	18.01	93.87	1.22
5	1,019,098	4.09	98.82	1,107,651	4.59	98.46	0.5
6	215,474	0.87	99.69	341,325	1.42	99.88	0.55
7	51,725	0.21	99.9	10,770	0.04	99.92	-0.17
8	19,417	0.08	99.97	17,858	0.07	99.99	-0.01
9	3,996	0.02	99.99				
10	2,057	0.01	100				
12	458	0	100	1,755	0.01	0.01	
Total	24,898,173	100		24,109,883	100		

Table 4.10: Distribution by household type

	HBS			SHIW			
TIPFAM	Freq.	Percent	Cum.	Freq.	Percent	Cum.	SHIW-HBS
Lone person with aged 35 or less	790,892	3.18	3.18	667,293	2.77	2.77	-0.41
Lone person with aged 35-64	3,006,973	12.08	15.25	2,242,159	9.3	12.07	-2.78
Lone person with aged 65 or more	3,746,461	15.05	30.3	3,253,380	13.49	25.56	-1.56

Couple without children with reference person aged 35 or less	359,307	1.44	31.74	412,720	1.71	27.27	0.27
Couple without children with reference person aged 35-64	1,960,631	7.87	39.62	1,864,970	7.74	35.01	-0.13
Couple without children with reference person aged 65 or more	2,725,666	10.95	50.57	3,445,928	14.29	49.3	3.34
Couple with 1 child	4,135,496	16.61	67.18	4,371,016	18.13	67.43	1.52
Couple with two children	3,804,988	15.28	82.46	3,966,771	16.45	83.88	1.17
Couple with three of more children	917,111	3.68	86.14	1,136,521	4.71	88.59	1.03
Single-parent	2,036,744	8.18	94.32	1,113,027	4.62	93.21	-3.56
Other typologies	1,413,904	5.68	100	1,636,098	6.79	100	1.11
Total	24,898,173	100		24,109,883	100		

Table 4.11: HH marital status distribution

HBS				SHIW			
staciv_hh	Freq.	Percent	Cum.	Freq.	Percent	Cum.	SHIW-HBS
Married	14,279,565	58.22	58.22	14,930,469	61.93	61.93	3.71
Single	4,178,597	17.04	75.26	3,395,081	14.08	76.01	-2.96
Separated/divorced; widower/widow.	6,066,691	24.74	100	5,784,333	24	100	-0.74
Total	24,524,853	100		24,109,883	100		

Table 4.12: Distribution of educational level of HH

HBS				SHIW			
studio_hh	Freq.	Percent	Cum.	Freq.	Percent	Cum.	SHIW-HBS
None	1,074,613	4.32	4.32	991,096	4.11	4.11	-0.21
Elementary school	5,640,965	22.66	26.97	5,248,367	21.77	25.88	-0.89
Middle school	8,681,545	34.87	61.84	8,858,791	36.74	62.62	1.87
High school	6,548,539	26.3	88.14	6,216,280	25.78	88.4	-0.52
Bachelor’s degree	2,678,544	10.76	98.9	2,504,990	10.39	98.79	-0.37
Post-graduate qualification.	273,967	1.1	100	290,359	1.2	100	0.1
Total	24,898,173	100		24,109,883	100		

Table 4.13: Distribution of HH by occupational status

HBS				SHIW			
Condprof_hh	Freq.	Percent	Cum.	Freq.	Percent	Cum.	SHIW-HBS
Employed	12,966,270	52.08	52.08	12,872,133	53.39	53.39	1.31
First-job seeker	79,439	0.32	52.4	67,600	0.28	53.67	-0.04
Homemaker or pensioner	10,559,066	42.41	94.81	10,294,199	42.69	96.36	0.28
Unemployed	677,936	2.72	2.72	720,365	2.99	2.99	0.27
Student	102,174	0.41	3.13	109,245	0.45	3.44	0.04
Other not employed (including well-off)	513,288	2.06	5.19	46,341	0.19	3.63	-1.87
Total	24,898,173	100		24,109,883	100		

Table 4.14: Distribution of HH by branch of activity

HBS	SHIW

sett_hh	Freq.	Percent	Cum.	Freq.	Percent	Cum.	SHIW-HBS
Agriculture	596,344	2.4	2.4	560,262	2.32	2.32	-0.08
Manufacturing	2,359,172	9.48	11.87	2,500,211	10.37	12.69	0.89
Building and construction	1,493,663	6	17.87	1,241,473	5.15	17.84	-0.85
Wholesale and retail trade, lodging and catering services	2,668,976	10.72	28.59	2,080,548	8.63	26.47	-2.09
Transport and communication	866,295	3.48	32.07	643,280	2.67	29.14	-0.81
Services of credit and insurance institutions	362,685	1.46	33.53	478,183	1.98	31.12	0.52
Real estate and renting services, other professional and business activities	1,100,328	4.42	37.94	714,934	2.97	34.09	-1.45
Domestic services and other private services to persons	1,355,895	5.45	43.39	1,816,695	7.54	41.63	2.09
General government, defence, education, health and other public services	2,789,956	11.21	54.6	2,762,635	11.46	53.09	0.25
Extra-territorial	50,892	0.2	54.8	27,176	0.11	53.2	-0.09
Not employed	11,253,967	45.2	100	11,284,486	46.8	100	1.6
Total	24,898,173	100		24,109,883	100		

Table 4.15: Distribution of HH by work status

	HBS			SHIW			
qualp_hh	Freq.	Percent	Cum.	Freq.	Percent	Cum.	SHIW-HBS
Blue-collar worker	4,378,757	17.59	17.59	4,779,580	19.82	19.82	2.23
Office worker or school teacher	4,338,265	17.42	35.01	3,923,688	16.27	36.1	-1.15
Junior manager/cadre	836,143	3.36	38.37	655,247	2.72	38.82	-0.64
Manager, senior official	618,234	2.48	40.85	405,089	1.68	40.5	-0.8
Member of the arts or professions	658,744	2.65	43.5	721,894	2.99	43.49	0.34
Sole proprietor	559,555	2.25	45.75	313,676	1.3	44.79	-0.95
Freelance	1,503,192	6.04	51.78	1,453,843	6.03	50.82	-0.01
Owner or member of a family business	41,616	0.17	51.95	402,076	1.67	52.49	1.5
Active shareholder/partner	31,764	0.13	52.08	217,040	0.9	53.39	0.77
Not employed.	11,931,903	47.92	100	11,237,750	46.61	100	-1.31
Total	24,898,173	100		24,109,883	100		

Table 4.16: Distribution by resident status

	HBS			SHIW			
godab1	Freq.	Percent	Cum.	Freq.	Percent	Cum.	SHIW-HBS
Home owner or with the right of redemption	18,320,473	73.63	73.63	16,591,747	68.82	68.82	-4.81
Tenant with right of usufruct	4,286,842	17.23	90.85	4,981,635	20.66	89.48	3.43
Use without charge	2,275,750	9.15	100	2,536,501	10.52	100	1.37
Total	24,883,065	100		24,109,883	100		

Table 4.17: Second dwellings

	HBS			SHIW			SHIW-HBS
nimm2f	Freq.	Percent	Cum.	Freq.	Percent	Cum.	
0	22,869,803	91.85	91.85	20,503,379	85.04	85.04	-6.81
1	1,829,075	7.35	99.2	2,822,783	11.71	96.75	4.36
2	166,400	0.67	99.87	548,273	2.27	99.02	1.6
3	27,442	0.11	99.98	152,366	0.63	99.65	0.52
4	5453	0.02	100	58,869	0.24	99.89	0.22
5				13,678	0.06	99.95	0.06
6				3,539	0.01	99.96	0.01
8				5,254	0.02	99.98	0.02
9				1,742	0.01	100	0.01
Total	24,898,173	100		24,109,883	100		

Table 4.18: Location of the main dwelling

	HBS			SHIW			
ubic1_1	Freq.	Percent	Cum.	Freq.	Percent	Cum.	SHIW-HBS
City	19,869,097	80.41	80.41	20,835,798	86.42	86.42	6.01
Small town, village	3,312,659	13.41	93.82	1,616,818	6.71	93.13	-6.7
Hamlet, detached houses, farm area	1,526,449	6.18	100	1,657,267	6.87	100	0.69
Total	24,708,205	100		24,109,883	100		

Below we report a series of figures comparing histogram distributions of consumption variables belonging to the two sources. In particular, as mentioned above, while HBS contains a detailed list of expenditures, SHIW provides a rougher disaggregation of overall consumption in some macro-variables: food consumption (CONSAL, Fig. 4.2), other non durable consumption (CONDIV, Fig. 4.3), their sum (Fig. 4.4), and transport expenditure (CDUR1, Fig. 4.5), and total durable consumption (CDUR, Fig. 4.6)[12]. As mentioned, this comparison required an accurate recodification of HBS variables (much more numerous and detailed) in order to get comparable aggregates among the two sources.

Generally speaking, SHIW distributions are much less smooth compared to HBS due to the lower accuracy of recording in the field of consumption. In fact, many modal values are thickened in correspondence of round amounts. Turning to specific distribution, while food and especially other non durables (Fig. 4.2, 4.3) as well as their sum (Fig. 4.4) fit rather well both in terms of means and the higher statistical moments, durable consumption is much more noisy and displays some inconsistencies. In particular, while the transport expenditure after adjustment appears fairly similar at least in terms of mean and variance (especially relative to the pre-adjustment situation,

[12] The variable containing the other non durable, CDUR2, which, together with CDUR1, sums up to CDUR, is not reported as it can be thought as the difference of the two.

Fig. 4.5), total durable expenditure (Fig. 4.6) is substantially different, recommend the other durable consumption being definitely dissimilar between the two samples. This evidence would suggest relying on total non durable consumption and, among the durables, on transport related expenses as variables for the match.

Finally, the distribution of the actual rents paid displays a definitely high degree of comparability, being a variable less subject to under-reporting or mis-reporting issues.

Figure 4.2: Food consumption distribution in the two surveys

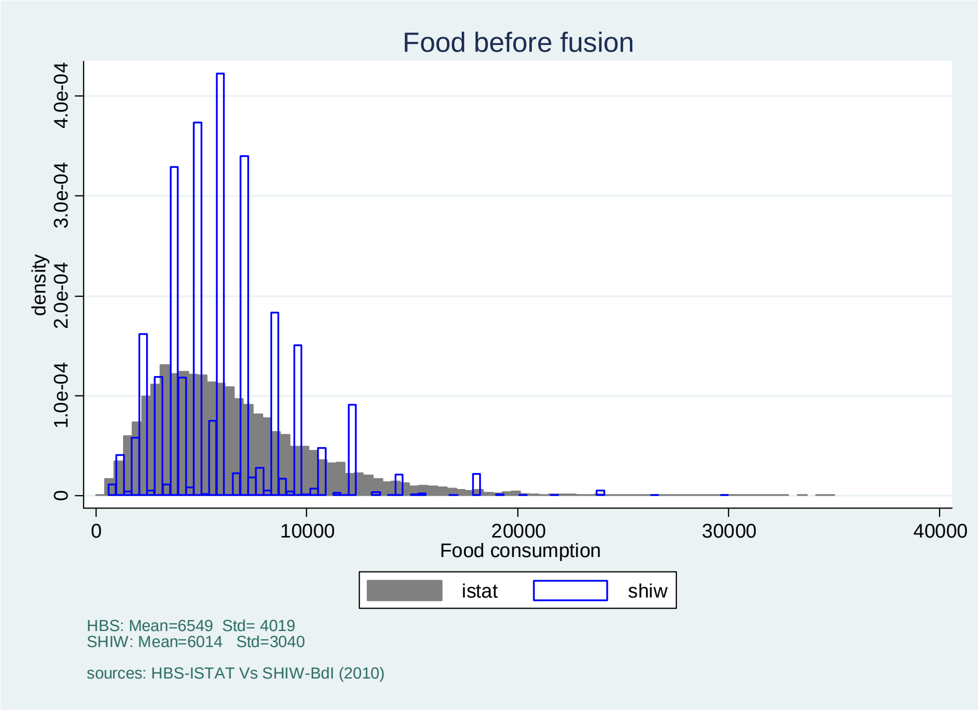


Figure 4.3: Non food non durable consumption distribution in the two surveys

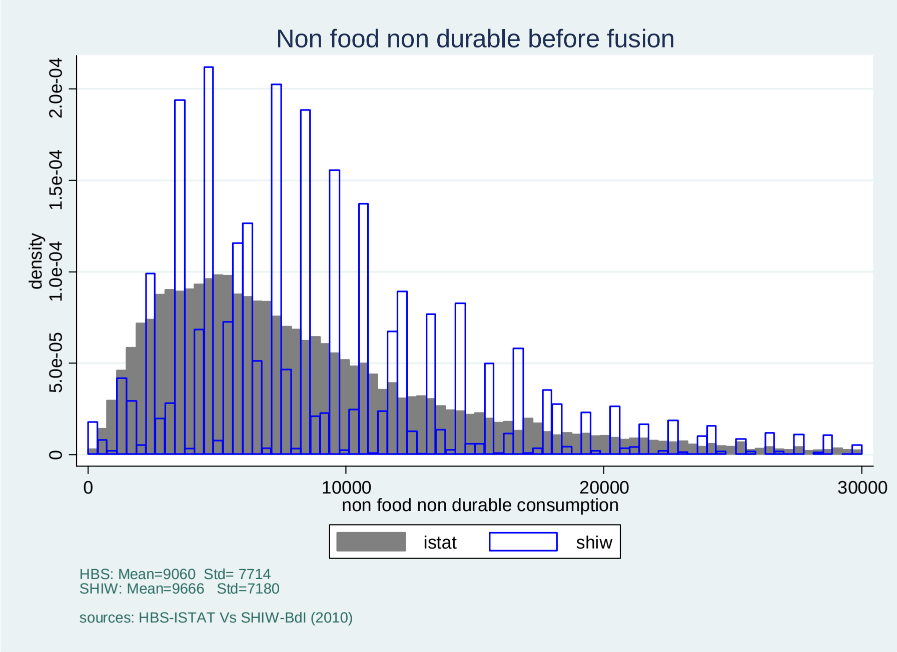
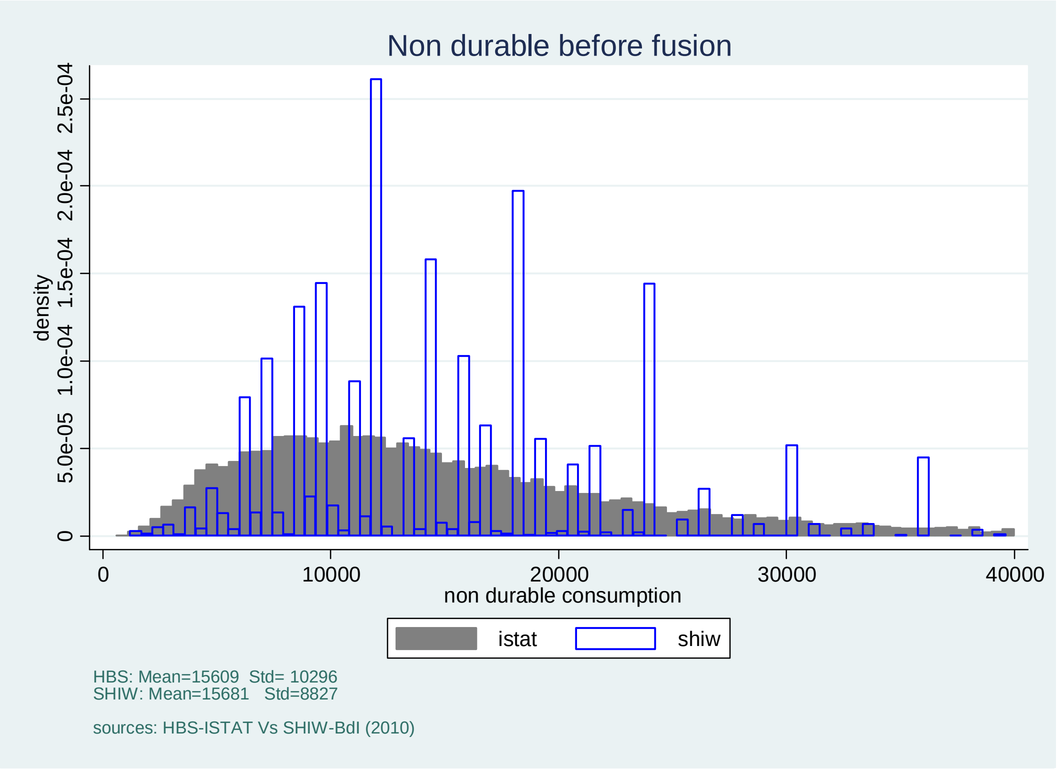


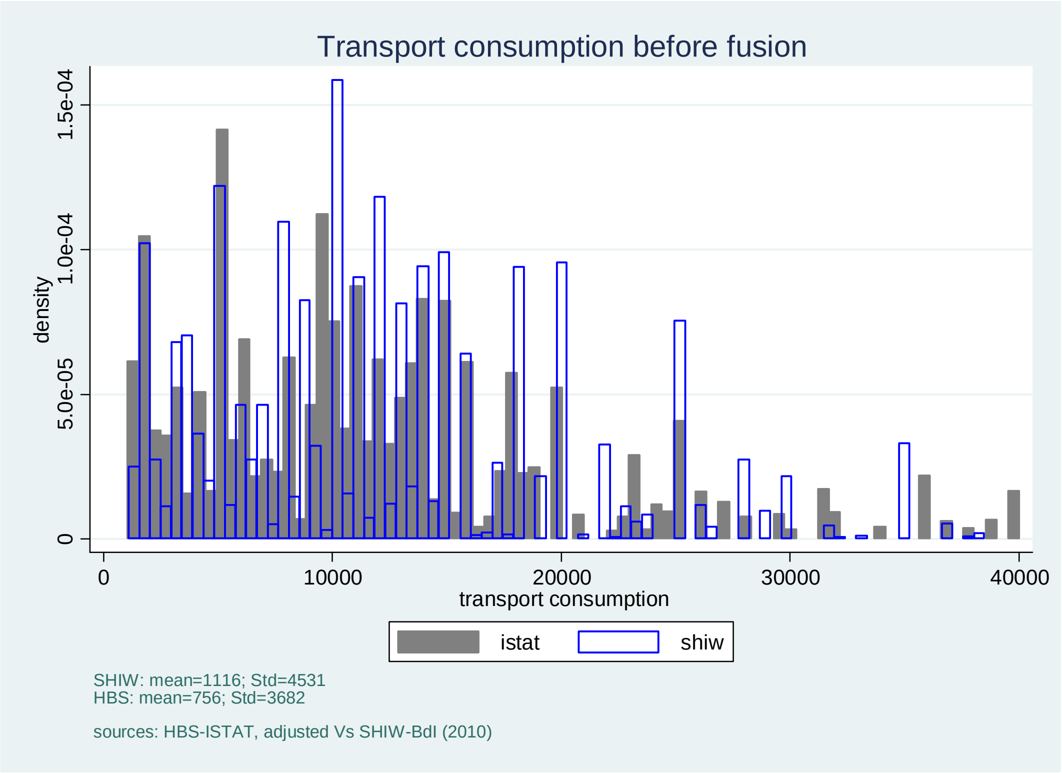
Figure 4.4: Total non durable consumption distribution in the two surveys



HBS: Mean=15609 Std= 10296
SHIW: Mean=15681 Std=8827

sources: HBS-ISTAT Vs SHIW-BdI (2010)

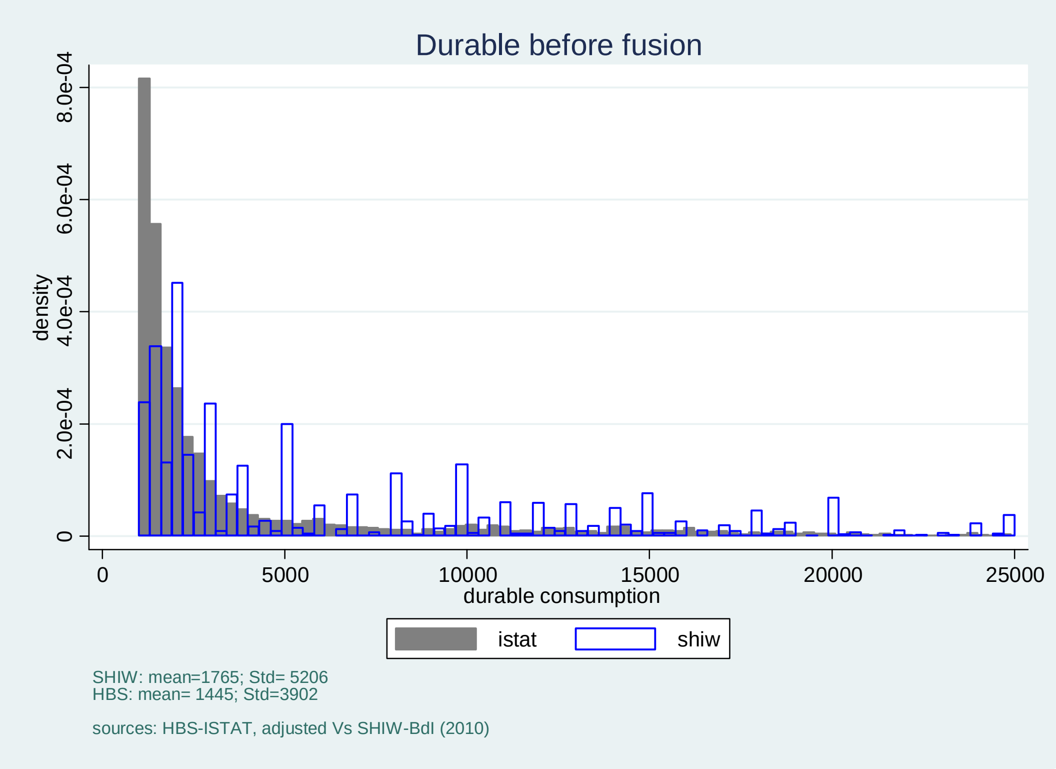
Figure 4.5: Transport durable consumption distribution in the two surveys, after HBS Monte Carlo adjustment



SHIW: mean=1116; Std=4531
HBS: mean=756; Std=3682

sources: HBS-ISTAT, adjusted Vs SHIW-BdI (2010)

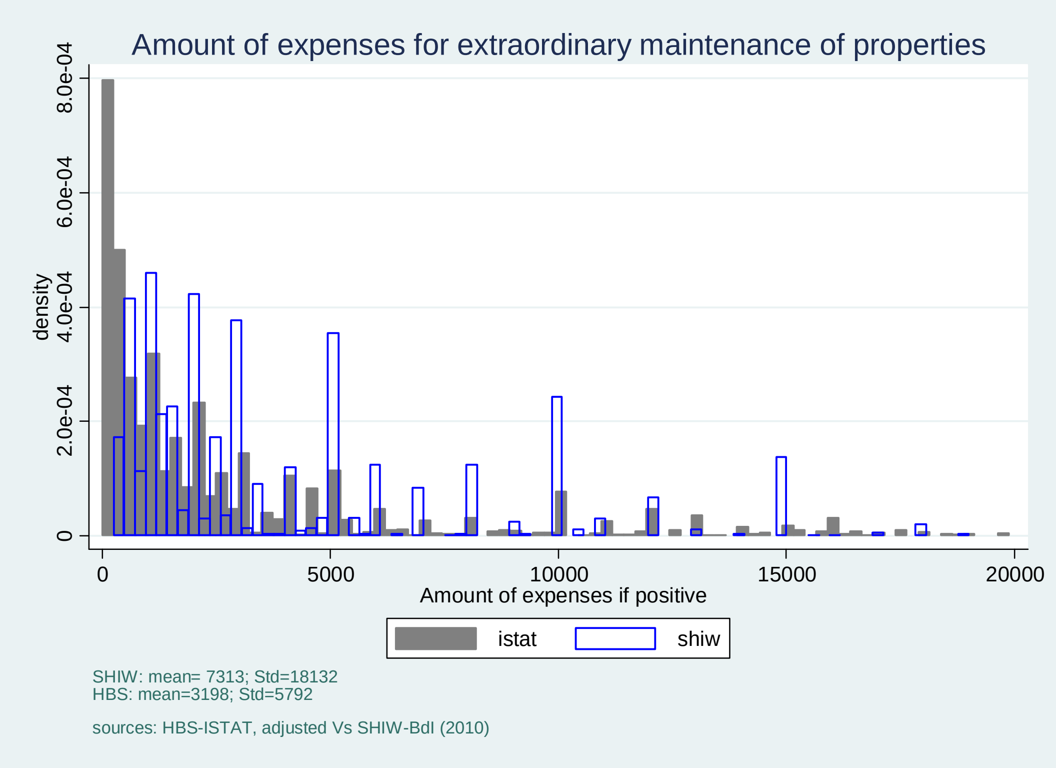
Figure 4.6: Total durable consumption distribution in the two surveys, after HBS Monte Carlo adjustment



SHIW: mean=1765; Std= 5206
HBS: mean= 1445; Std=3902

sources: HBS-ISTAT, adjusted Vs SHIW-BdI (2010)

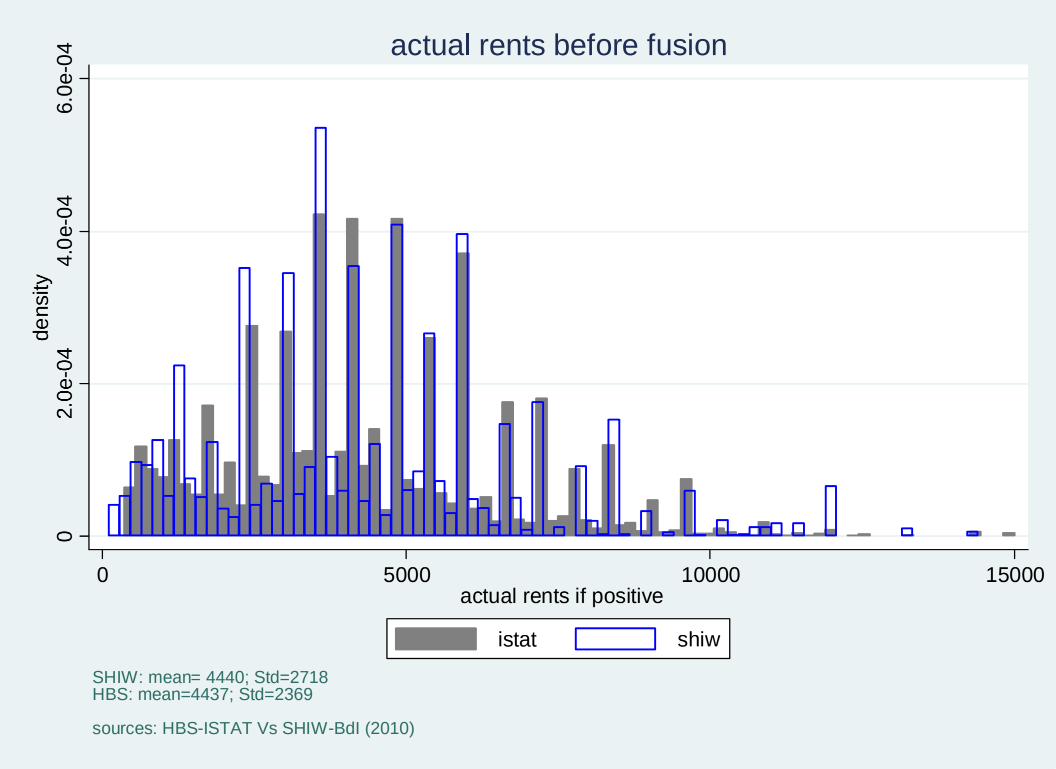
Figure 4.7: Amount of expenses for extraordinary maintenance of properties (Mastrip) in the two surveys, after HBS Monte Carlo adjustment



SHIW: mean= 7313; Std=18132
HBS: mean=3198; Std=5792

sources: HBS-ISTAT, adjusted Vs SHIW-BdI (2010)

Figure 4.8: Distribution of actual rents paid in the two surveys



SHIW: mean= 4440; Std=2718
HBS: mean=4437; Std=2369

sources: HBS-ISTAT Vs SHIW-BdI (2010)

5. Matching results and goodness tests

This section offers a selection of results according to the propensity score matching procedure with the *Mahalanobis* metric and the maximum manageable stratification[13] (i.e. 100 cells or strata), in some cases comparing the results with the ones obtained with different assumptions, such as lower number of cells and/or the use of a different metrics.

As shown in Table 5.1, the 100 cells stratification provides a lower feasible level of replication of donor units compared to a matching with fewer strata, thus accounting for a higher variation in the data. *Mahalanobis* metrics is preferred to nearest neighbour method since it seems to perform better in reproducing both conditional variability of target variables and the joint distribution and easily allows the match of all recipient units.

Table 5.1: Number of replicated HBS donor observations

Mahalanobis50cells			Mahalanobis80cells			Mahalanobis100cells		
Copies	observations	surplus	Copies	Observations	surplus	Copies	observations	Surplus
1	3824	0	1	3864	0	1	40330	0
2	2216	1108	2	2260	1130	2	22161108	1108

[13] Of course, the more one increases the number of strata, the greatest the deterministic part of the procedure; however, the sample size of each stratum gets smaller, up to some cells become too thin to be matched.

3	993	662	3	942	628	3	936624	624
4	516	387	4	452	339	4	384288	288
5	200	160	5	220	176	5	195156	156
6	102	85	6	156	130	6	9680	80
7	42	36	7	35	30	7	4236	36
8	8	7	8	16	14	8	87	7
9	18	16	9	18	16	10	3027	27
10	10	9	14	14	13	11	1110	10
11	11	10						
16	16	15						
17	17	16						

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 that first level validity can be assessed is by means of a simulation study (Rässler, 2002). Thus, we are not able 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.

In sum, we are able to test the validity of our data fusion at the fourth level; however, making some assumptions on the ***(*** $\boldsymbol{C^a}$***, I, W)*** distribution observed in SHIW, we can also make tests at the second and third level.

5.1 Lower level: the marginal distributions.

In order to check this level of validity we compare the unconditional distribution of several aggregates of consumption between the original HBS and the fused resulting file. We also show the comparison between a couple of conditional distributions, such as household consumption conditional on household types and household head age groups.

For the sake of our analyses, we recognize two definitions of total household consumption. The first, we refer to as "total matching consumption" (TMC), is one of the two dimensions on which we build the strata. 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 and imputed rents[14]. It represents, on average, about 70 percent of total household consumption

[14] These latter aggregates have a significant incidence on household income but their distributions are rather different between the two surveys, as they are related to durable goods and private wealth, which appear to be better represented

expenditures and - after the adjustment/imputation procedures illustred in the previous section - reveals a good degree of comparability between the two original surveys.

Then, we build the broadest definition of yearly household consumption, which collects all the observable flows of expenditure. We will refer to it as household "overall consumption" (OC).

The unconditional distribution of TMC in the fused file fits very closely the original HBS distribution up to about the 99th percentile (about 60,000 Euros). In the upper tail, it progressively deteriorates yet still appears still acceptable up to 100,000 Euros; a poor comparability between the two distributions clearly emerges in the very top tail (0.001%). This is due to a problem of common support. In fact, as shown in Table A.1 in the Appendix, the original HBS sample presents a smoother distribution all over the support, and even in the top tail. Such smoothness tends to vanish in the matching process due to the SHIW sample bumps.

In order to achieve a more continuos distribution and avoid bumps or multi-modality - at least in most part of the distribution - due to the lower accuracy of consumption recording in SHIW, we decided to slightly shock at random[15] the original SHIW sub-aggregates' values.

Table 5.2 shows a set of standard inequality indexes for this variable. The comparison displays a good preservation of the TMC inequality in the fused file. For instance, looking at the Gini, the difference between the two distribution is 0.002

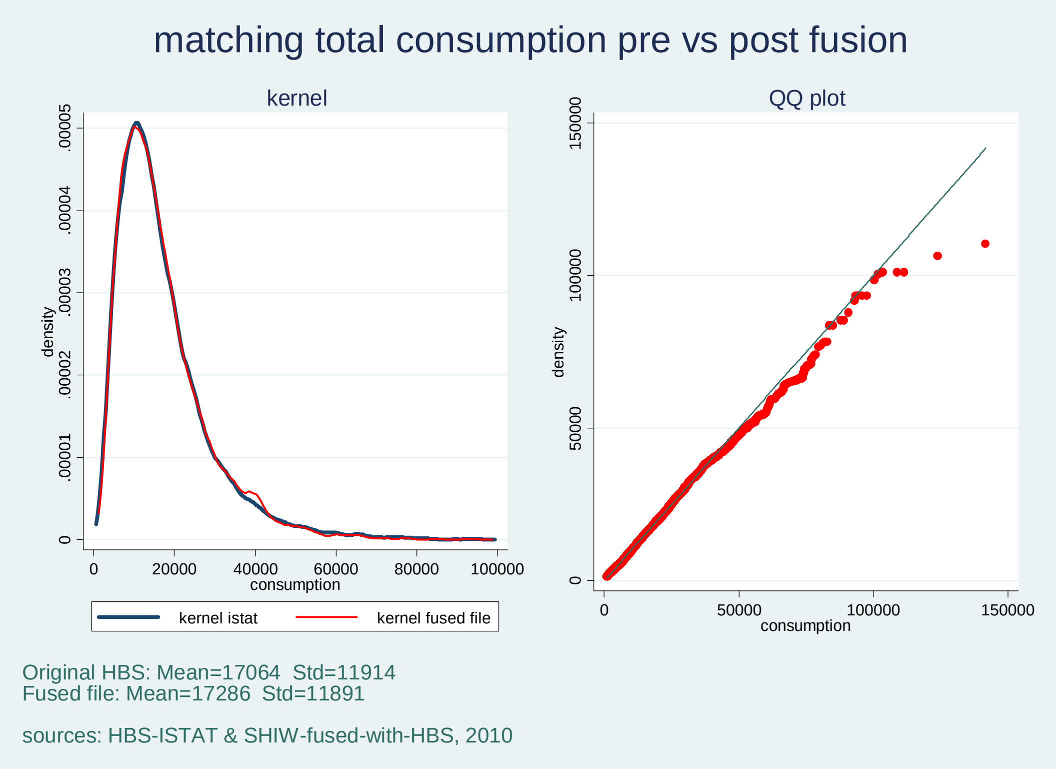
A slightly higher difference emerges with the General Entropy index with extreme parameters (-1 or 2). This evidence confirms the above-mentioned common support issues at the extreme tails of the distribution.

However, testing these differences between original HBS and the fused file in a framework of bootstrap inference (with 250 replications), it can be noticed that TMC mean, Gini and General Entropy in the two surveys are not statistically different from each other at least at the 5% level (Tab. 5.3).

in SHIW. Therefore, we decided not to use it for directly conditioning the matching. The amount of expenses for extraordinary maintenance of all property owned by the household has been kept out from the strata variables but it is used to condition the logistic propensity score estimates.

[15] 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 5.1: Unconditional distributions of TMC



Original HBS: Mean=17064 Std=11914
Fused file: Mean=17286 Std=11891

sources: HBS-ISTAT & SHIW-fused-with-HBS, 2010

Table 5.2: 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	
5.417	2.226	0,28542	2.373		5.492	2.284	0.416	2.370	
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(-1)	GE(-1)	GE(-1)	**Gini**	GE(-1)	GE(-1)	GE(-1)	GE(-1)	**Gini**
0.27433	0.20972	0.20322	0.24374	0.34884	0.25593	0.20336	0.19902	0.23660	0.34674

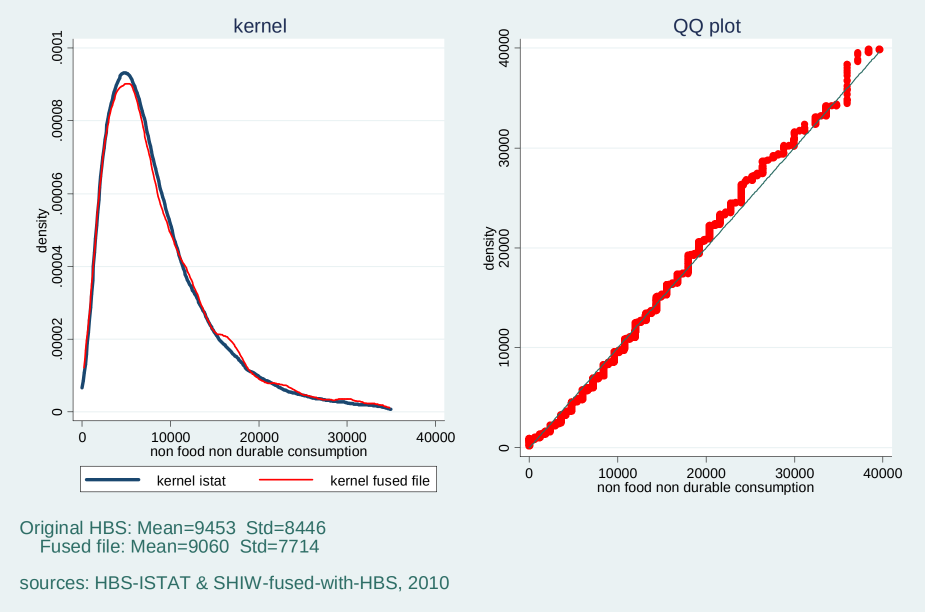
Table 5.3: Bootstrap inference on main distributive statistics for TMC

	Donor HBS					
	Number of obs = 22246					
	Replications = 250					
	Observed	Bootstrap			Normal-based	
	Coef.	Std. Err.	Z	P>z	[95% Conf. Interval]	
mean	17064.16	1.024.738	166.52.00	0.000	16863.32	17265.01
gini	.3488356	0.002321	150.03.00	0.000	.3442866	.3533846
ge0	.2097204	0.00281	74.62	0.000	.2042121	.2152286
ge1	.2032224	0.003123	65.08.00	0.000	.1971023	.2093425
ge2	.2437353	0.005585	43.65	0.000	.2327899	.2546807
	Fused file					
	Number of obs = 7951					
	Replications = 250					
	Observed	Bootstrap			Normal-based	

	Coef.	Std. Err.	Z	P>z	[95% Conf. Interval]	
mean	17286.4	1.973.998	87.57.00	0.000	16899.51	17673.3
gini	.3467364	0.004378	79.02.00	0.000	.338156	.3553168
ge0	.2033622	0.005093	39.93	0.000	.1933808	.2133436
ge1	.1990216	0.005761	34.55.00	0.000	.1877307	.2103124
ge2	2365978	0.009923	1,0166667	0.000	.2171486	.256047

Breaking down by sub-aggregates of consumption expenditure (Figures 5.2, 5.3 and 5.4), an acceptable preservation for non-durable consumption emerges, with slight differences in the first and second moments and with the usual deterioration in the top tail. For the actual rent paid (Fig. 5.5), the before-fusion dissimilarity account for a resulting distribution that is a mixture between HBS and SHIW, with first and second moments lower compared to the latter.

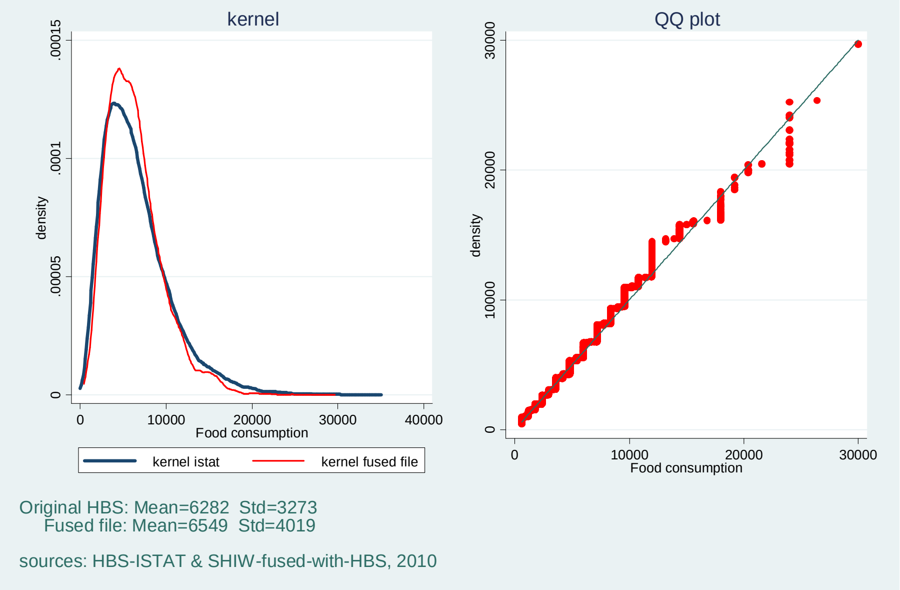
Figure 5.2: Non food, non durable consumption distribution in HBS and fused file



Original HBS: Mean=9453 Std=8446
Fused file: Mean=9060 Std=7714

sources: HBS-ISTAT & SHIW-fused-with-HBS, 2010

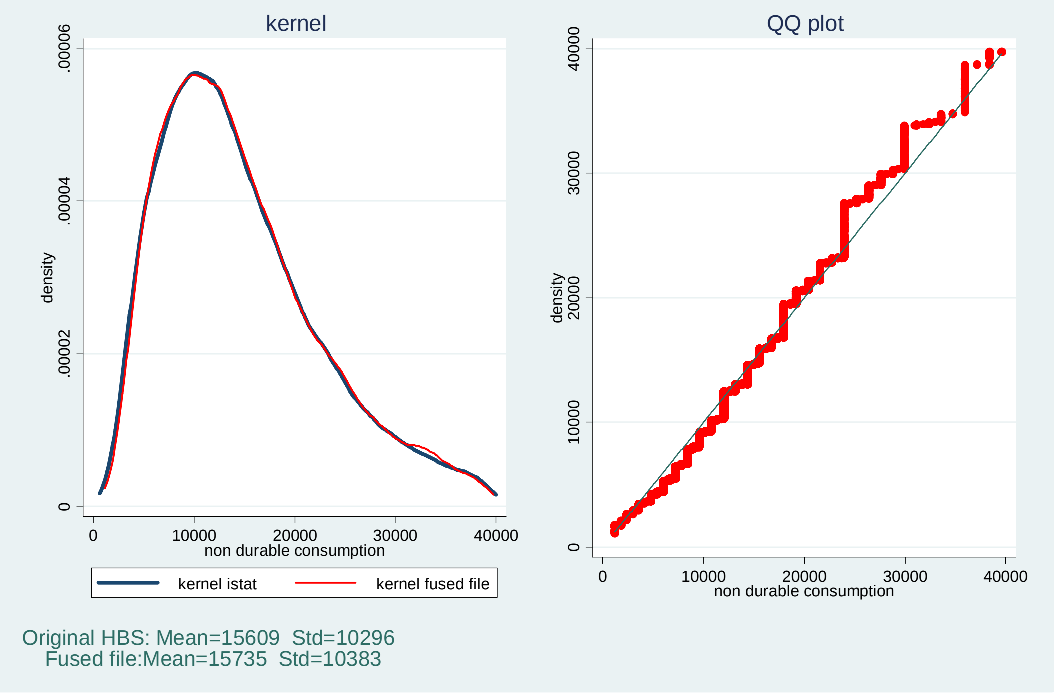
Figure 5.3: Food consumption distribution in HBS and fused file



Original HBS: Mean=6282 Std=3273
Fused file: Mean=6549 Std=4019

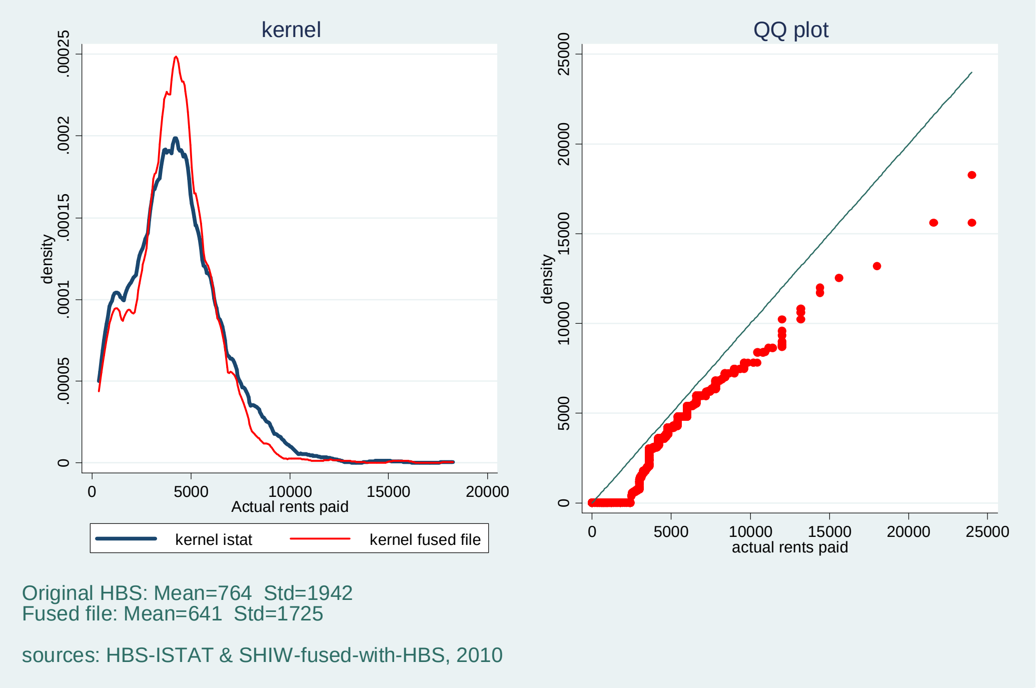
sources: HBS-ISTAT & SHIW-fused-with-HBS, 2010

Figure 5.4: Total non durable consumption distribution in HBS and fused file



Original HBS: Mean=15609 Std=10296
Fused file:Mean=15735 Std=10383

Figure 5.5: actual rents paid distribution in HBS and fused file



Original HBS: Mean=764 Std=1942
Fused file: Mean=641 Std=1725

sources: HBS-ISTAT & SHIW-fused-with-HBS, 2010

For durable consumption, included the expenses for extraordinary maintainance of household properties, as discussed in previous sections, our benchmark for comparing the resulting syntethic distribution is the SHIW file. In this case, despite the adjustments made before the matching, the post-matching distributions still present significant differences - summarized by a lower mean and variance in the syntethic distribution (Figures 5.6 and 5.7). This issue raises concerns on how to use the additional information on these items provided by HBS for the sake of inference and microsimulation.

As a last result on the unconditional distributions, we consider the overall aggregate of household consumption expenditure, included the items not directly conditioning the matching process. In this case, the structural differences - despite necessary adjustments and imputations to make the two surveys more comparable -, made the overall synthetic representation a mixture of the two original samples, both in mean and in distribution. (Figures 5.9 and 5.10; Table A.4).

In particular, the resulting distribution, on the one hand, and the original HBS and SHIW, on the other side, remain roughly comparable, though the issues of lacking of common support exacerbates, especially in the very top tail. In particular, the comparability worsen above 100,000 Euros with respect to both original files. In terms of overall inequality (Tab. 5.4), the resulting Gini for OC is about 0.31, one basis points above the original SHIW file and 2 above the original HBS. The significant dissimilarity in the inter-decile ratio 90/10 (4.3 vs 4 and 3.8 compared tho HBS and SHIW, respectively) signals that the main alterations pertain the tails.

Finally, the bootstrap inference (Tab. 5.5) reveals that both the resulting Gini and the overall mean are not statistically different from the SHIW figure at 5 percent level.

Figure 5.6: Total durable consumption distribution in SHIW and fused file

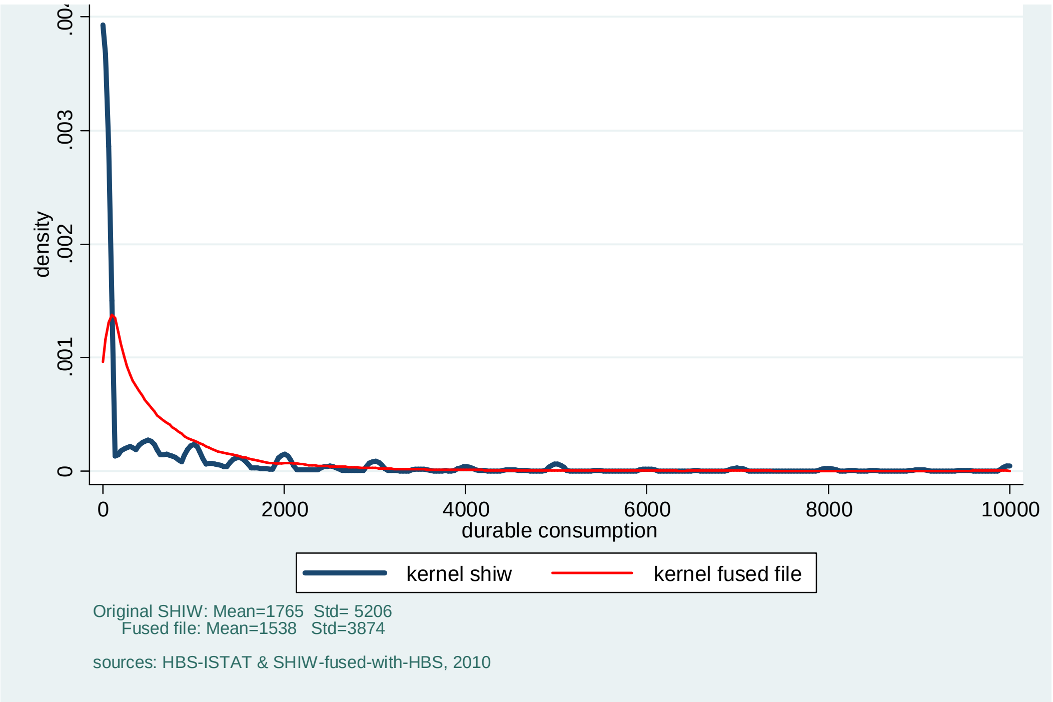
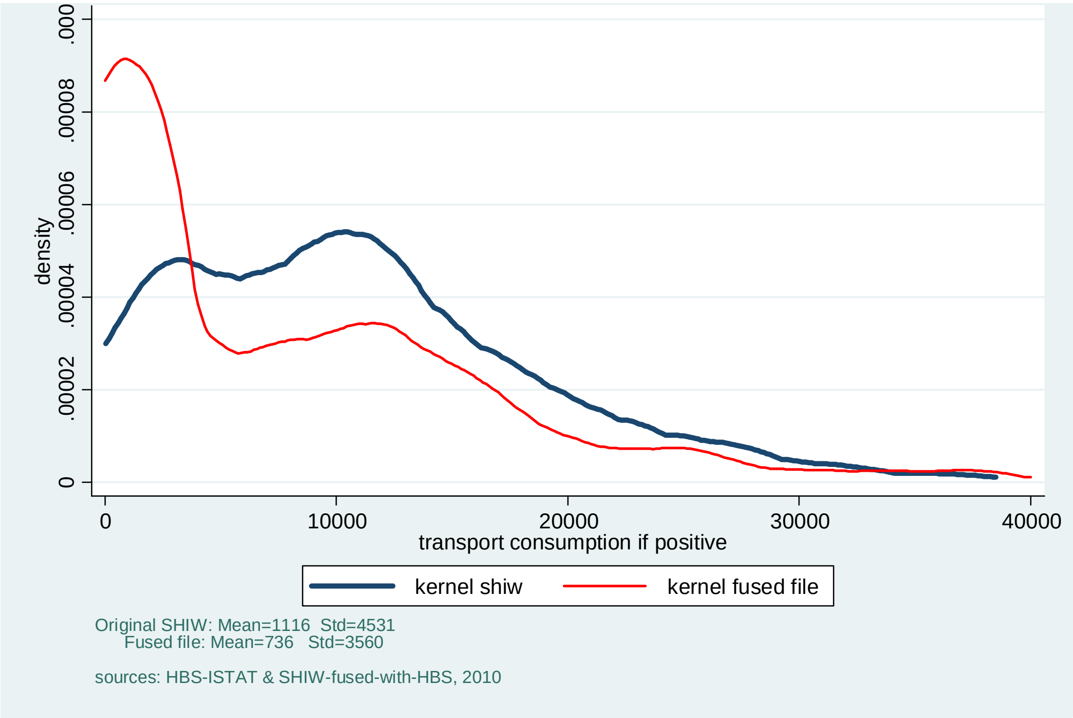


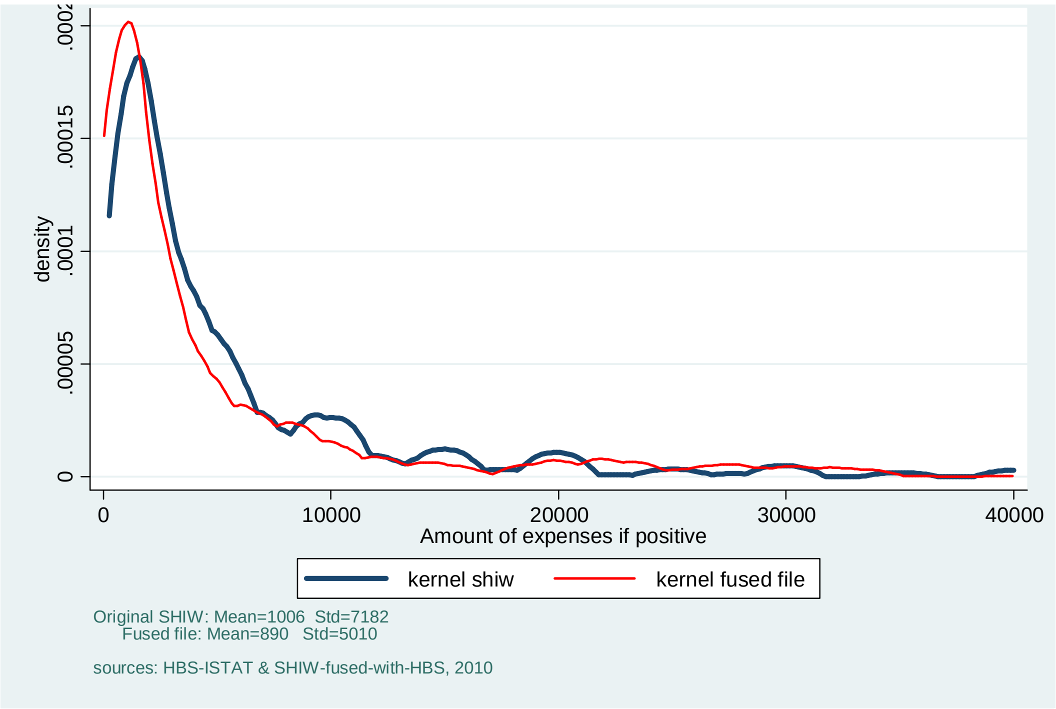
Figure 5.7: Transport expenditure distribution in SHIW and fused file



Original SHIW: Mean=1116 Std=4531
Fused file: Mean=736 Std=3560

sources: HBS-ISTAT & SHIW-fused-with-HBS, 2010

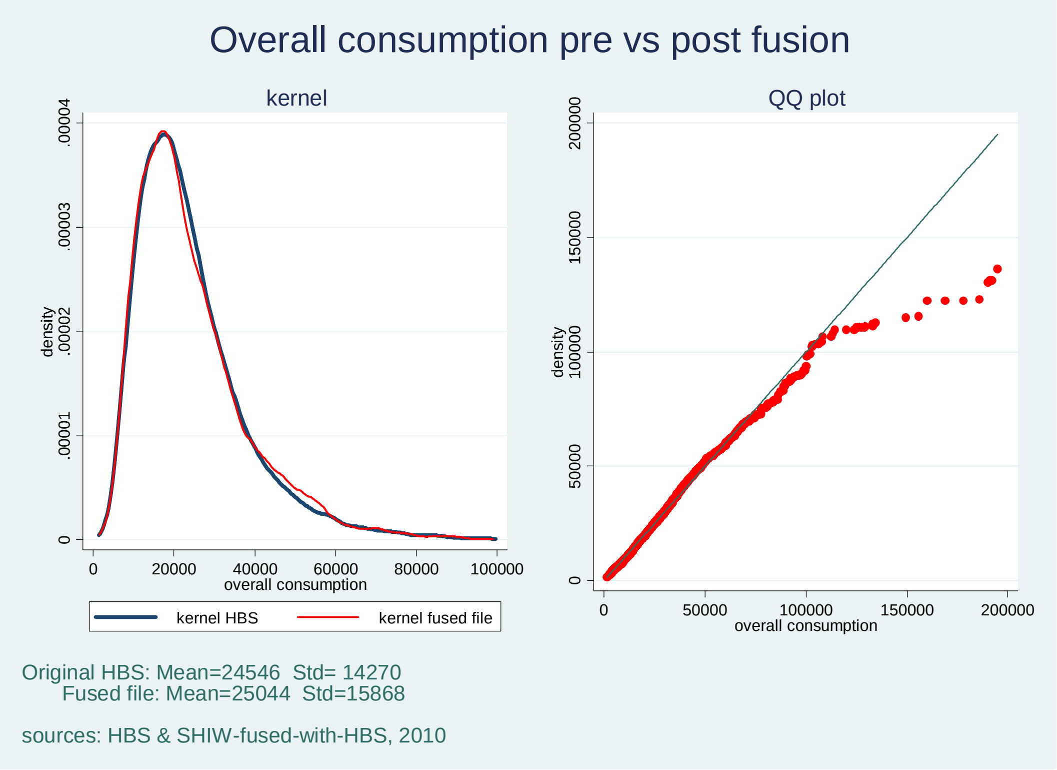
Figure 5.8: Extraordinary maintenance of properties expenditure distribution in SHIW and fused file



Original SHIW: Mean=1006 Std=7182
Fused file: Mean=890 Std=5010

sources: HBS-ISTAT & SHIW-fused-with-HBS, 2010

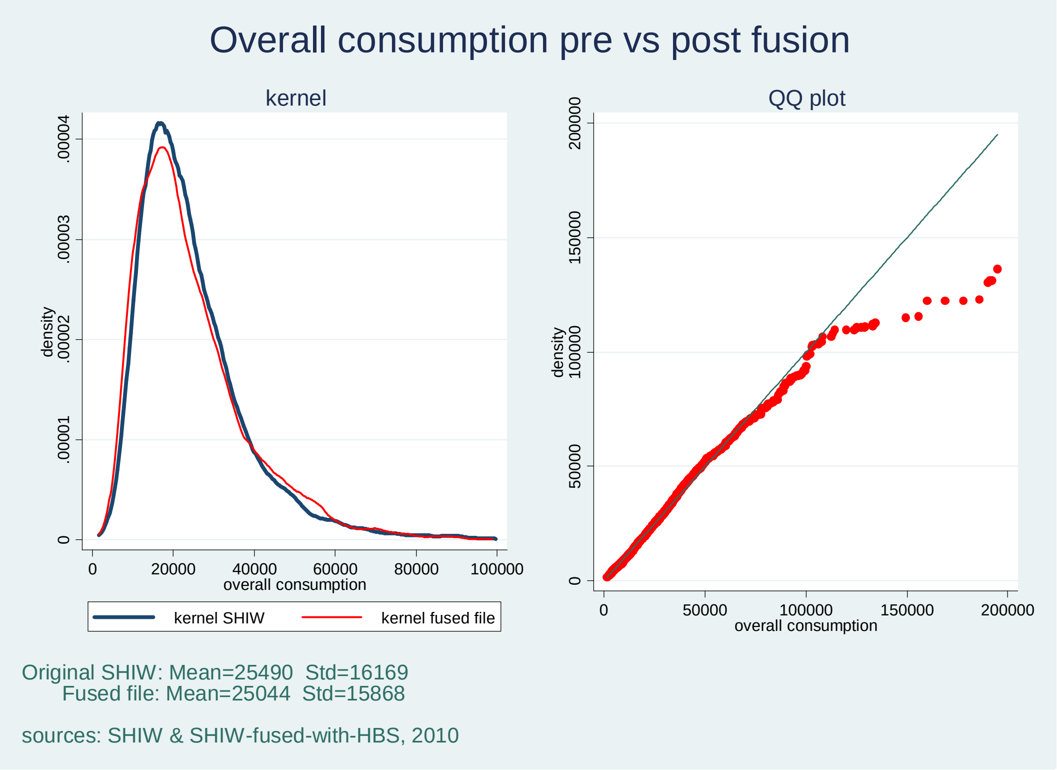
Figure 5.9: Overall consumption distribution in HBS and fused file


Overall consumption pre vs post fusion

Original HBS: Mean=24546 Std= 14270
Fused file: Mean=25044 Std=15868

sources: HBS & SHIW-fused-with-HBS, 2010

Figure 5.10: Overall consumption distribution in SHIW and fused file


Overall consumption pre vs post fusion

Original SHIW: Mean=25490 Std=16169
Fused file: Mean=25044 Std=15868

sources: SHIW & SHIW-fused-with-HBS, 2010

Table 5.4: Inequality indexes for overall consumption

Donor, original HBS

Percentile ratios				
p90/p10	p90/p50	p10/p50	p75/p25	
4.025	1.969	0.489	2.057	
Generalized Entropy indices GE(a) and Gini coefficient				
GE(-1)	GE(0)	GE(1)	GE(2)	Gini
0.17427	0.14723	0.14602	0.16899	0.29711

Recipient, original SHIW

Percentile ratios				
p90/p10	p90/p50	p10/p50	p75/p25	
3.814	1.995	0.523	1.974	
Generalized Entropy indices GE(a) and Gini coefficient				
GE(-1)	GE(0)	GE(1)	GE(2)	Gini
0.17178	0.15056	0.15797	0.2012	0.30189

Fused file

Percentile ratios				
p90/p10	p90/p50	p10/p50	p75/p25	
4.301	2.11	0.491	2.138	
Generalized Entropy indices GE(a) and Gini coefficient				
GE(-1)	GE(0)	GE(1)	GE(2)	Gini
0.19043	0.16315	0.16603	0.20073	0.31457

Table 5.5: Bootstrap inference on mean and Gini

Donor, original HBS

Bootstrap results — Number of obs = 22246; Replications = 250

	Observed Coef.	Bootstrap Std. Err.	z	P>z	Normal-based [95% Conf.	Interval]
mean	24546.01	118.3114	207.47	0	24314.13	24777.9
gini	0.2971128	0.0020186	147.19	0	.2931563	.3010692

Recipent, Original SHIW

Bootstrap results — Number of obs = 7951; Replications = 250

	Observed Coef.	Bootstrap Std. Err.	z	P>z	Normal-based [95% Conf.	Interval]
mean	25490.42	265.5709	95.98	0	24969.91	26010.93
gini	0.3018923	0.0039644	76.15	0	.2941222	.3096624

Fused file

Bootstrap results — Number of obs = 7951; Replications = 250

Observed	Bootstrap	Normal-based

	Coef.	Std. Err.	z	P>z	[95% Conf. Interval]	
mean	25044.91	268.8201	93.17	0	24518.03	25571.78
Gini	0.3145658	0.004479	70.23	0	.3057871	.3233445

Lastly, conditioning on some common observables, Tables A.2 and A.3 in the Appendix show the tabulation of total matching consumption by ten household typologies and twelve age groups.

In particular, looking at the bootstrap inference reported in the following (Tables A.7 and A.8) we can observe that only two sub-groups in the former condional distribution report mean and standard deviation which are statistically different from each other at 5 % level (i.e. single person over 65 and single-headed household with children), while conditioning on household head's age group, only the means for the subgroup 60-64 result to be statistically different from each other[16].

5.2 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 ***(X, Y, 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, $\widetilde{cov}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = cov(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})$. 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, since we are willing to assume the ***(C, I, W)*** distribution observed in SHIW as the 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 ***(C, I, W, (Z))*** by estimating a consumption equation[17] where the dependent variable is the fused consumption, and the explanatory variables (income, wealth variables and socio-demographic characteristics) from the recipient file. Then, we compare its coefficients with those of an estimated consumption equation in which the dependent variable is the original SHIW one. Following, we compute t-test and χ^2-test (such as Hausman test) as well as the absolute sum of estimated coefficient differences (ASEC) and the absolute mean of predicted values differences (AMPV), so as to obtain some metrics for evaluating different matching hypotheses. In practice, we evaluate the level of preservation of the joint distribution in the data fusion in terms of departure from the

[16] Of course, these exercises can be replicated conditioning different X variables on different Z common characteristics. For space reasons, we do not report all of them but results are available upon request.

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

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. In other terms, we look for the algorithm that minimize:

$$\tilde{f}(\boldsymbol{C}|\boldsymbol{I},\boldsymbol{W},\boldsymbol{Z}) - f(\boldsymbol{C}|\boldsymbol{I},\boldsymbol{W},\boldsymbol{Z})$$

$$\text{s.t. } N_{\text{fused}}=7951 \qquad (1)$$

Table 5.6, displays the comparison of the vector of estimated coefficients between the fused and original SHIW file, in a Hausman test framework resulting from a PSM procedure with the *Mahalanobis* metric and 100 cells stratification. As expected, the χ^2-test refuses the null that - jointly - the difference in coefficients is not systematic. Nevertheless, in terms of constrained minimization of ASEC and AMPV, this solution proves to superior to other trials we performed (see following Tab. 5.7), i.e. it minimize eq. (1)[18].

When we carry out the same test by using the OC as dependent variable for the consumption equation (Tab. 5.8), as expected, the similarity between $\tilde{f}(\boldsymbol{C}|\boldsymbol{I},\boldsymbol{W},\boldsymbol{Z})\ and\ f(\boldsymbol{C}|\boldsymbol{I},\boldsymbol{W},\boldsymbol{Z})$ worsen (both in terms of ASEC and AMPV) holding, however, a quite acceptable overall comparability. Table 5.9 suggests that, in terms of OC, PSM procedure with the *Mahalanobis* metric and 50 cells stratification seems slightly superior.

Such dissimilarity between TMC and OC has to be kept in mind in carrying out joint distributional analyses of consumption, income and wealth or micro simulation exercises of direct and indirect taxation.

Table 5.6: Hausman test on matching consumption functions in order to analyze the degree of preservation of main joint distributions passing from SHIW to the synthetic dataset

---- OLS Coefficients ----				
Dependent variable: ln{total matching consumption}				
Explanatory	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	fused	shiw	Difference	S.E.
ln{disposable income}	0.17597	0.1786894	-0.0027192	0.0000696
ln{real wealth}	0.011891	0.0104377	0.0014533	0.000022
ln{finacial wealth}	0.024256	0.022538	0.0017178	0.0000125
ln{financial debt}	0.009701	0.0102116	-0.0005102	0.0000103
ln{actual rent}	0.064578	0.0624558	0.0021218	0.0001232
ln{imputed rent}	0.072514	0.0640824	0.0084318	0.0000887
number of earners	0.122582	0.1089565	0.013625	0.000072
Age	0.00935	0.0066437	0.0027064	0.000019
age^2	-7.8E-05	-0.0000516	-0.0000261	1.64E-07
hh woman	-0.04234	-0.0494204	0.0070755	0.0001131
number of components	0.082058	0.0726669	0.0093914	0.000048
Marital status (omitted: Never married)				

[18] In fact, nearest neighbor distance with 50 cells does not allow the full match of all recipient units, even increasing indefinitely the caliper.

Married	-0.15303	-0.1136726	-0.0393603	0.0001506
Separeted, divorced	-0.13178	-0.1068528	-0.0249241	0.0001725
Widowed	-0.13479	-0.085807	-0.0489848	0.0001594
		Education (omitted: None)		
Primary	0.02437	0.0092761	0.0150936	0.0002179
lower-secondary	0.136597	0.0857644	0.0508323	0.0002332
upper-secondary	0.21275	0.1563315	0.0564184	0.0002458
Tertiary	0.281142	0.2246243	0.0565178	0.0002749
Postgraduate	0.212612	0.2057733	0.0068382	0.0004468
		Occupational status (omitted: in work)		
First-time job seeker	-0.09787	-0.0338735	-0.0639917	0.0011804
Housewife	-0.04926	-0.0403084	-0.0089542	0.0009658
Rentier	-0.00429	0.0169664	-0.0212565	0.0009067
Pensioner	-0.06931	-0.019354	-0.0499581	0.0009405
Unemployed	0.089339	0.035908	0.0534305	0.0010975
		Branch of activity (omitted:		
sett_2	-0.13811	-0.065163	-0.0729438	0.000288
sett_3	0.02262	0.0401009	-0.0174814	0.0001681
sett_4	-0.01862	0.0070418	-0.0256605	0.0002096
sett_5	-0.01251	0.0239975	-0.0365052	0.00019
sett_6	-0.03513	-7.43E-06	-0.0351177	0.0002603
sett_7	0.198664	0.1736905	0.0249739	0.0003015
		Properties		
dummy second swelligs	0.014416	0.0172619	-0.0028462	0.0001119
tenant	0.173947	0.1257934	0.0481538	0.0011959
with usufruct, use without charge	0.015915	-0.0009745	0.0168893	0.0001637
		Work status (omitted: blue-collar, freelance)		
office worker or school teacher	0.107315	0.087317	0.0199976	0.0002009
junior manager/cadre	0.175957	0.1294837	0.046473	0.0003115
manager, senior official self-employed	0.286131	0.3398319	-0.0537014	0.0003749
member of the arts or professions	0.242956	0.2136857	0.0292707	0.000307
sole proprietor	0.133515	0.1387231	-0.0052081	0.0003921
not employed	-0.00212	-0.0133101	0.0111942	0.0009312

Test: Ho: difference in coefficients not systematic

chi2(38) = (b-B)'[(V_b-V_B)^(-1)](b-B)

= 99770.01

Prob>chi2 = 0.0000

Table 5.7: Statistics on $\boldsymbol{f}(\text{TMC}|\boldsymbol{I},\boldsymbol{W},\boldsymbol{Z})$

	Mahalanobis 100 cells	Mahalanobis 50 cells	Nearest neighbour Caliper 50 cells
ASEC	0.992756	1.2102103	1.514979
AMPV	.0640927	.066517	.054987
N_{fused}	7951	7951	7861

Table 5.8: Hausman test on overall consumption functions in order to analyze the degree of preservation of main joint distributions passing from SHIW to the synthetic dataset

---- OLS Coefficients ----

dependent variable: ln{total consumption}

	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	fused	SHIW	Difference	S.E.
ln{disposable income}	0.1377799	0.158901	-0.02112	9.16E-06
ln{real wealth}	0.0104817	0.01022	0.000262	4.34E-06
ln{finacial wealth}	0.0199477	0.014852	0.005096	6.81E-07
ln{financial debt}	0.0084265	0.006067	0.002359	1.31E-06
ln{actual rent}	0.1224974	0.253765	-0.13127	5.38E-06
ln{imputed rent}	0.2000234	0.331718	-0.13169	9.11E-06
number of earners	0.0971476	0.061022	0.036126	6.27E-06
Age	0.0074373	0.005628	0.001809	2.02E-06
age^2	-0.0000581	-4E-05	-1.8E-05	1.11E-08
hh woman	-0.028637	-0.03394	0.005306	1.22E-05
number of components	0.0552746	0.054936	0.000338	3.69E-06
Marital status (omitted: Never married)				
Married	-0.1204895	-0.06025	-0.06024	1.74E-05
Separeted, divorced	-0.1087173	-0.06335	-0.04537	1.54E-05
Widowed	-0.1060093	-0.05836	-0.04765	7.82E-06
Education (omitted: None)				
Primary	0.072233	-0.01063	0.082861	0.000202
lower-secondary	0.183961	0.04056	0.143401	0.000216
upper-secondary	0.263791	0.083434	0.180356	0.000227
Tertiary	0.314167	0.155867	0.1583	0.000254
Postgraduate	0.191785	0.134191	0.057595	0.000413
Occupational status (omitted: in work)				
Housewife	0.0698654	-0.00799	0.077858	7.62E-06
Rentier	0.1820454	0.041154	0.140891	1.28E-05
Pensioner	0.2563974	0.091934	0.164464	7.09E-06
Unemployed	0.3050362	0.159456	0.14558	8.24E-06
Branch of activity (omitted:				
sett_2	-0.1817705	0.02337	-0.20514	2.61E-05
sett_3	-0.0069989	0.008094	-0.01509	2.96E-05
sett_4	-0.0390315	-0.02256	-0.01647	1.44E-05
sett_5	-0.0194485	-0.01595	-0.0035	5.48E-05
sett_6	-0.0524772	-0.00677	-0.0457	1.77E-05
sett_7	0.1518272	0.107524	0.044303	0.000014
Properties				
dummy second swelligs	0.0208821	0.152978	-0.1321	1.35E-05
tenant	0.7178538	0.755004	-0.03715	0.000088
with usufruct, use without charge	0.0143257	0.022958	-0.00863	2.45E-05
Work status (omitted: blue-collar, freelance)				
office worker or school teacher	0.0738156	0.041214	0.032602	8.88E-05
junior manager/cadre	0.1375652	0.077656	0.059909	9.58E-05

manager, senior official self-employed	0.2733969	0.171276	0.102121	0.000101
member of the arts or professions	0.1679273	0.112264	0.055664	0.000101
sole proprietor	0.1299994	0.074367	0.055632	0.000105
not employed	-0.1754578	0.032802	-0.20826	0.00071

Test: Ho: difference in coefficients not systematic

chi2(37) = (b-B)'[(V_b-V_B)^(-1)](b-B)= 8134658.82

Prob>chi2 = 0.0000

Table 5.9 Statistics on $f(\mathrm{OC}|\boldsymbol{I},\boldsymbol{W},\boldsymbol{Z})$

	Mahalanobis 100 cells	Mahalanobis 50 cells	Nearest neighbour Caliper 50 cells
ASEC	3.1903846	3.114997	1.514979
AMPV	.0842616	.0836663	.1278861
N_{fused}	7951	7951	7861

In terms of overall propensity to consume, the averages are very close to each other (about .89, Table 5.10). This value is quite in line also with other aggregates estimates (according to ISTAT, average savings propensity for the household sector in 2010 was 12.1%, indicating a decreasing trend for savings in Italy) and it is pretty unusual considering the traditional gap between macro and micro figures.

While the standard deviation is slightly higher in the fused file than in SHIW (.96 vs .90). Conditioning on household head's age group, Table A.5 in the Appendix shows that the greater differences are in groups 45-49 and over-74 while, conditioning on household typology, Figure A.6 signals the main differences among single persons over 65, couples without children with HH aged 35-64 and couples with three or more children.

Finally, figures 5.11 and 5.12 below show that our matching procedure results in a shape for the average propensity to consume which is more declining both in disposable income and net wealth (corresponding to a more concave consumption function) compared to the original SHIW picture. The investigation of these differences and the assessment of which is closer to the true (unknown) shape of such distributions in the population will deserve further analysis.

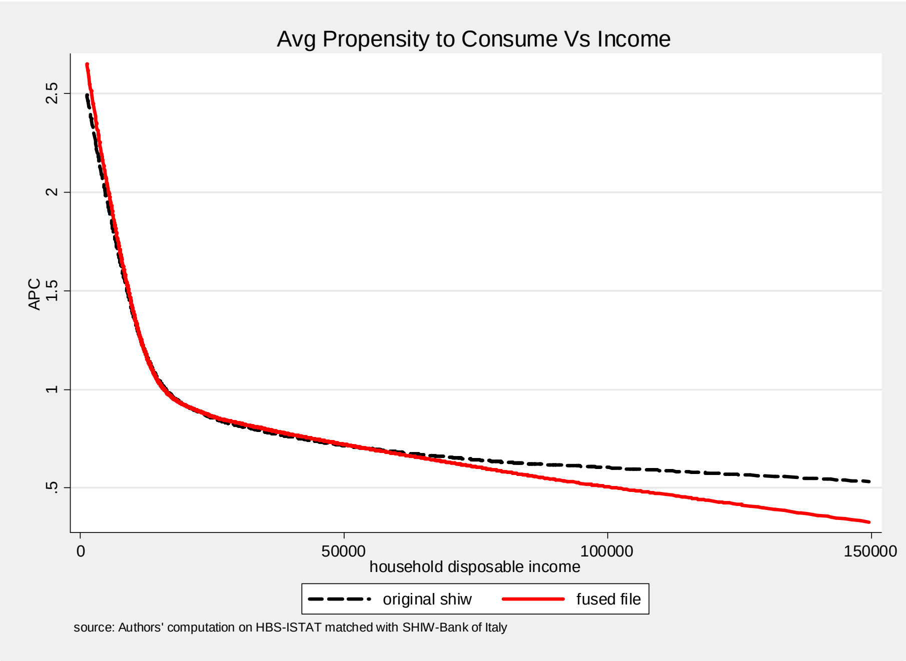
Table 5.10: Overall propensity to consume

Overall propensity to consume, original SHIW file				
	Percentiles	Smallest		
1%	0.3218299	-22.88571		
5%	0.4636188	0.04035		
10%	0.5353982	0.0904605	Obs	24053610
25%	0.6556474	0.1128187	Sum of Wgt.	24053610

50%	0.805501		***Mean***	0.8900568
		Largest	***Std. Dev***.	0.9060953
75%	0.9563276	18		
90%	1.201835	19.2	Variance	0.8210087
95%	1.45749	19.5	Skewness	3.765559

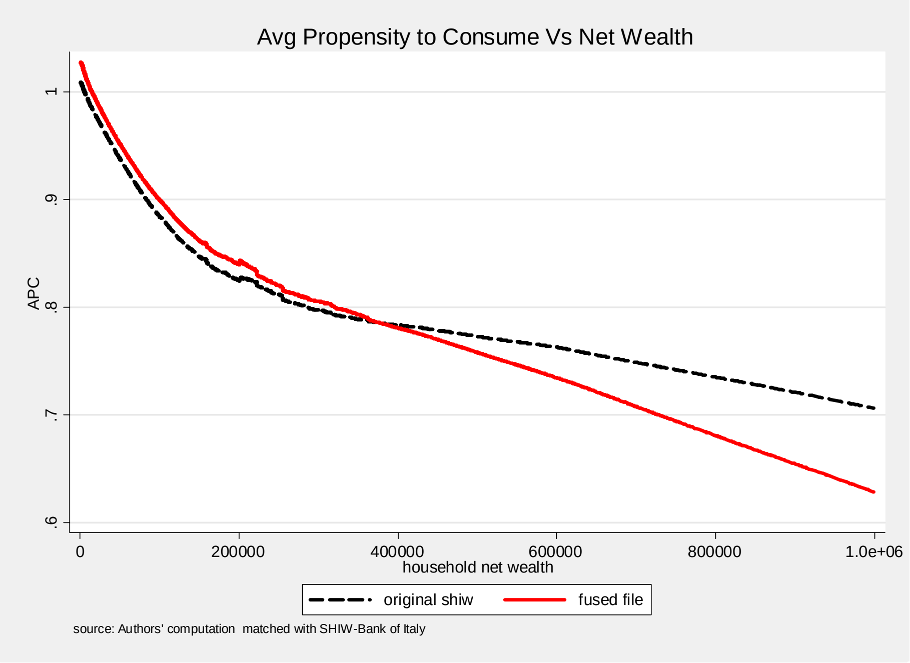
Overall propensity to consume, Fused file				
	Percentiles	Smallest		
1%	0.2469272	-25.90047		
5%	0.3886475	0.0626652		
10%	0.46623	0.1030742	Obs	24053610
25%	0.5983735	0.1195862	Sum of Wgt.	24053610
50%	0.7802094		***Mean***	0.8957129
		Largest	***Std. Dev***.	0.968635
75%	1.004484	16.47606		
90%	1.323618	17.18488	Variance	0.9382537
95%	1.652966	21.45784	Skewness	- 0.2861873

Figure 5.11 Average overall propensity to consume by household disposable income



source: Authors' computation on HBS-ISTAT matched with SHIW-Bank of Italy

Figure 5.12 Average overall propensity to consume by household net wealth



source: Authors' computation matched with SHIW-Bank of Italy

References:

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, Banca d'Italia, Roma.

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.

D'Agostino RB Jr. (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. e 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.

Dehejia, R.H. and S. Wahba (1999). "Causal Effects in Nonexperimental Studies: Reevaluation of the Evaluation of Training Programs", Journal of the American Statistical Association 94, 1053-1062.

Dehejia, R.H. and S. Wahba (2002). "Propensity Score Matching Methods for NonExperimental Causal Studies", Review of Economics and Statistics, 84(1), 151- 161.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd (1998) "Characterizing Selection Bias Using Experimental Data," Econometrica, 66(5), 1017-
1098.

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, Volume 33 (2004), Number 1&2, 153-171

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

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

Rosenbaum, P.R., and D.B. Rubin (1984) “Reducing Bias in Observational Studies using Subclassification 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., and Rowe, G. (1993), “Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption”, *Survey Methodology*, 19, 59-79.

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

Appendix

Figure A.1: Household head age group distribution SHIW vs HBS

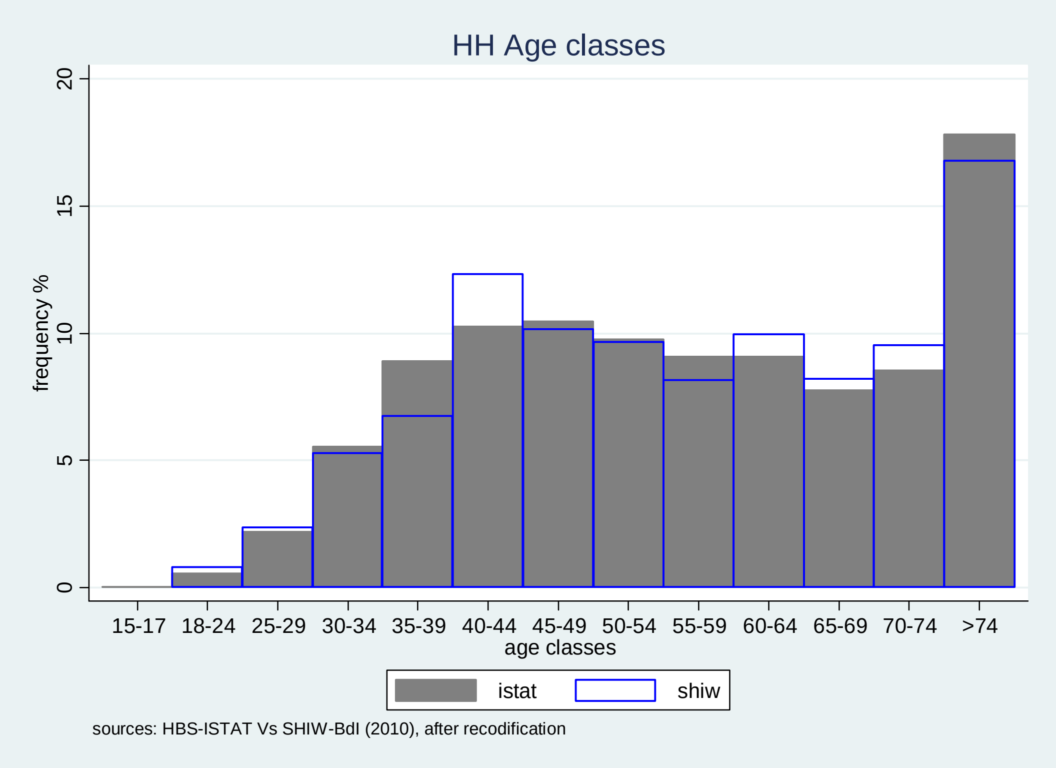


Figure A.2: Household typology distribution SHIW vs HBS

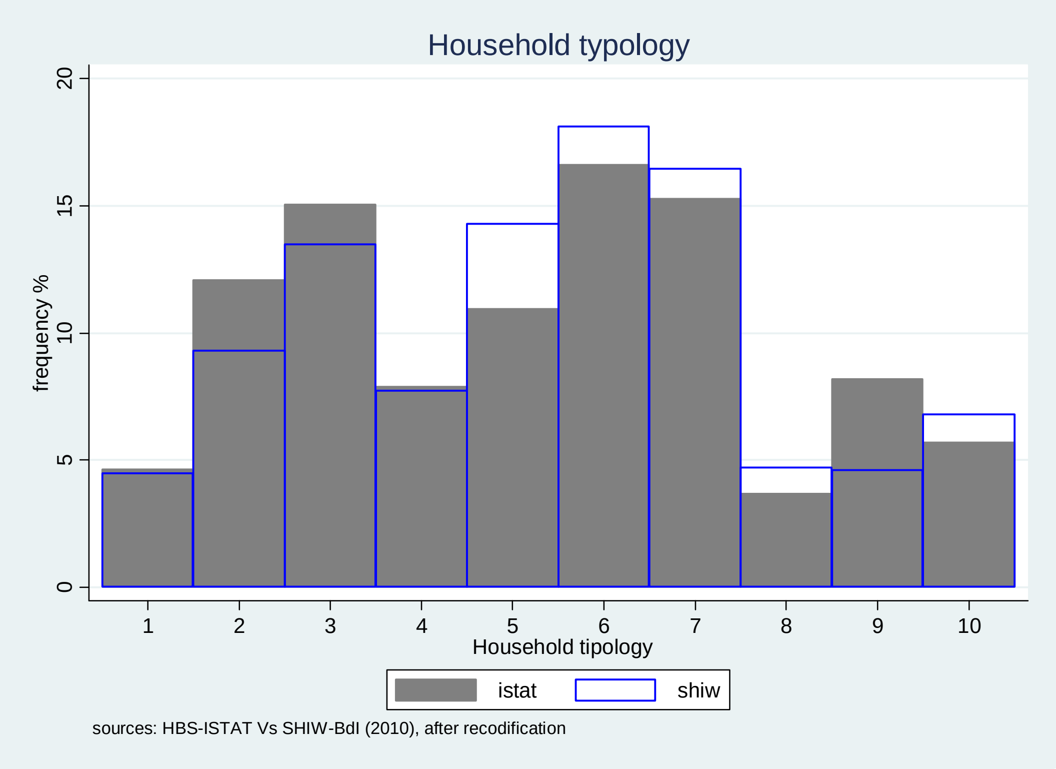


Figure A.3: Region of residence distribution SHIW vs HBS

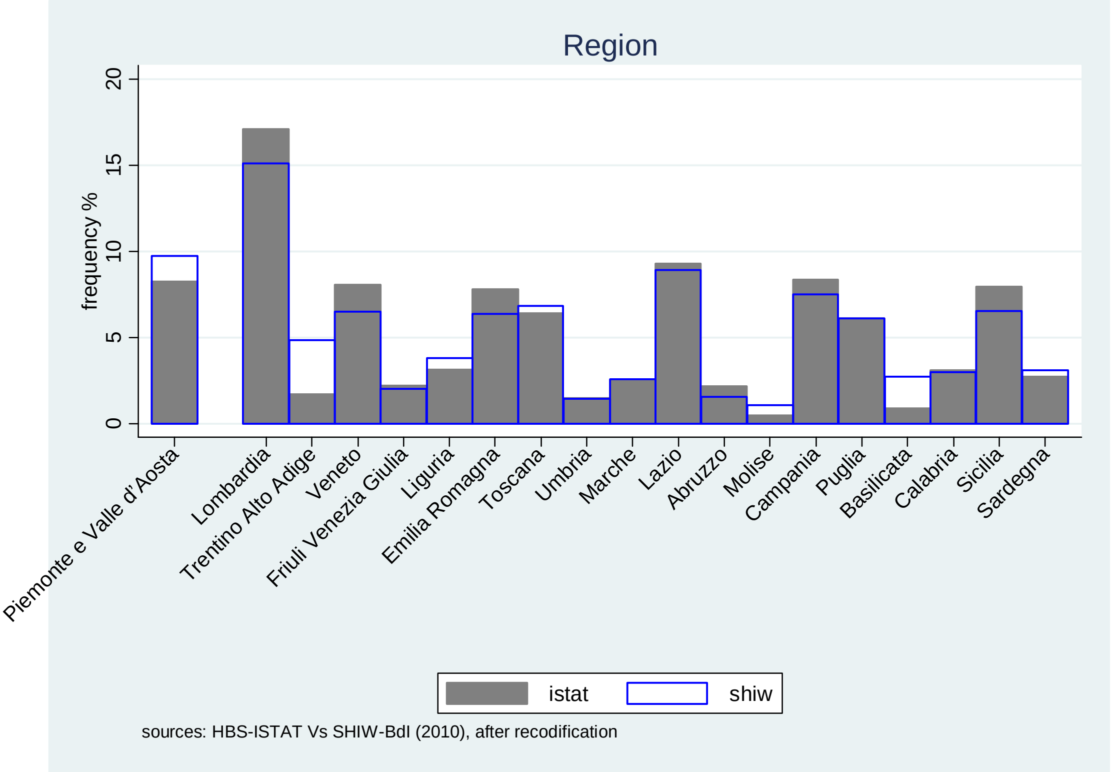
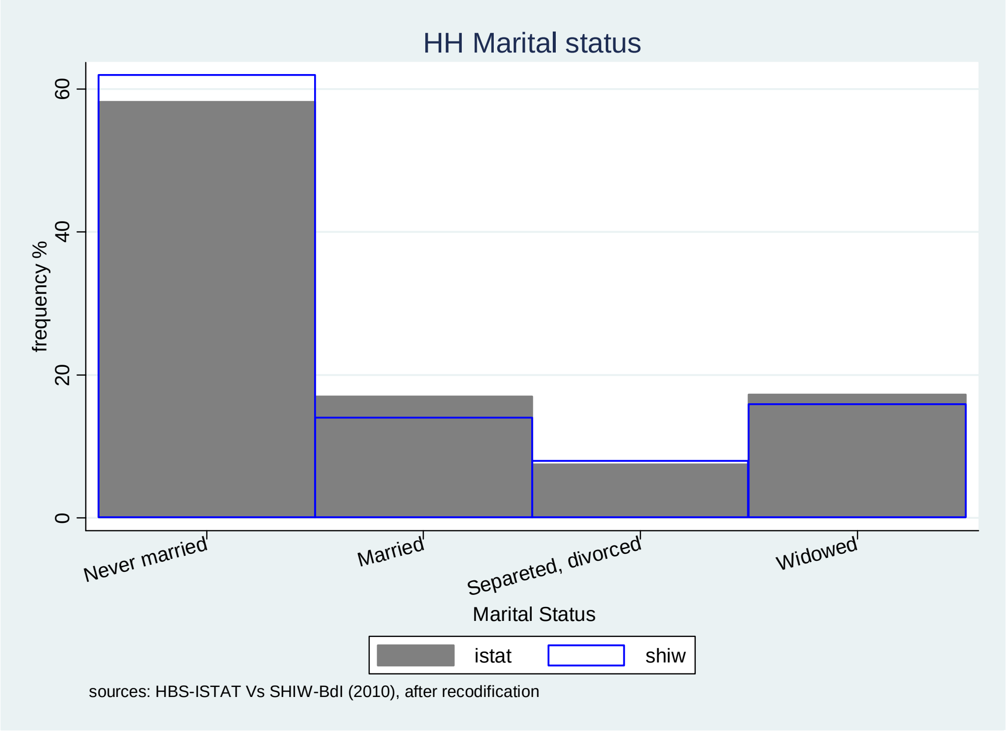
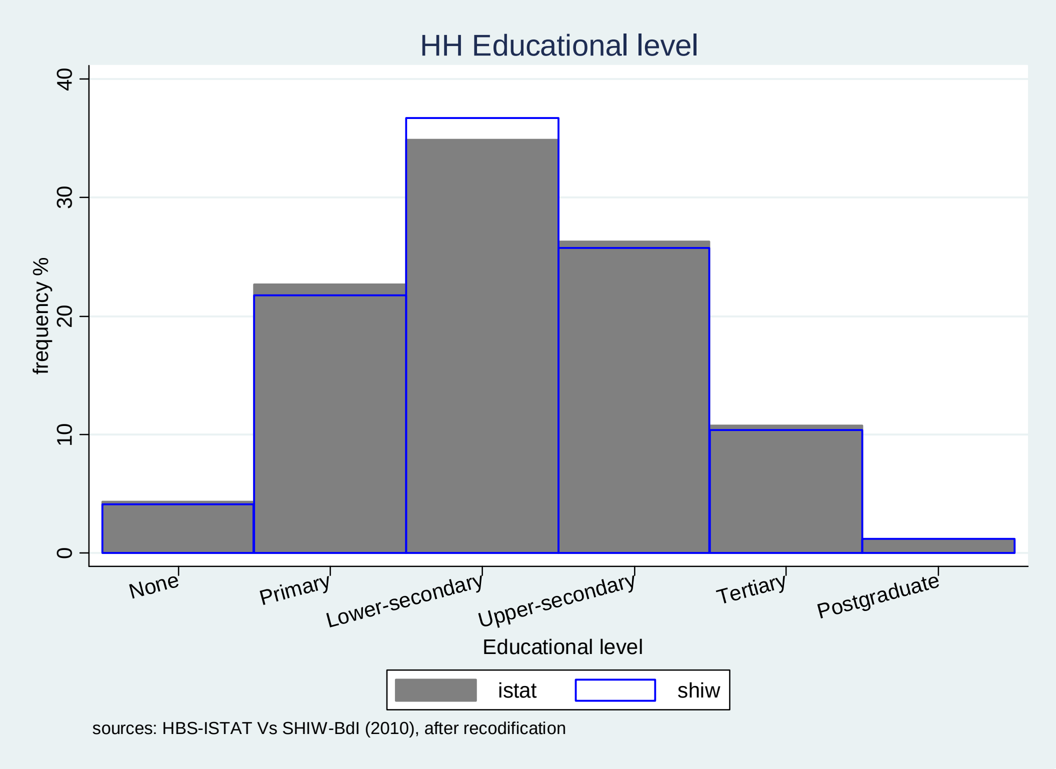


Figure A.4: HH marital status distribution SHIW vs HBS



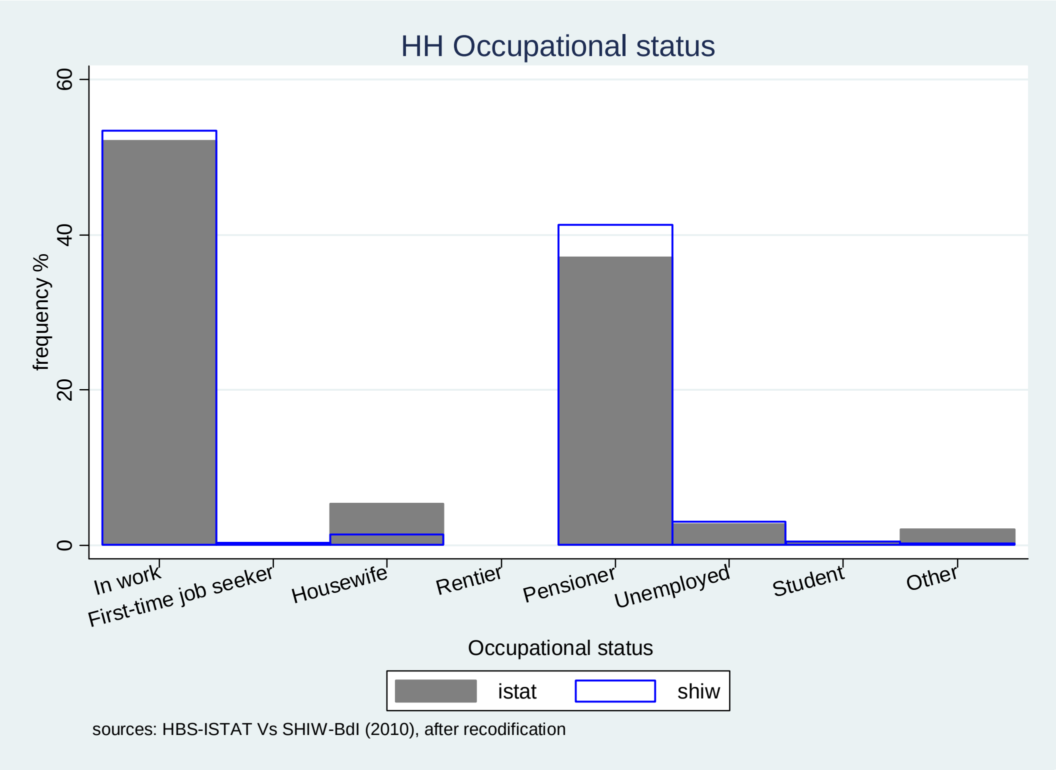
sources: HBS-ISTAT Vs SHIW-BdI (2010), after recodification

Figure A.5: HH educational level distribution SHIW vs HBS



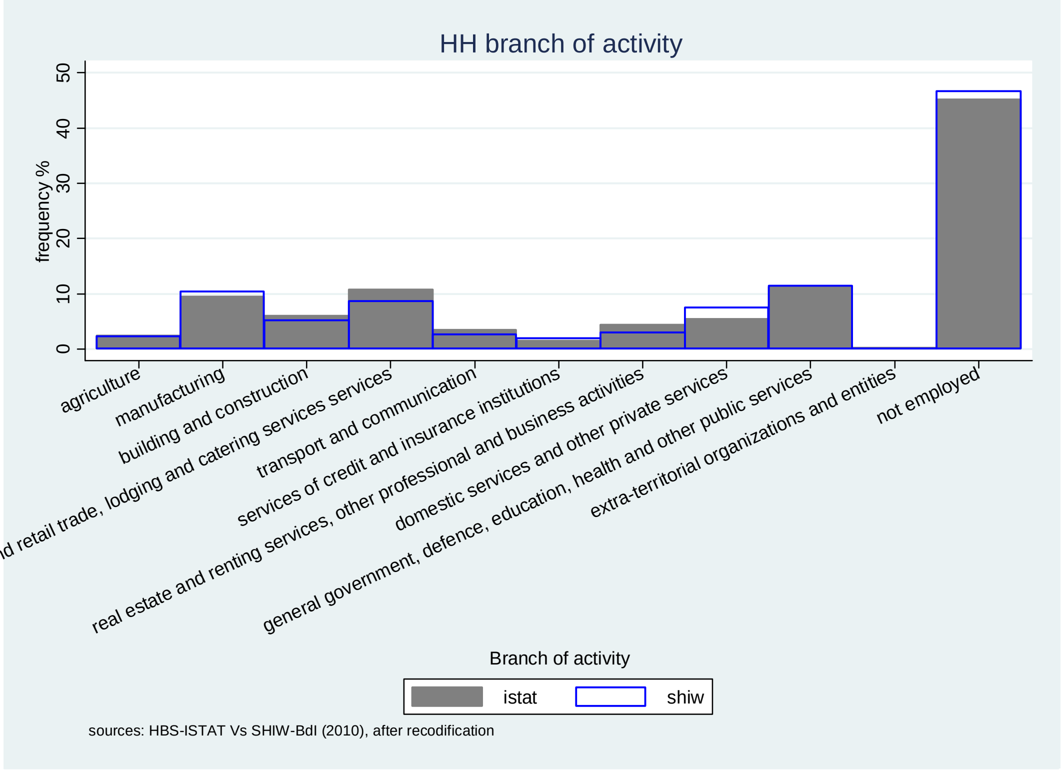
sources: HBS-ISTAT Vs SHIW-BdI (2010), after recodification

Figure A.6: HH occupational status distribution SHIW vs HBS



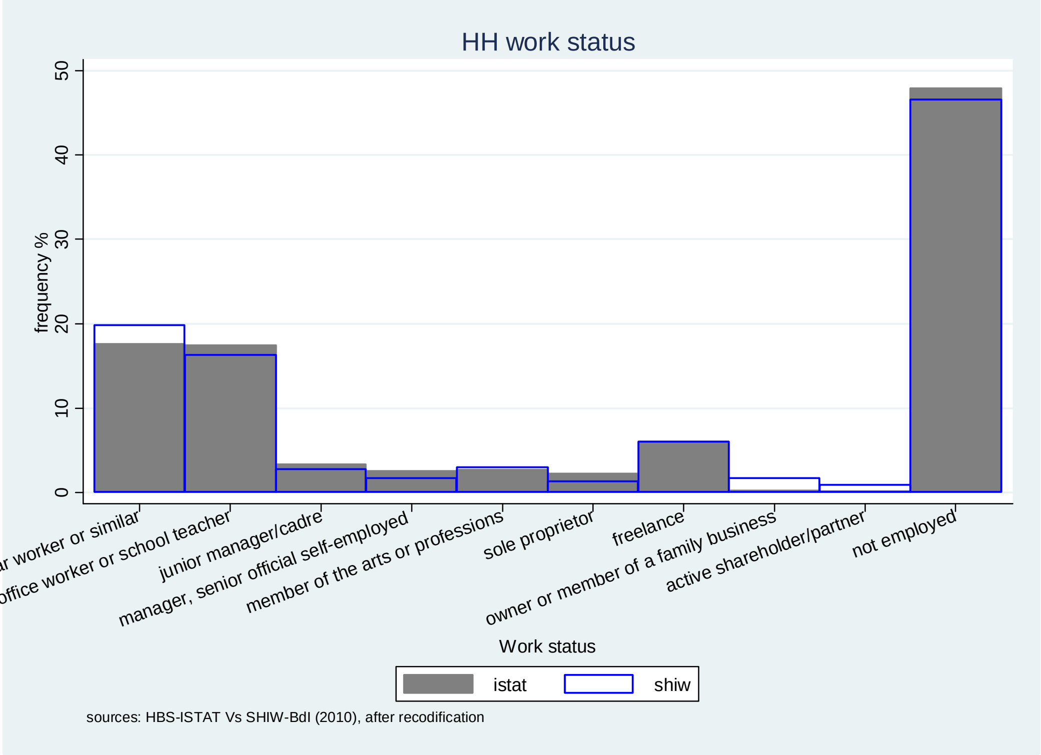
sources: HBS-ISTAT Vs SHIW-BdI (2010), after recodification

Figure A.7: HH branch of activity distribution SHIW vs HBS



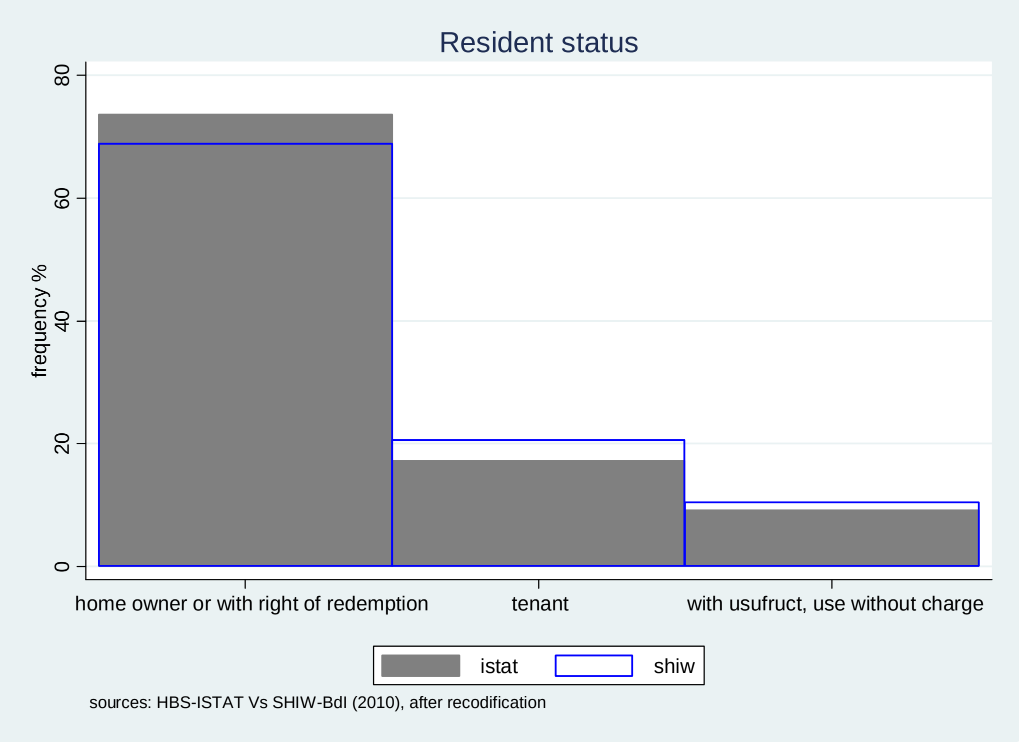
sources: HBS-ISTAT Vs SHIW-BdI (2010), after recodification

Figure A.8: HH work status distribution SHIW vs HBS



sources: HBS-ISTAT Vs SHIW-BdI (2010), after recodification

Figure A.9: Resident status distribution SHIW vs HBS



sources: HBS-ISTAT Vs SHIW-BdI (2010), after recodification

Table A.1: Total matching consumption distribution.
Original HBS donor file, original SHIW and fused mahalanobis 100 cells.

Total matching consumption, HBS original file	
Percentiles	Smallest

1%	2778,36	603,96		
5%	4474,94	833,52		
10%	5816	840,36	Obs	24898173
25%	9085,2	993,24	Sum of Wgt.	24898173
50%	14151,83		Mean	17064,16
		Largest	Std. Dev.	11914,04
75%	21562,14	112020,9		
90%	31502,39	132840	Variance	1,42E+08
95%	39479,36	138302,8	Skewness	2,192739
99%	61080,95	170324,5	Kurtosis	11,71125
Total matching consumption, SHIW original file				
	Percentiles	Smallest		
1%	3593,455	1194,01		
5%	5999,407	1194,498		
10%	7202,412	1201,139	Obs	24109883
25%	9760,163	1202,563	Sum of Wgt.	24109883
50%	14497,26		Mean	17530,33
		Largest	Std. Dev.	11350,17
75%	21604,83	120505		
90%	30495,6	124489,5	Variance	1,29E+08
95%	38068,75	136339,8	Skewness	2,538544
99%	59333,68	144002,3	Kurtosis	15,85671
Total matching consumption, Fused file				
	Percentiles	Smallest		
1%	3243,97	1230,55		
5%	4787,93	1239,14		
10%	5942,04	1343,52	Obs	24109883
25%	9181,06	1611,84	Sum of Wgt.	24109883
50%	14287,15		Mean	17286,4
		Largest	Std. Dev.	11891,18
75%	21755,72	101165,4		
90%	32635,92	101165,4	Variance	1,41E+08
95%	39806,34	106295,2	Skewness	2,021056
99%	61376,95	110282,4	Kurtosis	9,718745

Table A.2. Average conditional TMC by household typology

Total matching consumption by household typology				
Original HBS				
Household typology = 1: single person or couple without childrenn with HH under 35				
Obs	Mean	Std. Dev.	Min	Max
1150199	14991.18	11017.48	2094.53	110373
Household typology = 2: single person aged 35-64				
Obs	Mean	Std. Dev.	Min	Max
3006973	13480.45	9269.116	1230.55	93228.54
Household typology = 3: single person over 65				
Obs	Mean	Std. Dev.	Min	Max
3746461	9351.396	6456.601	603.96	60385.13
Household typology = 4: couple without children with HH aged 35-64				
Obs	Mean	Std. Dev.	Min	Max
1960631	19169.35	12034.21	1798.56	107711.8

Household typology = 5: couple without children with HH over 65				
Obs	Mean	Std. Dev.	Min	Max
2725666	15281.36	10159.86	1440.72	102189.7
Household typology = 6: couple with one child				
Obs	Mean	Std. Dev.	Min	Max
4135496	20391.91	11915.52	2491.63	112020.9
Household typology = 7: couple with two children				
Obs	Mean	Std. Dev.	Min	Max
3804988	22625.27	13575.97	1239.14	170324.5
Household typology = 8: couple with three or more children				
Obs	Mean	Std. Dev.	Min	Max
917111	23726.25	13899.59	4359.88	103836.5
Household typology = 9: single-headed household with children				
Obs	Mean	Std. Dev.	Min	Max
2036744	16451.79	10750.38	1883.92	101030.2
Household typology = 10: other typologies				
Obs	Mean	Std. Dev.	Min	Max
1413904	19188.33	12434.36	1713.42	95368.65

Table A.3: Average conditional TMC by age groups

Total matching consumption by HH age-group				
Original HBS				
HH age group = 18-24				
Obs	Mean	Std. Dev.	Min	Max
135314	12112.73	7657.018	3136.83	58562.58
HH age group = 25-29				
Obs	Mean	Std. Dev.	Min	Max
542576	15725.92	12436.52	2887.76	110373
HH age group = 30-34				
Obs	Mean	Std. Dev.	Min	Max
1368267	16707.88	10349.91	2094.53	79554.16
HH age group = 35-39				
Obs	Mean	Std. Dev.	Min	Max
2201803	18306.26	11038.22	1627.08	111440
HH age group = 40-44				
Obs	Mean	Std. Dev.	Min	Max
2539969	18852.73	11098.97	1239.14	111291.5
HH age group = 45-49				
Obs	Mean	Std. Dev.	Min	Max
2591974	20449.28	13653.43	2258.52	170324.5
HH age group = 50-54				
Obs	Mean	Std. Dev.	Min	Max
2416636	20817.45	14543.53	1798.56	138302.8
HH age group = 55-59				
Obs	Mean	Std. Dev.	Min	Max
2249510	20056.96	12238.76	1230.55	100487.1
HH age group = 60-64				
Obs	Mean	Std. Dev.	Min	Max
2245216	18170.12	11854.09	1645.24	107711.8
HH age group = 65-69				
Obs	Mean	Std. Dev.	Min	Max
1918863	15932.05	10640.1	1334.37	106295.2
HH age group = 70-74				
Obs	Mean	Std. Dev.	Min	Max

2111430	14440.96	10118.82	833.52	95724.49

HH age group >74

Obs	Mean	Std. Dev.	Min	Max
4409333	11311.7	8776.041	603.96	110282.4

Fused file

Household typology = 1: single person or couple without children with HH under 35

Obs	Mean	Std. Dev.	Min	Max
1080013	14054.21	9738.408	2094.53	62139.73

Household typology = 2: single person aged 35-64

Obs	Mean	Std. Dev.	Min	Max
2242159	12723.84	11077.03	1230.55	87854.77

Household typology = 3: single person over 65

Obs	Mean	Std. Dev.	Min	Max
3253380	8395.89	4631.958	1343.52	42346.44

Household typology = 4: couple without children with HH aged 35-64

Obs	Mean	Std. Dev.	Min	Max
1864970	19483.04	12938.08	3420.91	100487.1

Household typology = 5: couple without children with HH over 65

Obs	Mean	Std. Dev.	Min	Max
3445928	15966.31	9580.086	2513.64	101165.4

Household typology = 6: couple with one child

Obs	Mean	Std. Dev.	Min	Max
4371016	21131.23	12349.65	2713.44	110282.4

Household typology = 7: couple with two children

Obs	Mean	Std. Dev.	Min	Max
3966771	23018.35	12461.85	1239.14	98642.93

Household typology = 8: couple with three or more children

Obs	Mean	Std. Dev.	Min	Max
1136521	22576.54	14302.08	5928.92	85476.75

Household typology = 9: single-headed household with children

Obs	Mean	Std. Dev.	Min	Max
1113027	14157.99	7936.861	3498.55	52046.78

Household typology = 10: other typologies

Obs	Mean	Std. Dev.	Min	Max
1636098	17912.18	10480.18	2071.08	93333.54

TMC by HH age-group

Fused file

HH age group = 18-24

Obs	Mean	Std. Dev.	Min	Max
196308	13018.2	7419.034	2921.64	42236.53

HH age group = 25-29

Obs	Mean	Std. Dev.	Min	Max
568674	14757.1	9817.288	2393.52	46504.74

HH age group = 30-34

Obs	Mean	Std. Dev.	Min	Max
1277428	16158.12	9856.93	2094.53	62139.73

HH age group = 35-39

Obs	Mean	Std. Dev.	Min	Max
1627100	16680.78	10405.14	1614.96	72532.66

HH age group = 40-44

Obs	Mean	Std. Dev.	Min	Max
2973971	18395.17	11562.07	1239.14	87854.77

HH age group = 45-49				
Obs	Mean	Std. Dev.	Min	Max
2448799	21649.71	15447.22	1230.55	100487.1
HH age group = 50-54				
Obs	Mean	Std. Dev.	Min	Max
2328081	20957.73	12174.32	2855.4	98642.93
HH age group = 55-59				
Obs	Mean	Std. Dev.	Min	Max
1965062	21418.97	13850.38	1944.36	106295.2
HH age group = 60-64				
Obs	Mean	Std. Dev.	Min	Max
2400896	19625.72	12117.33	3337.39	91760.13
HH age group = 65-69				
Obs	Mean	Std. Dev.	Min	Max
1978281	16616.81	11490.12	2196.24	101165.4
HH age group = 70-74				
Obs	Mean	Std. Dev.	Min	Max
2297002	14291.92	9263.639	1611.84	93377.74
HH age group >74				
Obs	Mean	Std. Dev.	Min	Max
4048281	11515.87	7610.067	1343.52	110282.4

Table A.4: OC distribution.
Original HBS donor file, original SHIW and fused mahalanobis 100 cells.

Overall consumption, HBS original file				
	Percentiles	Smallest		
1%	5746.8	1419.96		
5%	8473.62	1614.96		
10%	10457.39	1651.68	Obs	24898173
25%	14843.05	1687.92	Sum of Wgt.	24898173
50%	21372.13		***Mean***	24546.01
		Largest	***Std. Dev.***	14270.13
75%	30526.04	136480.8		
90%	42091.01	155877.8	Variance	2.04E+08
95%	51462.27	176423.3	Skewness	1.922262
99%	75314.77	180319.5	Kurtosis	9.844588
Overall consumption, SHIW original file				
	Percentiles	Smallest		
1%	6084	1614		
5%	9600	1614		
10%	11300	2560	Obs	24109883
25%	15600	2760	Sum of Wgt.	24109883
50%	21600		***Mean***	25490.42
		Largest	***Std. Dev.***	16169.73
75%	30800	190000		

90%	43100	191000	Variance	2.61E+08
95%	54000	192000	Skewness	2.993276
99%	87550	195000	Kurtosis	20.34601
Overall consumption, Fused file				
Percentiles		Smallest		
1%	5923.8	1614.96		
5%	8592.439	1614.96		
10%	10296.6	2243.52	Obs	24109883
25%	14491.55	2298.43	Sum of Wgt.	24109883
50%	20984.62		***Mean***	25044.91
		Largest	***Std. Dev.***	15868.85
75%	30987.02	130301.6		
90%	44284.46	131314.8	Variance	2.52E+08
95%	54246.55	131314.8	Skewness	2.218352
99%	84935.41	136292.2	Kurtosis	11.11116

Table A.5: Overall average propensity to consume by HH age group

Age group	**SHIW**	**fused file**	**Diff**
18-24	1.264518	1.25865	0.005868
25-29	1.276727	1.262381	0.014346
30-34	0.8216971	0.7989159	0.022781
35-39	1.064749	1.076904	-0.01216
40-44	0.9385435	0.9581569	-0.01961
45-49	0.9947715	1.066917	-0.07215
50-54	0.8381405	0.8541405	-0.016
55-59	0.8580751	0.8829744	-0.0249
60-64	0.8053238	0.796044	0.00928
65-69	0.8058382	0.8183613	-0.01252
70-74	0.8309128	0.8226621	0.008251
>74	0.8412586	0.8045149	0.036744

Table A.6: Overall average propensity to consume by household typology

Household typology	SHIW	fused file	Diff
single person or couple without children with HH under 35	1.168715	1.132149	0.036566
single person aged 35-64	1.076473	1.057753	0.01872
single person over 65	0.883488	0.821588	0.0619
couple without children with HH aged 35-64	0.850123	0.934973	-0.08485
couple without children with HH over 65	0.811505	0.825037	-0.01353
couple with one child	0.820695	0.836867	-0.01617
couple with two children	0.886021	0.905178	-0.01916
couple with three or more children	1.017862	1.066607	-0.04875
single-headed household with children	0.866407	0.869302	-0.0029
other typologies	0.798209	0.803965	-0.00576

Table A.7: Bootstrap inference on total matching consumption means and standard deviations conditional on household typology

			donor HBS							Fused file			
Bootstrap	results	Number of obs	=	22246			Bootstrap	results	Number of obs	=	7951		
		Replications	=	250					Replications	=	250		
	Observed	Bootstrap			Normal-based		Observed	Bootstrap				Normal-based	
	Coeff.	Std. Err.	Z	P>z	[95% Conf. Interval]			Coeff.	Std. Err.	Z	P>z	[95% Conf. Interval]	
mean1	14991.18	465.5346	32.2	0	14078.74	15903.61	mean1	14054.21	755.8252	18.59	0	12572.82	15535.6
mean2	13480.45	228.6864	58.95	0	13032.24	13928.67	mean2	12723.84	656.5534	19.38	0	11437.02	14010.66
mean3	9351.396	152.1671	61.45	0	9053.154	9649.638	mean3	8395.89	187.4288	44.8	0	8028.536	8763.243
mean4	19169.35	355.4566	53.93	0	18472.67	19866.03	mean4	19483.04	925.9968	21.04	0	17668.12	21297.96
mean5	15281.36	237.8238	64.25	0	14815.23	15747.48	mean5	15966.31	418.9845	38.11	0	15145.12	16787.51
mean6	20391.91	234.7766	86.86	0	19931.76	20852.07	mean6	21131.23	547.6721	38.58	0	20057.81	22204.65
mean7	22625.27	286.2632	79.04	0	22064.21	23186.34	mean7	23018.35	516.8313	44.54	0	22005.38	24031.32
mean8	23726.25	564.8745	42	0	22619.11	24833.38	mean8	22576.54	1092.243	20.67	0	20435.78	24717.29
mean9	16451.79	311.7102	52.78	0	15840.85	17062.73	mean9	14157.99	587.3752	24.1	0	13006.76	15309.22
mean10	19188.33	374.2866	51.27	0	18454.74	19921.92	mean10	17912.18	635.2411	28.2	0	16667.13	19157.22
sd1	11017.48	1024.452	10.75	0	9009.586	13025.36	sd1	9738.408	906.5874	10.74	0	7961.529	11515.29
sd2	9269.116	416.3872	22.26	0	8453.012	10085.22	sd2	11077.03	1144.899	9.68	0	8833.072	13320.99
sd3	6456.601	231.3168	27.91	0	6003.228	6909.973	sd3	4631.958	203.2653	22.79	0	4233.565	5030.351
sd4	12034.21	485.7923	24.77	0	11082.07	12986.34	sd4	12938.08	2258.936	5.73	0	8510.648	17365.51
sd5	10159.86	365.148	27.82	0	9444.186	10875.54	sd5	9580.086	568.773	16.84	0	8465.312	10694.86
sd6	11915.52	399.9535	29.79	0	11131.62	12699.41	sd6	12349.65	811.1851	15.22	0	10759.76	13939.54
sd7	13575.97	510.3534	26.6	0	12575.69	14576.24	sd7	12461.85	501.1425	24.87	0	11479.63	13444.07
sd8	13899.59	688.4086	20.19	0	12550.33	15248.85	sd8	14302.08	1265.655	11.3	0	11821.44	16782.72
sd9	10750.38	515.9242	20.84	0	9739.186	11761.57	sd9	7936.861	813.1532	9.76	0	6343.11	9530.612
sd10	12434.36	477.6902	26.03	0	11498.1	13370.61	sd10	10480.18	658.298	15.92	0	9189.942	11770.42

Table A.8: Bootstrap inference on total matching consumption means and standard deviations conditional on household head's age group

donor HBS							Fused file						
Bootstrap	results	Number of obs	=	22246			Bootstrap results		Number of obs		=	7951	
		Replications	=	250					Replications	=	250		
	Observed	Bootstrap			Normal-based			Observed	Bootstrap			Normal-based	
	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]		Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
mean4	12112.73	918.4901	13.19	0	10312.53	13912.94	mean4	13018.2	1234.451	10.55	0	10598.72	15437.68
mean5	15725.92	793.0281	19.83	0	14171.61	17280.23	mean5	14757.1	1069.039	13.8	0	12661.82	16852.38
mean6	16707.88	394.5227	42.35	0	15934.63	17481.13	mean6	16158.12	736.6893	21.93	0	14714.23	17602
mean7	18306.26	327.5219	55.89	0	17664.33	18948.19	mean7	16680.78	715.2906	23.32	0	15278.83	18082.72
mean8	18852.73	299.2823	62.99	0	18266.15	19439.31	mean8	18395.17	616.3441	29.85	0	17187.16	19603.19
mean9	20449.28	344.9138	59.29	0	19773.26	21125.3	mean9	21649.71	990.1235	21.87	0	19709.11	23590.32
mean10	20817.45	454.9157	45.76	0	19925.83	21709.07	mean10	20957.73	554.9782	37.76	0	19870	22045.47
mean11	20056.96	300.1359	66.83	0	19468.71	20645.22	mean11	21418.97	816.4503	26.23	0	19818.76	23019.18
mean12	18170.12	316.1228	57.48	0	17550.53	18789.7	mean12	19625.72	555.1777	35.35	0	18537.59	20713.85
mean13	15932.05	295.9257	53.84	0	15352.05	16512.06	mean13	16616.81	748.4863	22.2	0	15149.8	18083.81
mean14	14440.96	281.8512	51.24	0	13888.55	14993.38	mean14	14291.92	471.3904	30.32	0	13368.01	15215.83
mean15	11311.7	159.9262	70.73	0	10998.25	11625.15	mean15	11515.87	288.5372	39.91	0	10950.34	12081.39
sd4	7657.018	1020.283	7.5	0	5657.301	9656.735	sd4	7419.034	720.4469	10.3	0	6006.984	8831.084
sd5	12436.52	1783.279	6.97	0	8941.352	15931.68	sd5	9817.288	938.3839	10.46	0	7978.09	11656.49
sd6	10349.91	550.6529	18.8	0	9270.647	11429.17	sd6	9856.93	730.4166	13.49	0	8425.339	11288.52
sd7	11038.22	409.6778	26.94	0	10235.27	11841.17	sd7	10405.14	896.1951	11.61	0	8648.63	12161.65
sd8	11098.97	378.965	29.29	0	10356.21	11841.72	sd8	11562.07	707.2309	16.35	0	10175.92	12948.22
sd9	13653.43	496.3395	27.51	0	12680.62	14626.23	sd9	15447.22	1638.095	9.43	0	12236.62	18657.83
sd10	14543.53	808.4567	17.99	0	12958.99	16128.08	sd10	12174.32	635.1639	19.17	0	10929.42	13419.22
sd11	12238.76	385.5837	31.74	0	11483.03	12994.49	sd11	13850.38	1064.193	13.01	0	11764.6	15936.16
sd12	11854.09	380.938	31.12	0	11107.47	12600.71	sd12	12117.33	665.4449	18.21	0	10813.08	13421.58
sd13	10640.1	418.1363	25.45	0	9820.57	11459.63	sd13	11490.12	809.848	14.19	0	9902.85	13077.4
sd14	10118.82	526.9358	19.2	0	9086.045	11151.6	sd14	9263.639	591.2931	15.67	0	8104.726	10422.55
sd15	8776.041	316.4693	27.73	0	8155.772	9396.309	sd15	7610.067	466.1778	16.32	0	6696.376	8523.759