**Noname manuscript No.**
(will be inserted by the editor)

# Statistical matching and uncertainty analysis in combining household income and expenditure data

**Pier Luigi Conti · Daniela Marella · Andrea Neri**

the date of receipt and acceptance should be inserted later

**Abstract** Among the goals of statistical matching, a very important one is the estimation of the joint distribution of variables not jointly observed in a sample survey but separately available from independent sample surveys. The absence of joint information on the variables of interest leads to uncertainty about the data generating model since the available sample information is unable to discriminate among a set of plausible joint distributions.

In the present paper a short review of the concept of uncertainty in statistical matching under logical constraints, as well as how to measure uncertainty for continuous variables is presented. The notion of matching error is related to an appropriate measure of uncertainty and a criterion of selecting matching variables by choosing the variables minimizing such an uncertainty measure is introduced. Finally, a method to choose a plausible joint distribution for the variables of interest via Iterative Proportional Fitting algorithm is described.

The proposed methodology is then applied to household income and expenditure data when extra sample information regarding the average propensity to consume is available. This leads to a reconstructed complete dataset where each record includes measures on income and expenditure.

P.L. Conti
Dipartimento di Scienze Statistiche, Sapienza Università di Roma; P.le A. Moro 5, 00185 Roma, Italy.
E-mail: pierluigi.conti@uniroma1.it

D. Marella
Dipartimento di Scienze della Formazione, Univeristà Roma TRE; Via D. Manin 53, 00185 Roma, Italy.
E-mail: daniela.marella@uniroma3.it

A. Neri
Banca d'Italia; Via Nazionale 91, 00184 Roma, Italy.
E-mail: andrea.neri@bancaditalia.it

**Keywords** Statistical matching · uncertainty · matching error · iterative proportional fitting

## 1 Introduction

Let $(Y, Z, X)$ be a three-dimensional variate, defined on an appropriate population, and let $\boldsymbol{s}_A$ and $\boldsymbol{s}_B$ be two independent samples of $n_A$ and $n_B$ records from $(Y, Z, X)$, respectively. The observational mechanism is such that $(i)$ only the variates $(Y, X)$ are observed in $\boldsymbol{s}_A$, and $(ii)$ only the variates $(Z, X)$ are observed in $\boldsymbol{s}_B$. The variable $X$ is *common* to the samples $\boldsymbol{s}_A$, $\boldsymbol{s}_B$, and plays the role of *matching variable*.

The main goal of statistical matching consists in estimating the joint distribution of $(Y, Z, X)$. Roughly speaking, two approaches have been considered. At first, techniques based on the conditional independence assumption between Y and Z given X (CIA assumption) were considered, see Okner (1972). Appropriateness of CIA is discussed in several papers. We cite, among others, Sims (1972) and Rodgers (1984). The second group of techniques uses external auxiliary information on the statistical relationship between Y and Z (e.g., an additional dataset where all the variables are jointly observed is available, as in Singh *et al.* (1993)).

As a matter of fact, the CIA is usually a misspecified assumption, while external auxiliary information is hardly ever available. The lack of joint information on the variables of interest is the cause of *uncertainty* about the model for $(Y, Z, X)$ since the available sample information is actually unable to discriminate among a set of plausible models for the variables of interest, see Conti *et al.* (2012), Conti *et al.* (2013), Conti *et al.* (2016a).

When extra sample information is available some models become illogical and must be excluded from the set of plausible distribution functions. As a consequence, the statistical model for the data becomes less uncertain. Clearly, each distribution in the class of plausible models can be taken as a surrogate of the actual joint distribution of $(Y, Z, X)$. Then, the statistical matching problem essentially consists in choosing a distribution in such a class.

The most favourable case, that for instance happens under CIA, occurs when the class of plausible distribution functions collapses into a single distribution function and the model is identifiable on the basis of sample data.

Our contribution to the existing literature is twofold. First of all, a method to choose a plausible joint distribution for the variables not jointly observed (that is, a *matching distribution*) from the set of equally plausible joint distributions for $(Y, Z, X)$ via the Iterative Proportional Fitting (IPF) algorithm is proposed. The reliability of the estimate provided by IPF is evaluated via an appropriate measure of uncertainty. Once a matching distribution has been estimated, a final dataset can be reconstructed in which each record includes measures on $(Y, Z, X)$. Secondly, we take into account the complexity of the sampling design (based on stratification, different level of clustering and inclusion probabilities proportional to an appropriate measure of size). The *i.i.d.*

assumption is hardly ever valid for sample surveys data, then the sample selection process must be taken into account in order to avoid misleading results; see Pfeffermann (1993) for an insightful discussion on this point.

Statistical matching in complex sample surveys is studied in Rubin (1986), Renssen (1998) and Wu (2004). The method proposed in Rubin (1986) has been seldom used in practice, since it requires the knowledge of inclusion probabilities of the units in one sample under the sampling design of the other sample. Renssen (1998) approach consists in calibrating the actual survey weights of the two distinct samples $s_A$ and $s_B$ to the common information in the two samples, in order to have compatible distributions on $(Y, X)$ and $(Z, X)$ estimated on $s_A$ and $s_B$, respectively. Further contributions are in Wu (2004). D'Orazio *et al.* (2009) show by simulation that the methods in Rubin (1986), Renssen (1998) and Wu (2004) are nearly equivalent in terms of estimable parameters and bounds of the uncertainty set of distributions. In Conti *et al.* (2016b) the choice of the estimators is based on attainment of good asymptotic properties going beyond the computation of finite sample size efficiency of the different estimators. For this reason, no comparison with the other estimators is proposed, given that their asymptotic behavior is not clear.

In Italy, the main sources used for households income and expenditures are the Banca d'Italia Survey on Household Income and Wealth (SHIW, for short) and the Italian National Statistical Household Budget Survey (HBS, thereafter), respectively. However, no single data source containing information on both expenditures and incomes currently exists.

Household-level data on income and consumption expenditure are widely used by policy makers and empirical researchers to provide insights into a number of areas. A first important field of research relates household's saving decisions. Many studies have focused on the reasons why people save, trying to quantify the importance of precautionary or pension accumulation motives (see among others Kennickell and Lusardi (2004), Guiso *et al.* (1992), Caballero and Ricardo (1990)).

A second area of research relates the reaction of household expenditure /saving to temporary and permanent income changes. These changes may reflect both external shocks such as financial distress, job losses, tax reforms and changes in the pension system, see Browining and Collado (2001), Browining and Collado (1996). Another field of research relates the analysis of household economic well-being. It is widely accepted that both income and consumption are not sufficient measures of achieved standards of living when considered separately. A better approach is to use both simultaneously.

Despite the importance of such topics, most countries do not have single sources of micro-data including high-quality disaggregated information on both incomes and expenditures. One of the main reasons is that collecting high-quality data on both topics requires a very large number of questions that would result in an excessive respondent burden. Quality expenditure data usually call for the use of diaries in which the household records all purchases made within a short period of time (at least for small and frequently purchased

items). The diary method minimizes the reliance on respondents' memories at a higher cost in terms of respondent burden. On the other hand, collecting high-quality information on income require asking each member of the household whether or not he/she has received a particular type of income. This should be done for all possible sources of income (self-employment, employment, pensions, return on assets, etc.). Moreover, it is also a good practice to collect additional data such as the type of work the respondent is engaged in, the type of pension received, the characteristics of a rented dwelling, and so on. As a consequence, since asking detailed questions on income and consumption in the same survey can be problematic, surveys tend to specialize in one of the two topics.

Browining *et al.* (2014) describe the alternative solutions available to economists in the existing literature to address this issue. One of the most widespread approach is to use statistical matching techniques to merge two or more sources of information, see D'Orazio *et al.* (2006a). These techniques usually are based on the CIA assumption. This approach has been widely used in the analysis of household's saving decisions. Skinner (1987) is the first to suggest imputing the total consumption expenditure of the Panel Survey of Income Dynamics respondent households (PSID), on the basis of the limited expenditure questions in the PSID and information from the Consumer Expenditure Survey. Cifaldi and Neri (2013) and Tedeschi (2013) use a similar approach to combine the information of SHIW with that coming from the HBS. Other studies have extended this procedure to allow for more flexible functional forms (Palumbo (1999)). For instance, Battistin *et al.* (2003) and Attanasio and Pistaferri (2014) model the relationship between total consumer expenditure and expenditure on a particular good as an inverse Engel curve.

The CIA assumption is particularly unappropriate when the matching relates consumption expenditure and income of households. In this paper an uncertainty analysis on the joint distribution of household income and expenditure under logical constraints regarding the average propensity to consume is performed, using SHIW and HBS datasets. The paper is organized as follows. Section 2 provides an overview on the uncertainty in statistical matching under logical constraints as well as how to measure uncertainty. Furthermore, the uncertainty is related to the matching error in order to evaluate how far is a matching distribution from the true distribution of the variables not jointly observed. Section 3 deals with the estimation of the uncertainty measures for complex survey data, as well as on choosing a *matching distribution*. In Section 4, the SHIW and HBS surveys are briefly described. In Section 5.1 an uncertainty analysis in combining household income and expenditure is performed and a new criterion for the matching variables selection is introduced. Finally, in Section 5.2 a method to pick a *matching distribution* from the set of plausible joint distributions for the variables of interest is proposed. Once such a joint distribution has been chosen, a "fused" SHIW dataset can be reconstructed.

## 2 Uncertainty in statistical matching

As previously stressed, the lack of joint information on the variables of interest is the cause of *uncertainty* about the model for $(Y, Z, X)$. Sub-section 2.1 is devoted to a short review of the concept of uncertainty in statistical matching under logical constraints, as well as how to measure uncertainty. In sub-section 2.2 the notion of matching error is introduced and related to the uncertainty measure in order to evaluate how far is a plausible joint distribution function for the variables not jointly observed (*matching distribution*) from the *true* distribution.

### 2.1 Uncertainty: definition and descriptive aspects

Let $\mathcal{U}_N$ be a finite population of $N$ units labeled by integers 1, ..., $N$, and denote by $Y$, $Z$, $X$ three characters of interest, taking values $y_i$, $z_i$, $x_i$, respectively, for unit $i$ $(i = 1, \ldots, N)$. Next, consider the indicators

$$I_{(y_i \leqslant y)} = \begin{cases} 1 \text{ if } y_i \leqslant y \\ 0 \text{ if } y_i > y \end{cases}, \quad i = 1, \ldots, N$$

and define similarly the indicators $I_{(z_i \leqslant z)}$ and $I_{(x_i \leqslant x)}$. The (finite) population (joint) distribution function (p.d.f.) of the three characters $Y$, $Z$, $X$ is:

$$H_N(y, z, x) = \frac{1}{N} \sum_{i=1}^{N} I_{(y_i \leqslant y)} I_{(z_i \leqslant z)} I_{(x_i \leqslant x)} \quad y, z, x \in \mathbb{R}.$$

Let

$$Q_N(x) = H_N(\infty, \infty, x), \;\; p_N(x) = Q_N(x) - Q_N(x^-) \tag{1}$$

be the marginal p.d.f. of $X$ and the proportion of population units such that $X = x$, respectively. From now on, we will assume that $X$ is a discrete character. Define further the conditional p.d.f.s

$$H_N(y, z\,|x) = \frac{1}{N\, p_N(x)} \sum_{i=1}^{N} I_{(y_i \leqslant y)} I_{(z_i \leqslant z)} I_{(x_i = x)}, \tag{2}$$

$$F_N(y\,|x) = H_N(y, \infty\,|x), \;\; G_N(z\,|x) = H_N(\infty, z|x). \tag{3}$$

Knowledge of the p.d.f.s $F_N(y\,|x)$, $G_N(z\,|x)$ *does not imply* knowledge of $H_N(y, z\,|x)$ (the most important exception occurs under CIA assumption). If only the p.d.f.s (3) were known, then one could only say that

$$\max(0, F_N(y|x) + G_N(y|x) - 1) \leqslant H_N(y, z|x) \leqslant \min(F_N(y|x), G_N(z|x)). \tag{4}$$

The bounds in (4) are the well-known *Fréchet bounds*. Fréchet bounds (4) can be improved when extra-sample information is available. In statistical practice, a kind of extra-sample information frequently available consists in

logical constraints, namely in restrictions on the support of $(Y, Z)|X$. Given $X = x$, the kind of constraints we consider is

$$a_x \leqslant f_x(y, z) \leqslant b_x, \tag{5}$$

where $f_x(y, z)$ is a monotone function of $y$ ($z$) for each $z$ ($y$). In case of *i.i.d.* observations, such constraints were first discussed in Conti *et al.* (2012), and used in Conti *et al.* (2013) in the special case of discrete ordinal variates $Y$, $Z$.

For instance, if $Y$ is the household expenditure, $Z$ is the household income, and $X$ the household size (*i.e.* the number of household components), using techniques of national accounting it is possible to produce fairly reasonable lower and upper bounds of the average propensity to consume (*apc*), namely of the ratio between consumption expenditure and income, for each household size. In this case $f_x(y, z) = Y/Z$.

Another kind of constraint frequently occurring in practice is $Y \geqslant Z$, given $X$. For instance, this is the case of Okner (1972) where $Y$ plays the role of total income and $Z$ plays the role of income subject to taxation.

Note that great caution is needed in defining the set of logical constraints. In fact, if the constraints are not compatible with the marginal information on $Y|X$ and $Z|X$ the set of plausible joint distributions is an empty set. Roughly speaking, as stressed in Vantaggi (2008), when logical constraints are used, global coherence of partial assessment drawn from different sources needed to be checked, and if coherence is not satisfied, incoherences have to be removed. The problem of resetting coherence can been faced by considering different criteria. D'Orazio *et al.* (2006b) restore consistency by using an iterative algorithm based on maximum likelihood estimates. Brozzi *et al.* (2012) adopt an approach based on the minimization of distances. Another plausible approach could be that of restoring consistency by changing as little as possible the marginal distributions.

Under the constraint (5), the Fréchet bounds (4) reduce to

$$K^x_{N-}(y, z) \leqslant H_N(y, z\,|x) \leqslant K^x_{N+}(y, z), \tag{6}$$

where using the notation $a \wedge b = min(a; b)$ it is not difficult to see that

$$K^x_{N-}(y, z) = \max(0, G_N(z\,|x) \wedge G_N(\gamma_y(a_x)\,|x) + F_N(y\,|x) \wedge F_N(\delta_z(b_x)\,|x) - 1, \\ F_N(y\,|x) + G_N(z\,|x) - 1) \tag{7}$$

$$K^x_{N+}(y, z) = \min(G_N(z\,|x), G_N(\gamma_y(a_x)\,|x), F_N(y\,|x), F_N(\delta_z(b_x)\,|x)) \tag{8}$$

with $\gamma_y(\cdot)$, $\delta_z(\cdot)$ being the inverse functions of $f_x(y, z)$ for fixed $y$ and $z$, respectively.

If $K^x_{N-}(y, z) \equiv K^x_{N+}(y, z)$ (for each $y$, $z$), then there is only *one* d.f. $H_N(y, z\,|x)$ satisfying (6). In this case, $H_N(y, z\,|x)$ is *identified*, and there is no uncertainty at all. The larger the distance between $K^x_{N-}(y, z)$ and $K^x_{N+}(y, z)$, the higher the uncertainty about $H_N(y, z\,|x)$.

Then, it is natural to use, as a measure of uncertainty on $H_N(y, z\,|x)$, a distance between $K^x_{N-}(y, z)$ and $K^x_{N+}(y, z)$. Using the same arguments

as in Conti *et al.* (2012), a simple measure of uncertainty on $H_N(y,\, z\,|x)$ *conditionally on* $x$ is

$$\Delta^x(F_N, G_N) = \frac{1}{N^2 p_N(x)^2} \sum_{i=1}^{N}\sum_{j=1}^{N} \left(K^x_{N+}(y_i,\, z_j) - K^x_{N-}(y_i,\, z_j)\right) I_{(x_i=x)} I_{(x_j=x)}$$
$$= \int_{R^2} \left(K^x_{N+}(y,\, z) - K^x_{N-}(y,\, z)\right) d[F_N(y\,|x) G_N(y\,|x)] \tag{9}$$

while an *unconditional uncertainty* measure of the $(Y,\, Z,\, X)$ joint distribution is

$$\Delta(F_N, G_N) = \sum_x \Delta^x(F_N, G_N) p_N(x). \tag{10}$$

Clearly, the unconditional uncertainty measure (10) is the average of the conditional uncertainty measures (9), w.r.t. the marginal distribution of $X$. An interesting property of the proposed uncertainty measures (either conditional or unconditional) is that their maximal value can be computed as shown in Proposition 1. Proof is in Appendix.

**Proposition 1** *The maximal value of uncertainty measures* $\Delta^x(F_N, G_N)(9)$ *and* $\Delta(F_N, G_N)$ (10) *is* $1/6 = 0.167$.

## 2.2 Matching error: the role of uncertainty measures in statistical matching

As previously stressed, even when the conditional p.d.f.s $F_N(y|x)$ and $G_N(z|x)$ are completely known, the lack of joint observations on the variables $(Y,\, Z,\, X)$ is the cause of uncertainty on $H_N(y,\, z|\, x)$. Roughly speaking, the available information is unable to discriminate among a set of plausible (joint) distributions for $(Y,\, Z)$ given $X$. The only thing we can say is that the true p.d.f. $H_N(y,\, z|\, x)$ belongs to the set

$$\mathcal{H}^x_N = \{H_N(y,\, z|\, x) : H_N(y,\, \infty|\, x) = F_N(y|x), H_N(\infty,\, z|\, x) = G_N(z|x),$$
$$a_x \leqslant f_x(y,\, z) \leqslant b_x\} \tag{11}$$

of all joint probability distributions of $(Y, Z)|X$ compatible with $F_N(y|x)$ and $G_N(z|x)$ and satisfying the imposed logical constraint. The measure of uncertainty (9) is, in a sense, a measure of the size of the class (11). If no further information are available, each d.f. in the class (11) is a plausible joint p.d.f. for $(Y,\, Z|\, X)$, *i.e.* is a plausible joint d.f. that matches $F_N(y|x)$ and $G_N(z|x)$ (*matching distribution*).

A statistical *matching procedure* essentially consists in picking a specific d.f. $\widetilde{H}_N(y,\, z|x)$ in the class $\mathcal{H}^x_N$ (11), and in using such a d.f. as if it was the "true" p.d.f. $H_N(y,\, z\,|x)$. $\widetilde{H}_N(y,\, z|x)$ is a *matching distribution* for $Y$ and $Z$ (given $X$), and plays the role of "blurred image" of the true p.d.f. $H_N(y,\, z\,|x)$.

Suppose now that a d.f. $\widetilde{H}_N(y,\, z|x)$ in the class $\mathcal{H}_N^x$ is used to match $F_N(y|x)$ and $G_N(z|x)$, but that the "true" d.f. of $(Y,\, Z|\, X)$ is $H_N(y,\, z|x)$, say. The discrepancy between $\widetilde{H}_N(y,\, z|x)$ and $H_N(y,\, z|x)$ is the *matching error*, that cannot be neither directly observed nor estimated on the basis of sample data. The notion of matching error is of basic importance in assessing the quality of matching procedures, because the smaller the matching error, the better the matching procedure.

Conditionally on $x$, the matching error at the point $(y,\, z)$ is

$$\epsilon_N^x(y,\, z) = |\widetilde{H}_N(y,\, z\,|x) - H_N(y,\, z\,|x)| \leqslant K^+(y,\, z\,|x) - K^-(y,\, z\,|x) \quad (12)$$

so that the overall matching error is given by

$$ME_x(\widetilde{H}_N, H_N) = \int \epsilon_N^x(y,\, z)\, dF_N(y\,|x) dG_N(z\,|x) \leqslant \Delta^x(F_N, G_N). \quad (13)$$

As a consequence, the uncertainty measure (9) can be interpreted as the maximal error occurring when the true p.d.f. $H_N(y,\, z|x)$ is replaced by a *matching distribution* $\widetilde{H}_N(y,\, z|x)$. Since $\Delta^x(F_N, G_N)$ only depends on the marginal d.f.s $F_N(y|x)$ and $G_N(z|x)$, it can be estimated on the basis of sample data in $s_A$ and $s_B$, respectively. In other words, the observed samples $s_A$, $s_B$ provide useful information on the maximal error occurring in matching $F_N(y|x)$ and $G_N(z|x)$, and hence on how reliable is the use of a *matching distribution*. This statement is strengthened by Proposition (1), that allows one to interpret how "small" or "large" is the value of the uncertainty measure if compared to its maximum 0.167.

A similar interpretation can also be given for the unconditional measure of uncertainty (10).

## 3 Estimating the uncertainty measures and choosing a matching distribution for complex survey data

In order to make inference on the uncertainty measures it is necessary to make assumptions on the sampling designs according to which the samples $\boldsymbol{s}_A$, $\boldsymbol{s}_B$ are drawn. Theoretical details are involved, and far from the goal of the present paper. For this reason, we confine ourselves to a short introduction. A wider theoretical treatment, with full details, is in Conti *et al.* (2016b). This section is devoted to the estimation of the uncertainty measures for complex survey data (sub-section 3.1). In sub-section 3.2 a method to choose a *matching distribution* for the variables of interest via IPF algorithm is proposed.

### 3.1 Plug-in estimates of uncertainty measures

For each unit $i$ of the finite population $\mathcal{U}_N$, let $D_{i,A}$ ($D_{i,B}$) be a Bernoulli random variable (r.v.), such that $i$ is in the sample $\boldsymbol{s}_A$ ($\boldsymbol{s}_B$) whenever $D_{i,A} = 1$

($D_{i,B} = 1$), whilst $i$ is not in $\boldsymbol{s}_A$ ($\boldsymbol{s}_B$) whenever $D_{i,A} = 0$ ($D_{i,B} = 0$). Let further $\pi_{i,A}$ ($\pi_{i,B}$) be the first order inclusion probabilities of the population units under the sampling design used to select $\boldsymbol{s}_A(\boldsymbol{s}_B)$.

The simplest approach to estimate the conditional uncertainty measure (9) consists in using a plug-in approach, *i.e.* in estimating $F_N(y|x)$ and $G_N(z|x)$ by their (Hájek) design-based estimators given by

$$\begin{aligned} \widehat{F}_H(y|x) &= \frac{\sum_{i=1}^N \frac{D_{i,A}}{\pi_{i,A}} I_{(y_i \leqslant y)} I_{(x_i=x)}}{\sum_{i=1}^N \frac{D_{i,A}}{\pi_{i,A}} I_{(x_i=x)}}, \\ \widehat{G}_H(y|x) &= \frac{\sum_{i=1}^N \frac{D_{i,B}}{\pi_{i,B}} I_{(z_i \leqslant z)} I_{(x_i=x)}}{\sum_{i=1}^N \frac{D_{i,B}}{\pi_{i,B}} I_{(x_i=x)}} \end{aligned} \tag{14}$$

and then in plugging such estimates in (9). In the sequel, we will denote by $\widehat{\Delta}_H^x$ the estimator of the uncertainty measure $\Delta^x(F_N, G_N)$ given by

$$\widehat{\Delta}_H^x = \frac{1}{\widehat{N}_A^x \widehat{N}_B^x} \sum_{i=1}^N \sum_{j=1}^N \left( \widehat{K}_+^x(y_i,\, z_j) - \widehat{K}_-^x(y_i,\, z_j) \right) \frac{D_{i,A}}{\pi_{i,A}} \frac{D_{i,B}}{\pi_{i,B}} I_{(x_i,x_j=x)} \tag{15}$$

where $I_{(x_i,x_j=x)} = I_{(x_i=x)} I_{(x_j=x)}$, $\widehat{K}_-^x$, $\widehat{K}_+^x$ are defined exactly as (7), (8), respectively, but with $F_N$ and $G_N$ replaced by the corresponding estimators $\widehat{F}_H$ and $\widehat{G}_H$, and $\widehat{N}_A^x$ and $\widehat{N}_B^x$ defined as

$$\widehat{N}_A^x = \sum_{i=1}^N \frac{D_{i,A}}{\pi_{i,A}} I_{(x_i=x)}, \quad \widehat{N}_B^x = \sum_{i=1}^N \frac{D_{i,B}}{\pi_{i,B}} I_{(x_i=x)} \tag{16}$$

We now turn to the problem of estimating the unconditional uncertainty measure. From the structure of (10), the following estimator can be defined

$$\widehat{\Delta}_H = \sum_{k=1}^K \widehat{\Delta}_H^{x^k} \widehat{p}_{H,AB}(x^k) \tag{17}$$

with

$$\widehat{p}_{H,AB}(x^k) = \tau_N^* \widehat{p}_{H,A}(x^k) + (1 - \tau_N^*) \widehat{p}_{H,B}(x^k) \tag{18}$$

where $\widehat{p}_{H,A}(x^k)$ and $\widehat{p}_{H,B}(x^k)$ are the Hájek estimators of $p_N(x^k)$ (for $k = 1, \ldots, K$) obtained from $\boldsymbol{s}_A$, $\boldsymbol{s}_B$, respectively, and $0 \leqslant \tau_N^* \leqslant 1$. As far as the value of $\tau_N^*$ is concerned, details are in Conti *et al.* (2016b), where the asymptotic normality of $\widehat{\Delta}_H^x$ and $\widehat{\Delta}_H$ is also proved. According to these asymptotic results the evaluation of the reliability of a matching distribution can be dealt with in terms of testing the hypotheses.

In Proposition 2 we confine ourselves to the asymptotic design-consistency (in the Brewer sense) of the estimators $\widehat{\Delta}_H^x$ and $\widehat{\Delta}_H$, which does not require any special regularity assumption on the sampling designs. Proof is in Appendix.

**Proposition 2** *The estimators* $\widehat{\Delta}_H^x(15)$ *and* $\widehat{\Delta}_H$ (17) *are* asymptotically design consistent.

### 3.2 Choosing a matching distribution

The goal of the present sub-section is to define a reasonable criterion to choose a *matching distribution* for $(Y, Z)|X$ in the class (11), with marginal d.f.s $F_N(y|x)$ and $G_N(z|x)$ replaced by their Hájek design-based estimators and satisfying the constraint (5). As already stressed, the smaller the estimate $\widehat{\Delta}_H^x$ of the uncertainty measure $\Delta^x(F_N, G_N)$, the closer the *matching distribution* to the true distribution of $Y$ and $Z$, given $X$.

We actually attack a slightly simplified version of this problem, where discretized versions of $Y$, $Z$ are considered. In order to select a *matching distribution* from $\mathcal{H}_N^x$ the following stepwise procedure can be used.

Step 1 The variables of interest $Y$ and $Z$ are first discretized by grouping their values in classes. Conditionally on $x$, denote by $Y_d$ and $Z_d$ the discrete counterparts of $Y$ and $Z$, where $Y_d$ has $r_x$ and $Z_d$ has $s_x$ outcomes, respectively. Furthermore, let $C^x$ be the contingency table with $r_x$ rows and $s_x$ columns and $m_{hj}^x$ the probability in cell $(h, j)$ of $C^x$, for $h = 1, \ldots, r_x$ and $j = 1, 2, \ldots, s_x$.

Step 2 Given $x$, the marginal probabilities $m_{h.}^x$ and $m_{.j}^x$, *i.e.* the probabilities that $Y_d$ falls into category $h$ and $Z_d$ falls into category $j$, respectively, can be estimated by

$$\widehat{m}_{h.}^x = \frac{\sum_{i=1}^N \frac{D_{i,A}}{\pi_{i,A}} I_{(y_i=h)} I_{(x_i=x)}}{\sum_{i=1}^N \frac{D_{i,A}}{\pi_{i,A}} I_{(x_i=x)}},\ \widehat{m}_{.j}^x = \frac{\sum_{i=1}^N \frac{D_{i,B}}{\pi_{i,B}} I_{(z_j=j)} I_{(x_i=x)}}{\sum_{i=1}^N \frac{D_{i,B}}{\pi_{i,B}} I_{(x_i=x)}} \qquad (19)$$

for $h = 1, 2, \ldots, r_x$ and $j = 1, 2, \ldots, s_x$.

Step 3 If the characters $Y$, $Z$ are discretized, then the constraints (5) become structural zeros in the contingency table $C^x$. The results is an incomplete table. The expected cell probabilities are then estimated *via* the iterative proportional fitting (IPF) algorithm.

The distribution obtained by the IPF algorithm matches both the estimated marginal d.f.s and the imposed constraint. However, if the constraint is not consistent the IPF algorithm does not converge. In this situation alternative solutions can be adopted.

(i) Change the initial estimated marginal probabilities $\widehat{m}_{h.}^x$ and $\widehat{m}_{.j}^x$ in order to restore consistency and estimate the expected cell probabilities *via* IPF.

(ii) Use the initial estimated marginal probabilities $\widehat{m}_{h.}^x$ and $\widehat{m}_{.j}^x$, and stop the IPF algorithm in a given number of iterations; this implies that just one marginal is matched.

(iii) Consider two consecutive iterations of the IPF algorithm and take the mean of the two fitted marginals.

## 4 The SHIW and HBS surveys

In Italy, the main sources used for estimating income and expenditures of households are the SHIW and HBS sample surveys. SHIW is conducted by Banca d'Italia every two years. Its main goal is to study the economic behaviors of Italian households. The sample for the SHIW survey is drawn in two stages, with municipalities and households as, respectively, the primary and secondary sampling units. The primary units are stratified by region and population size. Bigger municipalities (with more than $40,000$ inhabitants) are all included in the sample, while the smaller towns are selected using a probability proportional to size sampling (PPS). The individual households to be interviewed are then selected by simple random sampling. In the present paper we use the 2010 wave, whose sample consists of 7951 households and 387 municipalities. The main focus of the survey is the measurement of household income and wealth. The survey also includes some retrospective questions aimed at constructing a measure of total expenditure.

The HBS 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 (both durable and non-durable). The HBS survey is based on a two-stages sampling design similar to the SHIW survey. In the paper we use the 2010 wave. The sample is drawn in two stages with around 470 municipalities selected among two groups according to the population size at the first stage and 22227 households at the second stage. It is main goal is to measure total household consumption and its components.

Household income is defined as the combined disposable incomes of all people living in the household. It includes every form of income, *e.g.*, salaries and wages, self-employment income, retirement income, cash government transfers like unemployment benefits, and investment gains. The definition of household consumption used in the present paper includes the households' purchases of products for their everyday needs. It includes the expenditure for food and beverage, clothing and footwear, dwelling, fuels and electric power, for leisure, shows and education, for transport and communication, for health expenditures, and so on.

## 5 Beyond conditional independence: statistical matching between SHIW and HBS

The aim of this section is twofold. First of all, conditionally on $X$ in section 5.1 the maximal error arising from the combination of households income and expenditure under logical constraints regarding the propensity to consume, is studied. Furthermore, the criterion of selecting matching variables by choosing the variables minimizing such an error is introduced. Secondly, in section 5.2 a *matching distribution* for income and expenditure, that is a distribution lying in the class (11), is estimated on the basis of available sample data. An appli-

cation of statistical matching to household income and expenditure data is in Donatiello *et al.* (2016) where EU-SILC (EU Statistics on Income and Living Condition) 2012, with income reference year 2011, and the HBS 2011 are considered. However, in Donatiello *et al.* (2016) an uncertainty analysis is carried out discretizing the two variables of interest and no matching distribution is selected from the set of plausible joint distribution functions.

### 5.1 Uncertainty analysis: a new criterion to choose the matching variables

Roughly speaking, the literature highlights two main criteria for selecting the matching variables, see D'Orazio *et al.* (2006a). First of all, there must be both homogeneity in their statistical content and similarity in the distributions of the variables across the two surveys. Secondly, the variables must be significant in explaining variations in the target variables, in this case household expenditure and income. In the present section the criterion based on the unconditional uncertainty measure (17) is used to select the matching variables. A similar criterion based on an iterative procedure is proposed in D'Orazio *et al.* (2015), where categorical $X$, $Y$ and $Z$ variates are considered. However, in our approach the method of selecting matching variables by choosing the variables minimizing the uncertainty measure can be applied to both ordinal and continuous variables Y and Z.

The unconditional uncertainty measure is the average of the conditional uncertainty measures (9), w.r.t. the marginal distribution of X. Then, as $X$ changes, the unconditional uncertainty measure changes too. The criterion consists in choosing as matching variables those achieving the lowest level of uncertainty, namely the minimum "maximal error" occurring in combining household income and expenditure data. Clearly, the larger is the number of matching variables, the smaller is the number of observations used for estimating the conditional uncertainty measure with a reduction of the corresponding accuracy. Such a new criterion is not alternative but complementary to the previously described criteria. In our application, a set of variables have been considered as possible matching variables and have been harmonized across the two datasets. The set is composed by the variables: *ncomp=number of household components*, *area=geographical area of residence* and *condlav=occupational status*.

With regard to the first criterion, one of the main methods for evaluating the degree to which distributions of variables are similar across data sets is to compute a measure such as the Hellinger Distance (HD). It is generally considered that an HD of over 5% should raise concerns about the similarities in distributions. The HD is equal to 2.67, 2.43 and 5.47 for *ncomp*, *area* and *condlav*, respectively.

According to the second criterion, the common variables which should be used for matching are those that are statistically significant in explaining variations in both expenditure and income. Then an expenditure model was estimated on HBS data and an income model was estimated on SHIW data. Since

Table 1: Conditional Uncertainty Measure - X=number of household components

| *ncomp* | $n_{A,x}$ | $n_{B,x}$ | $a_x$ | $b_x$ | $r$ | $\widehat{\Delta}^x$ |
|---|---|---|---|---|---|---|
| 1 | 5851 | 1989 | 0.41 | 0.97 | 60 | 0.099 |
| 2 | 6292 | 2522 | 0.40 | 0.86 | 63 | 0.094 |
| 3 | 4758 | 1589 | 0.43 | 0.85 | 66 | 0.090 |
| 4+ | 5326 | 1851 | 0.49 | 0.99 | 66 | 0.087 |

both expenditure and income are highly positively skewed, the regression models were estimated on the logarithm of expenditure and income, respectively. Formally, the natural logarithm of household expenditure or household income, was modeled as a function of household characteristics. All the variables (*ncomp*, *area*, *condlav*) are statistically significant in explaining variations in both expenditure and income.

As far as the third criterion (based on the uncertainty measure (17)) is concerned, we assume that, conditionally on $X$, the constraints take the form $a_x \leqslant Y/Z \leqslant b_x$ where $Y$ and $Z$ denote the household expenditure and income, respectively. Then the ratio $apc = Y/Z$ represents the propensity to consume.

Since extra-sample information is not available, the bounds $a_x, b_x$ have been estimated by the ratio between the first quartile and the third quartile of expenditure in HBS and the median of income in SHIW, respectively, using the results in Cifaldi and Neri (2013), Tedeschi (2013), and Battistin *et al.* (2003). All these papers compare household expenditure data coming from the two surveys and show that SHIW underestimates households expenditure. This is also coherent with the fact that HBS is specialized on the measurement of household expenditure, while SHIW it is not. As a consequence, we may assume that for a given class of SHIW respondents (defined by their socio-demographic characteristics) the true expenditure lies between the SHIW and the HBS estimates. In order to define the bounds we prefer to use the quartiles of the expenditure distributions instead of the simple averages, obtaining more robust estimates.

We first develop a univariate uncertainty analysis to evaluate the effect on uncertainty measure of each possible matching variable independently. Next, we proceed to a bivariate analysis. Conditionally on $X = ncomp$, in Table 1 the sample sizes

$$n_{A,x} = \sum_{i=1}^{N} D_{i,A} I_{(x_i=x)}, \quad n_{B,x} = \sum_{i=1}^{N} D_{i,B} I_{(x_i=x)}, \tag{20}$$

the bounds $a_x$ and $b_x$, the percentage $r$ of sample observations that do not satisfy the constraint $a_x \leqslant apc \leqslant b_x$ and finally the conditional uncertainty measure are reported.

The same analysis has been performed also for both $X = area$ and $X = condlav$. The results are reported in Tables 2 and 3, respectively.

Table 2: Conditional Uncertainty Measure - X=area of residence

| *area* | $n_{A,x}$ | $n_{B,x}$ | $a_x$ | $b_x$ | $r$ | $\widehat{\Delta}^x$ |
|---|---|---|---|---|---|---|
| *North* | 9880 | 3477 | 0.42 | 0.95 | 63 | 0.094 |
| *Center* | 4157 | 1699 | 0.37 | 0.81 | 63 | 0.092 |
| *South and Islands* | 8190 | 2775 | 0.46 | 1.07 | 64 | 0.093 |

Table 3: Conditional Uncertainty Measure - X=occupational status

| *condlav* | $n_{A,x}$ | $n_{B,x}$ | $a_x$ | $b_x$ | $r$ | $\widehat{\Delta}^x$ |
|---|---|---|---|---|---|---|
| *Employed* | 8670 | 2605 | 0.46 | 0.93 | 65 | 0.089 |
| $Self-employed$ | 2510 | 784 | 0.40 | 0.85 | 67 | 0.083 |
| *Unemployed* | 582 | 251 | 0.67 | 1.49 | 74 | 0.065 |
| *Inactive* | 10465 | 4311 | 0.36 | 0.89 | 61 | 0.097 |

Conditionally on $X$, the value $\widehat{\Delta}^x$ in Tables 1,2 and 3 can be interpreted as the maximal error occurring when the true p.d.f. is replaced by a *matching distribution* belonging to the class (11). The larger error correspond to *Single* in Table 1, *North-Italy* in Table 2 and *Inactive* in Table 3, respectively.

As previously stressed, $r$ represents, in percentage terms, the effect of the constraint on the support reduction of the joint distribution of $(Y,Z)|X$. Clearly, the larger the reduction of support induced by a constraint, the larger the effect of the constraint on model uncertainty, *i.e.* the more informative the constraint. The average percentage of support reduction is equal to 63% for the *houselhold size* and the *geographical area of residence* and equal to 67% for the *occupational status*, respectively. Furthermore, as shown in Table 1,2 and 3 the admissible range for the *apc* is approximately the same as $X$ changes. These two factors helps to explain: (i) the strong reduction in the uncertainty measure when the constraint $a_x \leqslant apc \leqslant b_x$ is introduced; (ii) the small differences in the uncertainty measures as $X$ changes.

Table 4 shows the unconditional uncertainty measure (17) as the matching variables change. In order to assess the effect on the uncertainty measure coming from the introduction of an additional matching variable, the uncertainty analysis has been repeated for the following combinations : (*ncomp,area*), (*ncomp,condlav*). Roughly speaking, the constraint on *apc* halves the uncertainty on the data generating statistical model from 0.17 to 0.09, whatever the matching variables are.

From Table 4 the reduction of uncertainty as $X$ changes is approximately the same for different choices of $X$ variables. In conclusion, since the variable *condlav* has an HD larger the 5% and the uncertainty measure for *ncomp* is 0.092, we consider as final matching variable the *household size*.

Finally, conditionally on *household size* the same analysis has been repeated using alternative bounds for the *apc*. Conditionally on *household size*, the lower bound $a_x$ has been estimated using the 10th and the 20th percentile

Table 4: Overall Uncertainty Measure

| $X$ | $\widehat{\Delta}_H$ |
|---|---|
| *ncomp* | 0.092 |
| *area* | 0.093 |
| *condlav* | 0.091 |
| *ncomp,area* | 0.094 |
| *ncomp,condlav* | 0.092 |

Table 5: Conditional uncertainty measure as the constraint varies - X=ncomp

| $a_x$ | $\widehat{\Delta}^x$ | $a_x$ | $\widehat{\Delta}^x$ |
|---|---|---|---|
| 0.29 | 0.120 | 0.37 | 0.108 |
| 0.29 | 0.127 | 0.36 | 0.115 |
| 0.28 | 0.129 | 0.35 | 0.119 |
| 0.31 | 0.117 | 0.38 | 0.107 |

of the household propensity to consume distribution in SHIW, respectively. The upper bound $b_x$ is set equal to 1 for both cases.

Note that, the larger the set of possible values for the *apc* the smaller the reduction of the conditional uncertainty measure, that is less informative is the imposed constraint.The average percentage of support reduction is equal 45% and 53% for the 10th and 20th percentile, respectively.

## 5.2 Choosing a plausible distribution for the statistical matching between expenditure and income

The set of plausible d.f.s for $(Y, Z)|X$, given the sample information and the constraint $a_x \leqslant apc \leqslant b_x$ is $\mathcal{H}_N^x$, as defined in (11). This means that any d.f. in $\mathcal{H}_N^x$ can be used to estimate the true p.d.f. $H_N(y,\, z|\, x)$. Clearly, such a estimate can be used to perform the statistical matching between SHIW and HBS, that is to reconstruct a "fused file" in which each record includes measures on $(Y, Z, X)$.

In order to select a *matching distribution* from $\mathcal{H}_N^x$ the stepwise procedure described in Section 3.2 has been used. The discretization is performed by equal frequency binning. Then, conditionally on $X$ the thresholds of all bins are selected in such a way that all bins contain the same number of observations, equal to the square root of sample size. As a results, the size of each interval can be different. Conditionally on $X$, the number of bins $r_x$, $s_x$ for $Y$ and $Z$ are reported in Table 6. Furthermore, the discretization should be fine enough if a final SHIW dataset containing a continuous household expenditure value has to be reconstructed, since continuous data must be recovered from discretized data.

As far as step 3 is concerned, let $S^x$ be the set of cells consisting of all cells not containing structural zeros. In case of incomplete table, we can adopt the

Table 6: IPF results for X=number of household components

| $X$ | $r_x$ | $s_x$ | *accuracy level* |
|---|---|---|---|
| 1 | 75 | 43 | 0.0006 |
| 2 | 75 | 49 | 0.0003 |
| 3 | 67 | 37 | 0.0007 |
| 4+ | 70 | 41 | 0.0008 |

IPF to compute estimates expected cell values, except that the initial values must reflect the presence of structural zero cells, see Goodman (1968) and Bishop *et al.* (1975). This means that, in applying the IPF method the choice of the initial values must satisfy the quasi-independence relationship

$$m_{hj}^{x} = \delta_{hj} a_h^x b_j^x \tag{21}$$

for $h = 1, 2, \ldots, r_x$ and $j = 1, 2, \ldots, s_x$ where $\delta_{hj} = 1$ for cells $(h, j) \in S^x$ and $\delta_{hj} = 0$ otherwise. As initial values $\widehat{m}_{hj}^{0,x}$, that is at the $0th$ step of iterative algorithm, we set

$$\widehat{m}_{hj}^{0,x} = \delta_{hj} \widehat{m}_{h.}^{x} \widehat{m}_{.j}^{x} \tag{22}$$

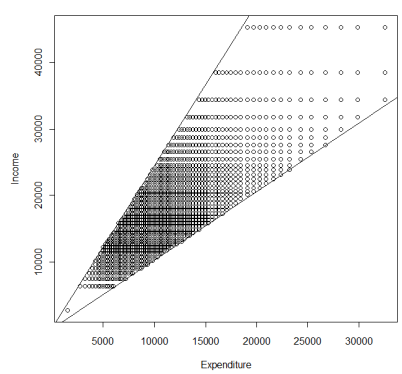
for all $(h, j) \in S^x$. Then IPF proportionally adjusts the values $\widehat{m}_{hj}^{t,x}$ in order to fit the marginals $\widehat{m}_{h.}^{x}$ and $\widehat{m}_{.j}^{x}$, respectively, until the desired level of accuracy is achieved. The fitted cells $\widehat{m}_{ij}^{x}$ represent a *matching distribution* for $(Y, Z)|X$. Conditionally on $X = ncomp$, in Table 6 the number of categories $r_x$, $s_x$ and the IPF achieved accuracy levels are reported. Furthermore, in Figures 1 and 2 the two-dimensional plot and the bivariate density estimate of the *matching distribution* is shown, respectively.

In Figure 1, conditionally on $X$, the two straight lines show the restriction on the support of the joint distribution of $(Y, Z)|X$ when the constraint $a_x \leqslant apc \leqslant b_x$ is introduced. Note that, in Figure 1 the frequency of the number of observations for each point is the largest integer less than or equal to $n_B \widehat{m}_{ij}^{x}$.

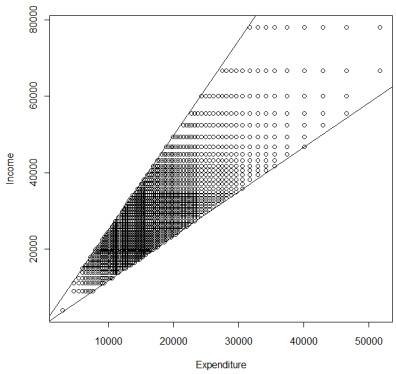
Once a *matching distribution* for $(Y, Z, X)$ has been estimated, a fused SHIW dataset can be reconstructed in which each record includes measures on $(Y, Z, X)$. Suppose that SHIW represents the recipient file and HBS the donor file. Conditionally on $X$, for each unit $k = 1, \cdots, n_B$ the following two step procedure can be applied: (i) given $(x_k, z_k)$ a categorical value for the expenditure $\tilde{y}_d$ is imputed choosing one of the plausible values of variable $Y_d$ with probabilities given by the IPF fitted cells $\widehat{m}_{ij}^{x} / \sum \widehat{m}_{.j}^{x}$; (ii) draw a donor unit in the class $C^x = \{i \in HBS : x_i = x,\ y_i \in \tilde{y}_d,\ a_x \leqslant y_i/z_k \leqslant b_x\}$ with probability proportional to sampling weights in HBS.

Note that, following Rässler (2002), four increasingly demanding levels of validity can be identified in the statistical matching problem: (i) preserving household values, (ii) preserving joint distributions, (iii) preserving correlation structures, (iv) preserving marginal distributions.

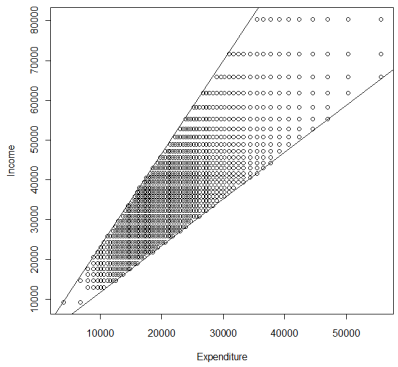
As stressed in Rässler (2002) the only way the first level validity can be assessed is by means of a simulation study, since the true household expenditure



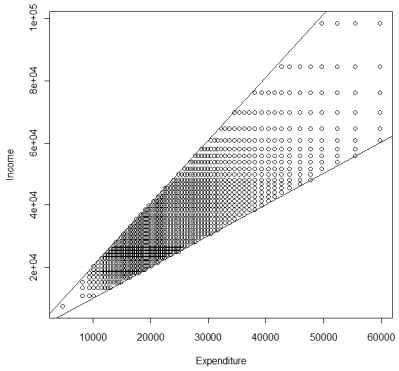

(a)




(b)

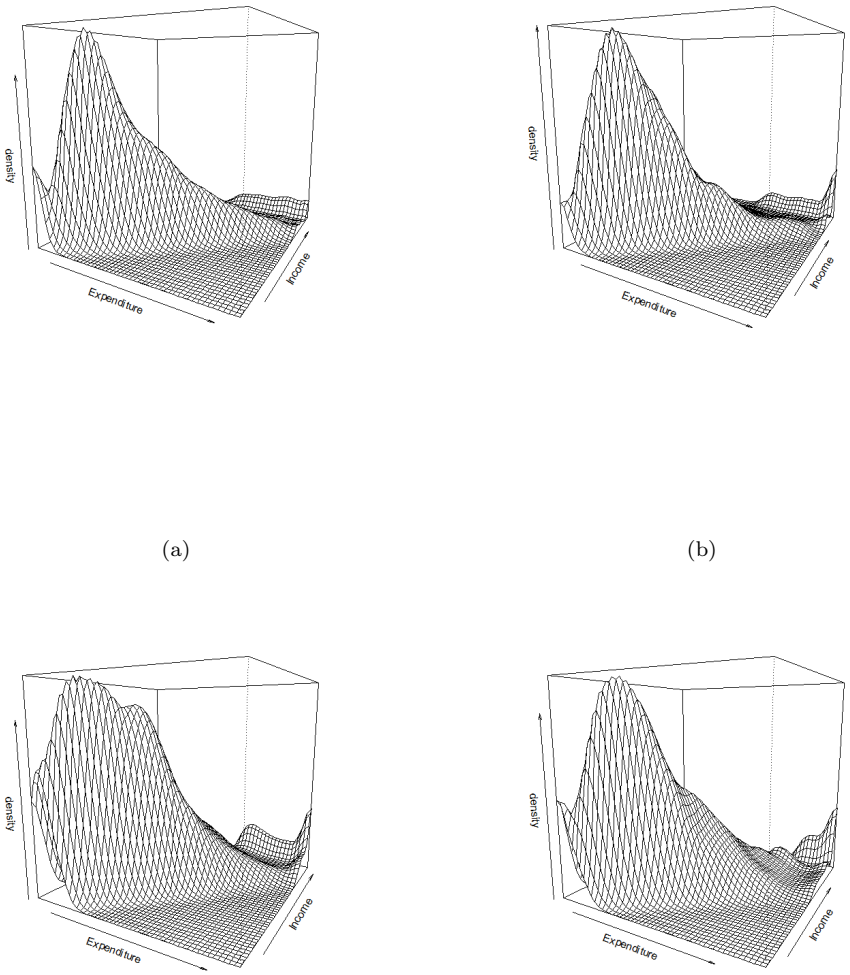



(c)




(d)

Fig. 1: Two-dimensional plots of matching distributions under the constraints $a_x \leqslant apc \leqslant b_x$ (a) x=1. (b) x=2. (c) x=3. (d) x=4+.




(a) (b)

(c) (d)

Fig. 2: Bivariate density estimates of matching distributions under the constraints $a_x \leqslant apc \leqslant b_x$ (a) x=1. (b) x=2. (c) x=3. (d) x=4+.

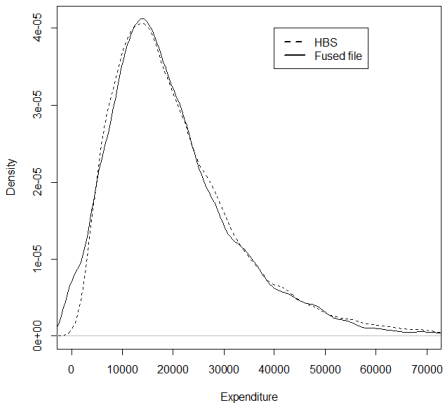



Fig. 3: Kernel density of overall expenditure in HBS and "fused" file

values are unknown. The second level requires the knowledge of the $(Y, Z, X)$ joint distribution. This distribution is unknown but, as previously stressed, the uncertainty measure can be used to asses how far is the *matching distribution* from the true joint distribution. Then, the smaller the uncertainty measure the more the *matching distribution* preserves the true joint distribution. Conditionally on the household size and under the constraint $a_x \leqslant apc \leqslant b_x$, this error is equal to 0.092.

In order to test the validity at the third level, the correlation observed in the original SHIW dataset between income and expenditure is 0.65, in the "fused" resulting SHIW dataset the correlation between imputed expenditure and income is 0.70.

Finally, as far as the fourth level of validity in Figure 3, the Kernel density of overall expenditure in HBS and in the "fused" dataset is reported. As expected, the procedure preserves the marginal distribution of expenditure in the "fused" dataset, as a consequence of IPF algorithm that proportionally adjust the initial values in order to fit the marginal distributions of income and expenditure in SHIW and HBS, respectively.

Then, the procedure proposed to choose a matching distribution in the class (11) always respects the fourth level of validity.

The same considerations hold when the bounds $a_x$ and $b_x$ are estimated as in Table 5.

## 6 Conclusions

In this paper an uncertainty analysis in combining household income and expenditure data under constraints regarding the average propensity to consume has been performed. The analysis allowed us: (i) to introduce a new criterion

to choose the matching variables in performing the statistical matching. The criterion consists in choosing as matching variables those achieving the lowest level of uncertainty, namely the minimum "maximal error" occurring in combining household income and expenditure data; (ii) to select a *matching distribution* from the class of equally plausible joint distributions (11) via IPF algorithm. As previously stressed, the matching distribution estimated via IPF algorithm is not preferable to another belonging to the class $\mathcal{H}_N^x$, its reliability is evaluated via the proposed uncertainty measure. Clearly, in order to apply the IPF algorithm the continuous variables of interest income and expenditures have been first discretized by grouping their values in classes.

Finally, once a *matching distribution* has been estimated, it can be used to impute expenditure microdata in SHIW. This leads to a "reconstructed complete dataset", characterized by an intrinsic matching error. By practitioners, although it can be used for inferential purposes, it cannot be considered as a genuine complete dataset, but only a "blurred image" of the actual joint distribution. The amount of blur is expressed by the uncertainty measure studied in the paper.

## Appendix

**Proof of Proposition 1** Taking into account that

1. $K_{N+}^x(y,\, z) \leqslant \min(F_N(y|x),\, G_N(z|x))$;
2. $K_{N-}^x(y,\, z) \geqslant \max(0,\, F_N(y|x) + G_N(z|x) - 1)$;

it is not difficult to see that

$$\begin{aligned}\Delta^x(F_N, G_N) \leqslant & \int_{R^2} \{\min(F_N(y|x),\, G_N(z|x)) \\ & - \max(0,\, F_N(y|x) + G_N(z|x) - 1)\}\; dF_N(y\,|x) dG_N(z\,|x) \\ \approx & \int_0^1 \int_0^1 \{min(u,\, v) - \max(0,\, u + v - 1)\}\; dudv \\ = & \frac{1}{6}. \end{aligned} \tag{23}$$

In other terms, the maximal value of the conditional measure of uncertainty (9) is essentially $1/6 \approx 0.167$. As an easy consequence of Proposition 1, also the unconditional uncertainty measure computed as in (10) takes the value $1/6$.

**Proof of Proposition 2** The following two statements hold:

$$\widehat{\Delta}_H^{*x} \xrightarrow{p} \Delta^x(F_N,\, G_N) \;\; as\; k \to \infty \tag{24}$$

$$\widehat{\Delta}_H^{*} \xrightarrow{p} \Delta(F_N,\, G_N) \;\; as\; k \to \infty \tag{25}$$

Asymptotic analysis requires to define how the samples sizes $n_A$, $n_B$ and the population size $N$ go to infinity. As in Brewer (1979) (cfr. also Little (1983)), this will be done as follows:

1. $k$ replicates of the original population are formed.
2. From each replicate, an independent sample $\boldsymbol{s}_A$ ($\boldsymbol{s}_B$) of size $n_A$ ($n_B$) is selected, according to the sampling design $P_A$ ($P_B$). Using notation introduced above, let $D^j_{i,A}$ ($D^j_{i,B}$) be a Bernoulli r.v. taking the value 1 if unit $i$ is included in the sample drawn from the $j$th replicate of the population ($j = 1, \ldots, k$) according to the sampling design $P_A$ ($P_B$), and the value 0 otherwise.
3. The $k$ populations are aggregated to a population of size $N^* = kN$. We will denote by $F_{N^*}(y\,|x)$, $G_{N^*}(z\,|x)$, $p_{N^*}(x)$ the conditional p.d.f.s of $Y$ and $Z$ given $X = x$ and the proportion of units such that $X = x$, respectively.
4. The $k$ samples drawn with the sampling design $P_A$ ($P_B$) are aggregated to a sample $\boldsymbol{s}^*_A$ ($\boldsymbol{s}^*_B$) of $n^*_A = kn_A$ ($n^*_B = kn_B$) units.
5. The quantities $F_{N^*}(y\,|x)$, $G_{N^*}(z\,|x)$, $p_{N^*}(x)$ are estimated by their Hájek estimators, as defined in sub-section 3.1, and based on $n^*_A$ and $n^*_B$ sample units. Such estimates are denoted by $\widehat{F}^*_H(y\,|x)$, $\widehat{G}^*_H(z\,|x)$, $\widehat{p}^*_H(x)$, respectively. Then, the uncertainty measures are estimated accordingly. We will denote by $\widehat{\Delta}^{*x}_H$ ($\widehat{\Delta}^*_H$) the estimate of the conditional (unconditional) measure of uncertainty.
6. $k$ is allowed to tend to infinity.

First of all, it is immediate to see that

$$F_{N^*}(y\,|x) = F_N(y\,|x), \;\; G_{N^*}(z\,|x) = G_N(z\,|x), \;\; p_{N^*}(x) = p_N(x).$$

In the second place, from

$$\widehat{F}^*_H(y\,|x) = \frac{\sum_{i=1}^N \left\{ \frac{1}{k} \sum_{j=1}^k \frac{D^j_{i,A}}{\pi_{i,A}} \right\} I_{(y_i \leqslant y)} I_{(x_i = x)}}{\sum_{i=1}^N \left\{ \frac{1}{k} \sum_{j=1}^k \frac{D^j_{i,A}}{\pi_{i,A}} \right\} I_{(x_i = x)}}$$

and using the law of large numbers

$$\frac{1}{k} \sum_{j=1}^k \frac{D^j_{i,A}}{\pi_{i,A}} \tag{26}$$

converges in probability to 1 as $k$ goes to infinity, then it is not difficult to see that $\widehat{F}^*_H(y\,|x)$ converges in probability to $F_N(y\,|x)$ as $k$ tends to infinity, for each $x$ and uniformly in $y$. In the same way, it is possible to show that $\widehat{G}^*_H(z\,|x)$ converges in probability to $G_N(z\,|x)$ as $k$ tends to infinity, for each $x$ and uniformly in $z$. Since the functional $\Delta^x(F_N, G_N)$ is continuous in the sup-norm, (24) is proved. In the same way, (25) can be proved.

## References

Attanasio, O., Pistaferri, L. (2014) Consumption inequality over the last half century: some evidence using the new PSID consumption measure. American Economic Review, 104, 122–126

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

Bishop, Y.M., Fienberg S.E., Holland, P.W. (1975) Discrete Multivariate Analysis. Springer, New-York

Blundell, R., Pistaferri, L., Preston, I. (2008) Consumption inequality and partial insurance. American Economic Review, 98, 1887–1921

Brewer, K.R.W. (1979) A Class of Robust Designs for Large-Scale Surveys. Journal of American Statistical Association, 74, 911–915

Browning, M. and Collado, M.D. (1996) Assessing the effectiveness of saving incentives. Journal of Economic Perspectives, 10, 4, 73–90

Browning, M., Collado, M.D. (2001) The Response of Expenditures to Anticipated Income Changes: Panel Data Estimates. American Economic Review, 91 ,3,681–692

Browning, M., Crossley, T.F., Winter, J. (2014) The Measurement of Household Consumption Expenditures. Annual Review of Economics, 6, 1, 475–501

Brozzi, A., Capotorti, A., Vantaggi, B. (2012) Incoherence correction strategies in statistical matching. International Journal of Approximate Reasoning, 53, 1124–1136

Caballero, R.J and Ricardo, J. (1990) Consumption puzzles and precautionary savings. Journal of Monetary Economics, 25, 1, 113-136

Cifaldi, G. and Neri, A. (2013) Asking income and consumption questions in the same survey: what are the risks? Bank of Italy, Economic Research and International Relations Area. Economic working papers, 908

Conti, P.L., Marella, D., Scanu, M. (2012) Uncertainty analysis in statistical matching. Journal of Official Statistics, 28, 69–88

Conti P.L., Marella D., Scanu M. (2013) Uncertainty Analysis for statistical matching of ordered categorical variables. Computational Statistics & Data Analysis, 68, 311–325

Conti, P.L., Marella, D., Scanu, M. (2016) How far from identifiability? A systematic overview of the statistical matching problem in a non-parametric framework. Communications in Statistics-Theory and Methods. DOI : 10.1080/03610926.2015.1010005

Conti, P.L., Marella, D., Scanu, M. (2016) Statistical matching analysis for complex survey data with applications. Journal of the American Statistical Association. DOI 10.1080/01621459.2015.1112803

Conti, P.L. (2014) On the estimation of the distribution function of a finite populatrion under high entropy sampling designs, with applications. Sankhya B, 76, 2, 234–259

Donatiellio, G., D'Orazio, M., Frattarola, D., Rizzi, A., Scanu, M., Spaziani, M. (2016) The role of the conditional independence assumption in statistically matching income and consumption. Statistical Journal of the IAOS, vol. Preprint, no. Preprint, 1–9

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

D'Orazio, M., Di Zio, M., Scanu, M. (2006) Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints. Journal of Offcial Statistics, 22, 137-157

DOrazio, M., Zio, M. D., Scanu, M. (2009). Uncertainty intervals for nonidentifiable parameters in statistical matching". In 57th Session of the International Statistical Institute, Durban (South Africa), August 2009.

D'Orazio, M., Di Zio, M., Scanu, M. (2015) The use of uncertainty to choose the matching variables in statistical matching. NTTS 2015 New Techniques and Technologies for Statistics and Exchange of Technology and Know-how, Brussels 10–12 March, 2015

Goodman, L.A. (1968) The analysis of cross-classified data: independence, quasi-independence, and interaction in contingency tables with or without missing cells. Journal of American Statistical Association, 63, 1091–1131

Guiso, L., Jappelli, T., Terlizzese, D. (1992) Earnings uncertainty and precautionary saving. Journal of Monetary Economics, 30, 2, 307-337

Kennickell, A. and Lusardi, A. (2004) Disentangling the Importance of the Precautionary Saving Motive. NBER working papers series, 10888, 1–64

Little, R.J.A. (1983) Estimating a Finite Population Mean From Unequal Probability Samples. Journal of American Statistical Association, 78, 596–604

Okner,B.A. (1972).Constructing a new database from existing microdata sets: the 1966 merge file. Annals of Economic and Social Measurement, 1, 325342.

Palumbo, M.G. (1999) Uncertain medical expenses and precautionary saving near the end of the life cycle. Review of Economic Studies, 66, 2, 395–421

Pfeffermann, D. (1993) The Role of Sampling Weights When Modeling Survey Data. Internatinal Statistical Review, 61, 317–337

Rässler, S. (2002) Statistical Matching: a Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. New York, Springer

Renssen, R.H. (1998) Use of Statistical Matching Techniques in Calibration Estimation. Survey Methodology, 24, 171–183

Rodgers, W.L. (1984) An Evaluation of Statistical Matching. Journal of Business and Economic Statistics, 2, 91–102

Rubin, D.B. (1986) Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. Journal of Business and Economic Statistics, 4, 87–94

Sims, C.A. (1972) Comments and Rejoinder (On Okner (1972)). Annals of Economic and Social Measurement, 1, 343–345, 355–357

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, 5979.

Skinner, J. (1987) A superior measure of consumption from the Panel Study of Income Dynamic. Economic Letters, 23, 213–216

Tedeschi, S. and Pisano, E. (2013) Data Fusion Between Bank of Italy-SHIW and ISTAT-HBS. MPRA Paper. RePEc:pra:mprapa:51253

Vantaggi, B. (2008) Statistical matching of multiple sources: A look through coherence. International Journal of Approximate Reasoning, 49, 701–711

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. The Canadian Journal of Statistics, 32, 112.