BANCA D'ITALIA

EUROSISTEMA

Temi di Discussione

(Working Papers)

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

by Pier Luigi Conti, Daniela Marella and Andrea Neri

July 2015

Number 1018

BANCA D'ITALIA
EUROSISTEMA

Temi di discussione

(Working papers)

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

by Pier Luigi Conti, Daniela Marella and Andrea Neri

Number 1018 - July 2015

The purpose of the Temi di discussione *series is to promote the circulation of working papers prepared within the Bank of Italy or presented in Bank seminars by outside economists with the aim of stimulating comments and suggestions.*

The views expressed in the articles are those of the authors and do not involve the responsibility of the Bank.

ISSN 1594-7939 (print)
ISSN 2281-3950 (online)

Printed by the Printing and Publishing Division of the Bank of Italy

STATISTICAL MATCHING AND UNCERTAINTY ANALYSIS IN COMBINING HOUSEHOLD INCOME AND EXPENDITURE DATA

by Pier Luigi Conti*, Daniela Marella* and Andrea Neri**

Abstract

The availability of microdata on both income and expenditure is highly recommended if one wants to assess the distributional consequences of policy changes. In Italy, the main sources used for estimating household income and expenditure are the Bank of Italy's Survey on Household Income and Wealth and the Italian National Institute of Statistics Household Budget Survey. However, there is no single data source containing information on both expenditure and income. The problem is generally overcome with statistical matching procedures based on the conditional independence (CIA) assumption. The aim of this paper is to present a method to combine information coming from different databases relaxing the CIA assumption. In particular we propose a method to combine household income and expenditure data under logical constraints regarding the average propensity to consume. We also propose an estimate of a plausible joint distribution function for household income and expenditure.

JEL Classification: C15, C14, C42.
Keywords: statistical matching, uncertainty, matching error, iterative proportional fitting.

Contents

*Dipartimento di Scienze Statistiche, Sapienza Università di Roma, **Dipartimento di Scienze della Formazione, Università Roma Tre, *** Banca d'Italia, Statistical Analyses Directorate.

1 Introduction

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)). These studies are relevant to policy since the reasons why people save influence their reaction to policies that imply more responsibility into insuring themselves against income or health /pension risks. Other researchers have focused on the role of other determinants of saving decisions (see for instance Jappelli and Pagano (1998)). Household financial education has been found to be a strong predictor of risk diversification, portfolio allocations and in the end of the levels of savings (see among others Banks and Oldfield (2007), Lusardi and Mitchell (2014)). Such a result has also important implications for welfare since individuals are increasingly in charge of their own financial security, especially after retirement.

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 for instance Browning and Collado (2001), Browning and Collado (1996)).

Another field of research relates the analysis of household economic well-being. It is widely accepted that neither income nor consumption are sufficient measures of achieved standards of living when considered separately. A better approach is to use both simultaneously (Meyer and Sullivan (2003)).

Despite the importance of such topics, most countries do not have single sources of micro-data including high-quality disaggregated information on both income and consumption expenditure. 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 burden on the respondent. 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.

Collecting high-quality information on income require asking all members of the household whether or not they had 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.

Browning *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.* (2006)). These techniques usually are based on the Conditional Independence Assumption (CIA, for short). Appropriateness of CIA is often questionable and discussed in several papers (see, among others, Sims (1972) and Rodgers (1984)).

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. The method is based on the propensity score method proposed by Rosenbaum and Rubin (1983). Cifaldi and Neri (2013) and Tedeschi and Pisano (2013) use a similar approach to combine the information of the Survey on Household Income and Wealth (SHIW, for short) with that coming from the Household Budget Survey (HBS, thereafter). 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. Clearly, the CIA assumption is particularly unappropriate when the matching relates consumption expenditure and income of households.

Our contribution to the existing literature is twofold. First, we propose a new method to combine information on income and expenditure coming from independent sample surveys. The method goes beyond the CIA assumption. Removing this assumption introduces "intrinsic" *uncertainty* (see Conti *et al.* (2012), Conti *et al.* (2013a), Conti *et al.* (2013b)). Our proposal is to choose a plausible joint distribution for the variables not jointly observed (that is, a *matching distribution*) from a set of equally plausible joint distribution via the Iterative Proportional Fitting (IPF) algorithm using as starting model the quasi-independence model and introducing logical constraints based on extra-sample information to restrict the set of possible choices. A criterion of selecting matching variables by choosing the variables minimizing the maximal error that can occur when combining data available in distinct sample surveys is then discussed.

Secondly we take into account the complexity of 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. Statistical matching in complex sample surveys is studied in Rubin (1986), Renssen (1998). Yet, to the best of our knowledge, previous economic applications fail to consider such an issue. We draw on two surveys of Italian households: SHIW and HBS.

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 the uncertainty analysis in combining household income and expenditure under logical constraints regarding the average propensity to consume 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

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*.

Generally speaking, no joint observation of (Y, Z, X) is available. The main goal of statistical matching, at a macro level, is the estimation of the joint distribution of (Y, Z, X) on the basis of available sample data. The major drawback is that Y, Z, X are not jointly observed, so that, unless special assumptions are made, the statistical model for the joint distribution of (Y, Z, X) is usually *unidentifiable*.

This produces *uncertainty* on the statistical 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 $1, \ldots, N$, and denote by Y, Z, X three characters of interests, 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.* (2013a) 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 $f_x(y,\, z) = Y/Z$.

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

$$\begin{aligned} K^x_{N-}(y,\, z) \;&=\; \max(0,\, G_N(z\,|x) \wedge G_N(\gamma_y(a_x)\,|x) + F_N(y\,|x) \wedge F_N(\delta_z(b_x)\,|x) - 1, \\ &\qquad F_N(y\,|x) + G_N(z\,|x) - 1) \\ K^x_{N+}(y,\, z) \;&=\; \min(G_N(z\,|x), G_N(\gamma_y(a_x)\,|x),\, F_N(y\,|x),\, F_N(\delta_z(b_x)\,|x)) \end{aligned}$$

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

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

$$
\begin{aligned}
\Delta^x(F_N, G_N) &= \frac{1}{N^2 p_N(x)^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(K^x_{N+}(y_i,\, z_j) - K^x_{N-}(y_i,\, z_j) \right) I_{(x_i=x)} I_{(x_j=x)} \\
&= \int_{R^2} \left(K^x_{N+}(y,\, z) - K^x_{N-}(y,\, z) \right) d[F_N(y\,|x) G_N(y\,|x)] \qquad (7)
\end{aligned}
$$

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

$$
\Delta(F_N, G_N) = \sum_x \Delta^x(F_N, G_N) p_N(x). \qquad (8)
$$

Clearly, the unconditional uncertainty measure (8) is the average of the conditional uncertainty measures (7), 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 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

$$
\mathcal{H}^x_N = \{ H_N(y,\, z|\, x) : H_N(y,\, \infty|\, x) = F_N(y|x), H_N(\infty,\, z|\, x) = G_N(z|x), a_x \leqslant f_x(y,\, z) \leqslant b_x \} \quad (9)
$$

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 (7) is, in a sense, a measure of the size of the class (9). If no further information are available, each d.f. in the class

(9) 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 $\mathcal{H}_N^x$ (9), and in using such a d.f. as if it was the "true" p.d.f. $H_N(y,\, z\,|x)$. Such a d.f. $\widetilde{H}_N(y,\, z|x)$ is a *matching distribution* for Y and Z (given X), and plays the role of "surrogate" 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 can neither be directly observed nor estimated on the basis of sample data. The notion of matching error is of basic importance in assessing the quality of the 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

$$\epsilon_N^x(y,\, z) = |\widetilde{H}_N(y,\, z\,|x) - H_N(y,\, z\,|x)| \leqslant K^{+}(y,\, z\,|x) - K^{-}(y,\, z\,|x) \tag{10}$$

so that the overall matching error is given by

$$ME_x(\widetilde{H}_N, H_N) = \int \epsilon_N^x(y,\, z)\, dF_N(y\,|x) dG_N(z\,|x) \leqslant \Delta^x(F_N, G_N). \tag{11}$$

As a consequence, the uncertainty measure (7) 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 the use of a *matching distribution* is. 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 (8).

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.* (2014). 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 (7) 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

$$\widehat{F}_H(y|x) = \frac{\sum_{i=1}^N \frac{D_{i,A}}{\pi_{i,A}} I_{(y_i \leqslant y)} I_{(x_i = x)}}{\sum_{i=1}^N \frac{D_{i,A}}{\pi_{i,A}} I_{(x_i = x)}}, \quad \widehat{G}_H(y|x) = \frac{\sum_{i=1}^N \frac{D_{i,B}}{\pi_{i,B}} I_{(z_i \leqslant z)} I_{(x_i = x)}}{\sum_{i=1}^N \frac{D_{i,B}}{\pi_{i,B}} I_{(x_i = x)}} \tag{12}$$

and then in plugging such estimates in (7). In the sequel, we will denote by $\widehat{\Delta}_H^x$ the estimator of the uncertainty measure $\Delta^x(F_N, G_N)$.

We now turn to the problem of estimating the unconditional uncertainty measure. From the structure of (8), the following estimator can be defined

$$\widehat{\Delta}_H = \sum_{k=1}^{K} \widehat{\Delta}_H^{x^k} \widehat{p}_{H,AB}(x^k) \tag{13}$$

with

$$\widehat{p}_{H,AB}(x^k) = \tau_N^* \widehat{p}_{H,A}(x^k) + (1-\tau_N^*) \widehat{p}_{H,B}(x^k) \tag{14}$$

where $\widehat{p}_{H,A}(x^k)$ and $\widehat{p}_{H,B}(x^k)$ are the Hájek estimators of $p_N(x^k)$ (for $k = 1, \dots, K$) obtained from $\boldsymbol{s}_A$, $\boldsymbol{s}_B$, respectively, and $0 \leqslant \tau_N^* \leqslant 1$. As far as the value of τ_N^* is concerned, details are in (cfr. Conti *et al.* (2014)), where the asymptotic normality of $\widehat{\Delta}_H^x$ and $\widehat{\Delta}_H$ is also proved. 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 $\widehat{\Delta}_H^x$ and $\widehat{\Delta}_H$ 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 (9), with marginal d.f.s $F_N(y|x)$ and $G_N(z|x)$ replaced by their estimators (12) 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^x_{hj} 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 $m^x_{h.}$ and $m^x_{.j}$, *i.e.* the probabilities that Y_d falls into category h and Z_d falls into category j, respectively, can be estimated by

$$\widehat{m}^x_{h.} = \frac{\sum_{i=1}^N \frac{D_{i,A}}{\pi_{i,A}} I_{(y_i=h)} I_{(x_i=x)}}{\sum_{i=1}^N \frac{D_{i,A}}{\pi_{i,A}} I_{(x_i=x)}}, \quad \widehat{m}^x_{.j} = \frac{\sum_{i=1}^N \frac{D_{i,B}}{\pi_{i,B}} I_{(z_j=j)} I_{(x_i=x)}}{\sum_{i=1}^N \frac{D_{i,B}}{\pi_{i,B}} I_{(x_i=x)}} \tag{15}$$

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.

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 (9), is estimated on the basis of available sample data.

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.* (2006). 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 a new criterion based on the unconditional uncertainty measure (13) is proposed to select the matching variables.

The unconditional uncertainty measure is the average of the conditional uncertainty measures (7), 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 variable the one achieving the lowest level of uncertainty, namely the minimum "maximal error" occurring in combining household income and expenditure data. 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 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 (13)) 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 and Pisano (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

$$n_{A,x} = \sum_{i=1}^{N} D_{i,A} I_{(x_i=x)}, \quad n_{B,x} = \sum_{i=1}^{N} D_{i,B} I_{(x_i=x)}, \tag{16}$$

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 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

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 (9). 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 (13) 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.

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

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

Results are in Table 5. 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.

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

a_x-10th percentile	$\widehat{\Delta}^x$-10th percentile	a_x-20th percentile	$\widehat{\Delta}^x$-20th percentile
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

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 (9). 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 an 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. 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 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

$$m_{hj}^x = \delta_{hj} a_h^x b_j^x \tag{17}$$

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

$$\widehat{m}_{hj}^{0,x} = \delta_{hj} \widehat{m}_{h.}^x \widehat{m}_{.j}^x \tag{18}$$

for all $(h,j) \in S^x$. Then IPF proportionally adjusts the values $\widehat{m}^{t,x}_{hj}$ in order to fit the marginals $\widehat{m}^x_{h.}$ and $\widehat{m}^x_{.j}$, respectively, until the desired level of accuracy is achieved. The fitted cells $\widehat{m}^x_{ij}$ 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.

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

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 $n_B\widehat{m}^x_{ij}$.

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 $\widehat{m}^x_{ij}/\sum \widehat{m}^x_{.j}$; (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.

Figure 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+.*

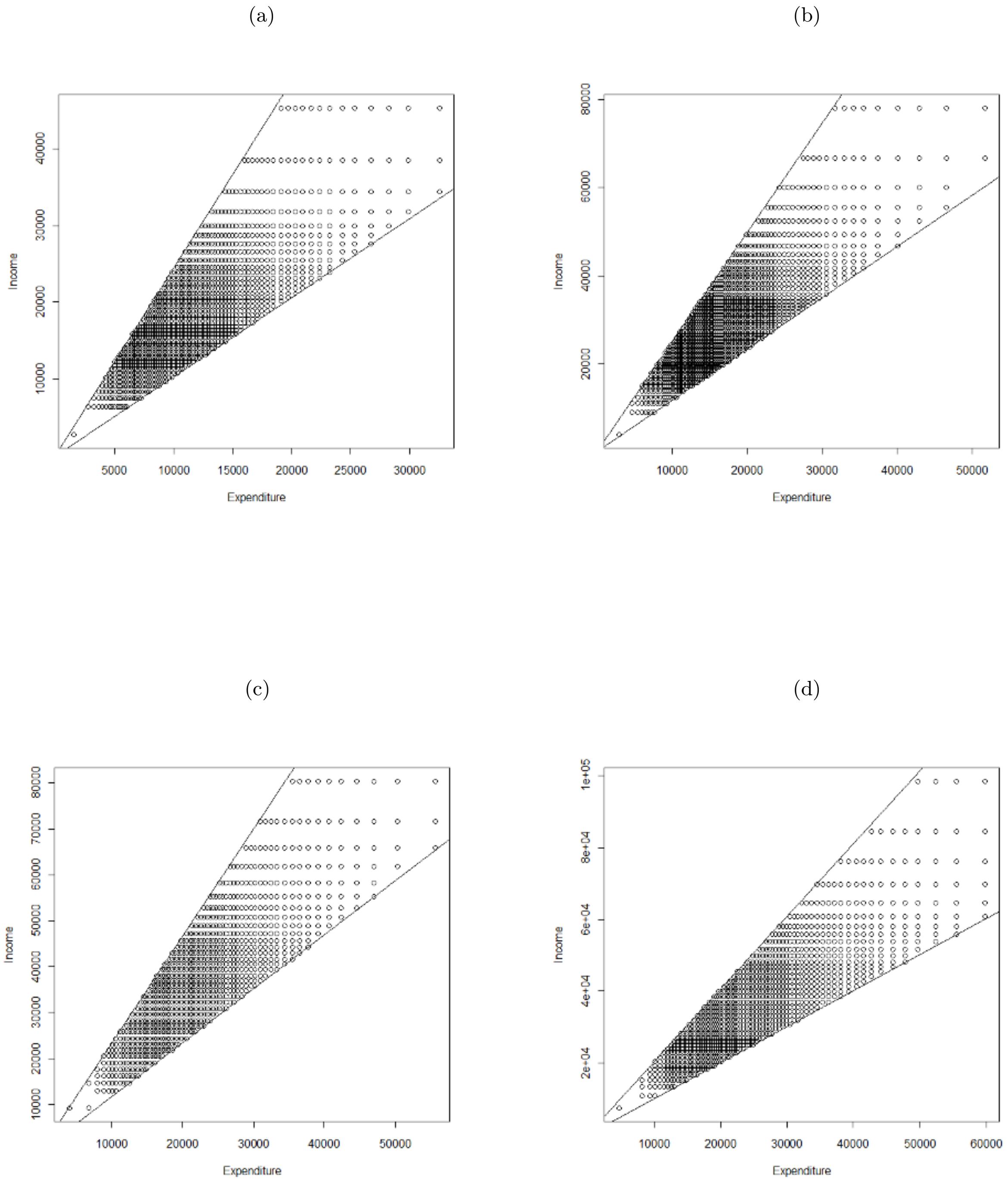
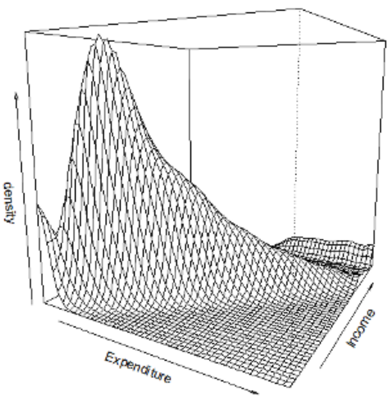
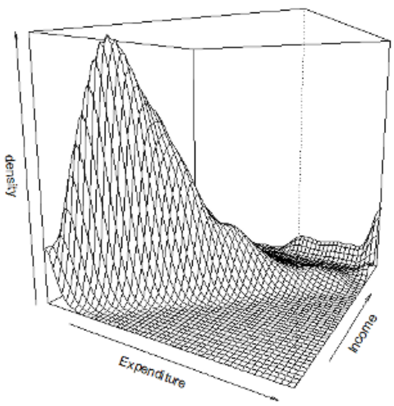


Figure 2: *Bivariate density estimates of matching distributions under the constraints $a_x \leqslant apc \leqslant b_x$ (a) x=1. (b) x=2. (c) x=3. (d) x=4+.*

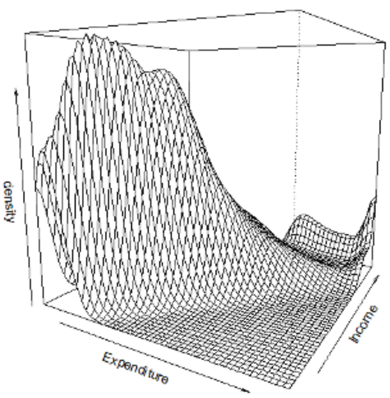
(a)



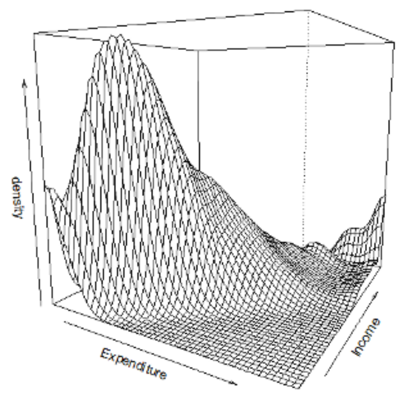
(b)



(c)



(d)



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 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 file between income and expenditure is 0.65, in the "fused" resulting SHIW file 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" file is reported. As expected, the procedure preserves the marginal distribution of expenditure in the "fused" file, 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.

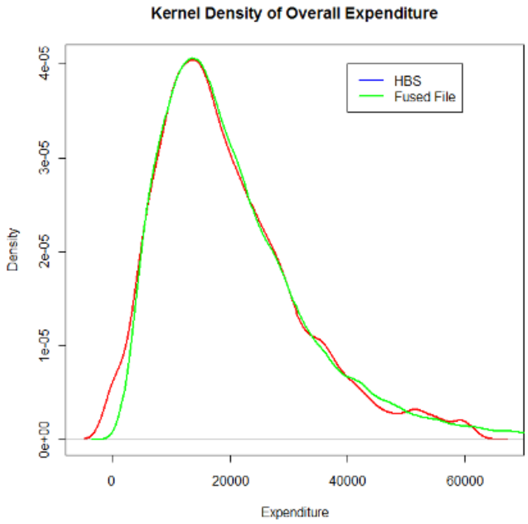
Then, the procedure proposed to choose a matching distribution in the class (9) always respects the fourth level of validity.

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. (ii) to select a *matching distribution* from the class (9) via IPF. Its quality is evaluated via its matching error. Finally, once a *matching distribution* has been estimated, it can been used to impute expenditure microdata in SHIW. This leads to a "reconstructed complete dataset", characterized by an intrinsic matching error. By practitioners, although it

Figure 3: *Kernel density of overall expenditure in HBS and "fused" file*



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 bounds (6) The kind of constraints we consider is $a_x \leqslant f(Y,Z) \leqslant b_x$ given $X = x$, where $f(Y,Z)$ is a monotone function of Y (Z) for each Z (Y).

Let $\gamma_y(\cdot)$ and $\delta_z(\cdot)$ be the inverse functions of $f(Y,Z)$ for fixed y and z, respectively. Without loss of generality, suppose that $f(y,z)$ is an increasing function of y for fixed z and a decreasing function of z for fixed y. Then, we have

$$\begin{aligned} H(y,z|x) &= P(Z \leqslant z, Y \leqslant y\,|x) \\ &= P(Z \leqslant z, Y \leqslant y, f(Y,Z) \leqslant b_x, f(Y,Z) \geqslant a_x|x) \\ &= P(Z \leqslant z, Z \leqslant \gamma_y(a_x), Y \leqslant y, Y \leqslant \delta_z(b_x)|x) \end{aligned}$$

$$
\begin{aligned}
&= P(Z \leqslant (z \wedge \gamma_y(a_x)), Y \leqslant (y \wedge \delta_z(b_x))|x) \\
&= H(z \wedge \gamma_y(a_x), y \wedge \delta_z(b_x)|x) \qquad (19)
\end{aligned}
$$

Hence, the Fréchet bounds become

$$
\begin{aligned}
K_+^x(y,\, z) &= U^x(G(z \wedge \gamma_y(a_x)|x), F(y \wedge \delta_z(b_x)|x)) \\
&= \min(G(z \wedge \gamma_y(a_x)|x), F(y \wedge \delta_z(b_x)|x)) \\
&= \min(G(z|x), G(\gamma_y(a_x)|x), F(y|x), F(\delta_z(b_x)\,|x)) \qquad (20)
\end{aligned}
$$

$$
\begin{aligned}
K_-^x(y,\, z) &= L^x(G(z \wedge \gamma_y(a_x)|x), F(y \wedge \delta_z(b_x)|x)) \\
&= \max(0, G(z \wedge \gamma_y(a_x)|x) + F(y \wedge \delta_z(b_x)|x) - 1) \\
&= \max(0, G(z|x) \wedge G(\gamma_y(a_x)|x) + F(y|x) \wedge F(\delta_z(b_x)|x) - 1). \qquad (21)
\end{aligned}
$$

Proof of Proposition 1 Taking into account that $K_{N+}^x(y,\, z) \leqslant \min(F_N(y|x),\, G_N(z|x))$ and $K_{N-}^x(y,\, z) \geqslant \max(0,\, F_N(y|x) + G_N(z|x) - 1)$, it is not difficult to see that

$$
\begin{aligned}
\Delta^x(F_N, G_N) &\leqslant \int_{R^2} \{\min(F_N(y|x),\, G_N(z|x)) \\
&\quad - \max(0,\, F_N(y|x) + G_N(z|x) - 1)\}\ dF_N(y\,|x) dG_N(z\,|x) \\
&\approx \int_0^1 \int_0^1 \{min(u,\, v) - \max(0,\, u+v-1)\}\ dudv \\
&= \frac{1}{6}. \qquad (22)
\end{aligned}
$$

In other terms, the maximal value of the conditional measure of uncertainty (7) is essentially $1/6 \approx 0.167$. As an easy consequence of Proposition 1, also the unconditional uncertainty measure computed as in (8) takes the value $1/6$.

Proof of Proposition 2 The following two statements hold:

$$\widehat{\Delta}^{*x}_H \xrightarrow{p} \Delta^x(F_N, G_N) \;\; as\; k \to \infty \tag{23}$$

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

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 jth replicate of the population ($j = 1, \dots, k$) 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

$$\frac{1}{k} \sum_{j=1}^k \frac{D^j_{i,A}}{\pi_{i,A}} \tag{25}$$

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, (23) is proved. In the same way, (24) 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.

Banks,J and Oldfield,Z (2007). “Understanding pensions: Cognitive function, numerical ability and retirement saving”, Fiscal Studies, **28**,2,143//170.

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.

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. and 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.

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

Cifaldi, G., 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)(a). “Uncertainty Analysis for statistical matching of ordered categorical variables”. Computational Statistics & Data Analysis, **68**, 311–325.

Conti, P.L., Marella, D., Scanu, M. (2013) (b). “How far from identifiability? A systematic overview of the statistical matching problem in a non-parametric framework”. Communications in Statistics-Theory and Methods. To appear.

Conti, P.L., Marella, D., Scanu, M. (2014). “Uncertainty analysis in statistical matching for complex survey data”. Submitted.

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

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

Jappelli,T.and Pagano, M.(1998). "The Determinants of Savings: Lessons from Italy", CSEF Working Papers,1.

Kennickell,A., 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.

Lusardi,A and Mitchell,O.S (2014). "The Economic Importance of Financial Literacy: Theory and Evidence", Journal of Economic Literature, **52**,1,5–44.

Meyer,B D. and Sullivan,J X. (2003). "Measuring The Well-Being Of The Poor Using Income And Consumption", Journal of Human Resources, **38**,1180–1220.

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.

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, 91102.

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

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

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

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

RECENTLY PUBLISHED "TEMI" (*)

N. 994 – *Trade liberalizations and domestic suppliers: evidence from Chile*, by Andrea Linarello (November 2014).

N. 995 – *Dynasties in professions: the role of rents*, by Sauro Mocetti (November 2014).

N. 996 – *Current account "core-periphery dualism" in the EMU*, by Tatiana Cesaroni and Roberta De Santis (November 2014).

N. 997 – *Macroeconomic effects of simultaneous implementation of reforms after the crisis*, by Andrea Gerali, Alessandro Notarpietro and Massimiliano Pisani (November 2014).

N. 998 – *Changing labour market opportunities for young people in Italy and the role of the family of origin*, by Gabriella Berloffa, Francesca Modena and Paola Villa (January 2015).

N. 999 – *Looking behind mortgage delinquencies*, by Sauro Mocetti and Eliana Viviano (January 2015).

N. 1000 – *Sectoral differences in managers' compensation: insights from a matching model*, by Emanuela Ciapanna, Marco Taboga and Eliana Viviano (January 2015).

N. 1001 – *How does foreign demand activate domestic value added? A comparison among the largest euro-area economies*, by Rita Cappariello and Alberto Felettigh (January 2015).

N. 1002 – *Structural reforms and zero lower bound in a monetary union*, by Andrea Gerali, Alessandro Notarpietro and Massimiliano Pisani (January 2015).

N. 1003 – *You've come a long way, baby. Effects of commuting times on couples' labour supply*, by Francesca Carta and Marta De Philippis (March 2015).

N. 1004 – *Ownership networks and aggregate volatility*, by Lorenzo Burlon (March 2015).

N. 1005 – *Strategy and tactics in public debt manamgement*, by Davide Dottori and Michele Manna (March 2015).

N. 1006 – *Inward foreign direct investment and innovation: evidence from Italian provinces*, by Roberto Antonietti, Raffaello Bronzini and Giulio Cainelli (March 2015).

N. 1007 – *The macroeconomic effects of the sovereign debt crisis in the euro area*, by Stefano Neri and Tiziano Ropele (March 2015).

N. 1008 – *Rethinking the crime reducing effect of education? Mechanisms and evidence from regional divides*, by Ylenia Brilli and Marco Tonello (April 2015).

N. 1009 – *Social capital and the cost of credit: evidence from a crisis*, by Paolo Emilio Mistrulli and Valerio Vacca (April 2015).

N. 1010 – *Every cloud has a silver lining. The sovereign crisis and Italian potential output*, by Andrea Gerali, Alberto Locarno, Alessandro Notarpietro and Massimiliano Pisani (June 2015).

N. 1011 – *Foreign direct investment and firm performance: an empirical analysis of Italian firms*, by Alessandro Borin and Michele Mancini (June 2015).

N. 1012 – *Sovereign debt and reserves with liquidity and productivity crises*, by Flavia Corneli and Emanuele Tarantino (June 2015).

N. 1013 – *Bankruptcy law and bank financing*, by Giacomo Rodano, Nicolas Serrano-Velarde and Emanuele Tarantino (June 2015).

N. 1014 – *Women as 'gold dust': gender diversity in top boards and the performance of Italian banks*, by Silvia Del Prete and Maria Lucia Stefani (June 2015).

N. 1015 – *Inflation, financial conditions and non-standard monetary policy in a monetary union. A model-based evaluation*, by Lorenzo Burlon, Andrea Gerali, Alessandro Notarpietro and Massimiliano Pisani (June 2015).

N. 1016 – *Short term inflation forecasting: the M.E.T.A. approach*, by Giacomo Sbrana, Andrea Silvestrini and Fabrizio Venditti (June 2015).

(*) Requests for copies should be sent to:
Banca d'Italia – Servizio Struttura economica e finanziaria – Divisione Biblioteca e Archivio storico –
Via Nazionale, 91 – 00184 Rome – (fax 0039 06 47922059). They are available on the Internet www.bancaditalia.it.

2012

F. CINGANO and A. ROSOLIA, *People I know: job search and social networks*, Journal of Labor Economics, v. 30, 2, pp. 291-332, **TD No. 600 (September 2006).**

G. GOBBI and R. ZIZZA, *Does the underground economy hold back financial deepening? Evidence from the italian credit market*, Economia Marche, Review of Regional Studies, v. 31, 1, pp. 1-29, **TD No. 646 (November 2006).**

S. MOCETTI, *Educational choices and the selection process before and after compulsory school*, Education Economics, v. 20, 2, pp. 189-209, **TD No. 691 (September 2008).**

P. PINOTTI, M. BIANCHI and P. BUONANNO, *Do immigrants cause crime?*, Journal of the European Economic Association , v. 10, 6, pp. 1318–1347, **TD No. 698 (December 2008).**

M. PERICOLI and M. TABOGA, *Bond risk premia, macroeconomic fundamentals and the exchange rate*, International Review of Economics and Finance, v. 22, 1, pp. 42-65, **TD No. 699 (January 2009).**

F. LIPPI and A. NOBILI, *Oil and the macroeconomy: a quantitative structural analysis,* Journal of European Economic Association, v. 10, 5, pp. 1059-1083, **TD No. 704 (March 2009).**

G. ASCARI and T. ROPELE, *Disinflation in a DSGE perspective: sacrifice ratio or welfare gain ratio?,* Journal of Economic Dynamics and Control, v. 36, 2, pp. 169-182, **TD No. 736 (January 2010).**

S. FEDERICO, *Headquarter intensity and the choice between outsourcing versus integration at home or abroad*, Industrial and Corporate Chang, v. 21, 6, pp. 1337-1358, **TD No. 742 (February 2010).**

I. BUONO and G. LALANNE, *The effect of the Uruguay Round on the intensive and extensive margins of trade,* Journal of International Economics, v. 86, 2, pp. 269-283, **TD No. 743 (February 2010).**

A. BRANDOLINI, S. MAGRI and T. M SMEEDING, *Asset-based measurement of poverty*, In D. J. Besharov and K. A. Couch (eds), Counting the Poor: New Thinking About European Poverty Measures and Lessons for the United States, Oxford and New York: Oxford University Press, **TD No. 755 (March 2010).**

S. GOMES, P. JACQUINOT and M. PISANI, *The EAGLE. A model for policy analysis of macroeconomic interdependence in the euro area,* Economic Modelling, v. 29, 5, pp. 1686-1714, **TD No. 770 (July 2010).**

A. ACCETTURO and G. DE BLASIO, *Policies for local development: an evaluation of Italy's "Patti Territoriali",* Regional Science and Urban Economics, v. 42, 1-2, pp. 15-26, **TD No. 789 (January 2006).**

E. COCOZZA and P. PISELLI, *Testing for east-west contagion in the European banking sector during the financial crisis,* in R. Matoušek; D. Stavárek (eds.), Financial Integration in the European Union, Taylor & Francis, **TD No. 790 (February 2011).**

F. BUSETTI and S. DI SANZO, *Bootstrap LR tests of stationarity, common trends and cointegration,* Journal of Statistical Computation and Simulation, v. 82, 9, pp. 1343-1355, **TD No. 799 (March 2006).**

S. NERI and T. ROPELE, *Imperfect information, real-time data and monetary policy in the Euro area,* The Economic Journal, v. 122, 561, pp. 651-674, **TD No. 802 (March 2011).**

A. ANZUINI and F. FORNARI, *Macroeconomic determinants of carry trade activity,* Review of International Economics, v. 20, 3, pp. 468-488, **TD No. 817 (September 2011).**

M. AFFINITO, *Do interbank customer relationships exist? And how did they function in the crisis? Learning from Italy*, Journal of Banking and Finance, v. 36, 12, pp. 3163-3184, **TD No. 826 (October 2011).**

P. GUERRIERI and F. VERGARA CAFFARELLI, *Trade Openness and International Fragmentation of Production in the European Union: The New Divide?,* Review of International Economics, v. 20, 3, pp. 535-551, **TD No. 855 (February 2012).**

V. DI GIACINTO, G. MICUCCI and P. MONTANARO, *Network effects of public transposrt infrastructure: evidence on Italian regions,* Papers in Regional Science, v. 91, 3, pp. 515-541, **TD No. 869 (July 2012).**

A. FILIPPIN and M. PACCAGNELLA, *Family background, self-confidence and economic outcomes,* Economics of Education Review, v. 31, 5, pp. 824-834, **TD No. 875 (July 2012).**

2013

A. MERCATANTI, *A likelihood-based analysis for relaxing the exclusion restriction in randomized experiments with imperfect compliance*, Australian and New Zealand Journal of Statistics, v. 55, 2, pp. 129-153, **TD No. 683 (August 2008).**

F. CINGANO and P. PINOTTI, *Politicians at work. The private returns and social costs of political connections*, Journal of the European Economic Association, v. 11, 2, pp. 433-465, **TD No. 709 (May 2009).**

F. BUSETTI and J. MARCUCCI, *Comparing forecast accuracy: a Monte Carlo investigation*, International Journal of Forecasting, v. 29, 1, pp. 13-27, **TD No. 723 (September 2009).**

D. DOTTORI, S. I-LING and F. ESTEVAN, *Reshaping the schooling system: The role of immigration*, Journal of Economic Theory, v. 148, 5, pp. 2124-2149, **TD No. 726 (October 2009).**

A. FINICELLI, P. PAGANO and M. SBRACIA, *Ricardian Selection*, Journal of International Economics, v. 89, 1, pp. 96-109, **TD No. 728 (October 2009).**

L. MONTEFORTE and G. MORETTI, *Real-time forecasts of inflation: the role of financial variables*, Journal of Forecasting, v. 32, 1, pp. 51-61, **TD No. 767 (July 2010).**

R. GIORDANO and P. TOMMASINO, *Public-sector efficiency and political culture*, FinanzArchiv, v. 69, 3, pp. 289-316, **TD No. 786 (January 2011).**

E. GAIOTTI, *Credit availablility and investment: lessons from the "Great Recession"*, European Economic Review, v. 59, pp. 212-227, **TD No. 793 (February 2011).**

F. NUCCI and M. RIGGI, *Performance pay and changes in U.S. labor market dynamics*, Journal of Economic Dynamics and Control, v. 37, 12, pp. 2796-2813, **TD No. 800 (March 2011).**

G. CAPPELLETTI, G. GUAZZAROTTI and P. TOMMASINO, *What determines annuity demand at retirement?,* The Geneva Papers on Risk and Insurance – Issues and Practice, pp. 1-26, **TD No. 805 (April 2011).**

A. ACCETTURO e L. INFANTE, *Skills or Culture? An analysis of the decision to work by immigrant women in Italy,* IZA Journal of Migration, v. 2, 2, pp. 1-21, **TD No. 815 (July 2011).**

A. DE SOCIO, *Squeezing liquidity in a "lemons market" or asking liquidity "on tap"*, Journal of Banking and Finance, v. 27, 5, pp. 1340-1358, **TD No. 819 (September 2011).**

S. GOMES, P. JACQUINOT, M. MOHR and M. PISANI, *Structural reforms and macroeconomic performance in the euro area countries: a model-based assessment,* International Finance, v. 16, 1, pp. 23-44, **TD No. 830 (October 2011).**

G. BARONE and G. DE BLASIO, *Electoral rules and voter turnout,* International Review of Law and Economics, v. 36, 1, pp. 25-35, **TD No. 833 (November 2011).**

O. BLANCHARD and M. RIGGI, *Why are the 2000s so different from the 1970s? A structural interpretation of changes in the macroeconomic effects of oil prices*, Journal of the European Economic Association, v. 11, 5, pp. 1032-1052, **TD No. 835 (November 2011).**

R. CRISTADORO and D. MARCONI, *Household savings in China,* in G. Gomel, D. Marconi, I. Musu, B. Quintieri (eds), The Chinese Economy: Recent Trends and Policy Issues, Springer-Verlag, Berlin, **TD No. 838 (November 2011).**

A. ANZUINI, M. J. LOMBARDI and P. PAGANO, *The impact of monetary policy shocks on commodity prices*, International Journal of Central Banking, v. 9, 3, pp. 119-144, **TD No. 851 (February 2012).**

R. GAMBACORTA and M. IANNARIO, *Measuring job satisfaction with CUB models,* Labour, v. 27, 2, pp. 198-224, **TD No. 852 (February 2012).**

G. ASCARI and T. ROPELE, *Disinflation effects in a medium-scale new keynesian model: money supply rule versus interest rate rule,* European Economic Review, v. 61, pp. 77-100, **TD No. 867 (April 2012).**

E. BERETTA and S. DEL PRETE, *Banking consolidation and bank-firm credit relationships: the role of geographical features and relationship characteristics,* Review of Economics and Institutions, v. 4, 3, pp. 1-46, **TD No. 901 (February 2013).**

M. ANDINI, G. DE BLASIO, G. DURANTON and W. STRANGE, *Marshallian labor market pooling: evidence from Italy,* Regional Science and Urban Economics, v. 43, 6, pp.1008-1022, **TD No. 922 (July 2013).**

G. SBRANA and A. SILVESTRINI, *Forecasting aggregate demand: analytical comparison of top-down and bottom-up approaches in a multivariate exponential smoothing framework,* International Journal of Production Economics, v. 146, 1, pp. 185-98, **TD No. 929 (September 2013).**

A. FILIPPIN, C. V, FIORIO and E. VIVIANO, *The effect of tax enforcement on tax morale,* European Journal of Political Economy, v. 32, pp. 320-331, **TD No. 937 (October 2013).**

2014

G. M. TOMAT, *Revisiting poverty and welfare dominance*, Economia pubblica, v. 44, 2, 125-149, **TD No. 651 (December 2007).**

M. TABOGA, *The riskiness of corporate bonds*, Journal of Money, Credit and Banking, v.46, 4, pp. 693-713, **TD No. 730 (October 2009).**

G. MICUCCI and P. ROSSI, *Il ruolo delle tecnologie di prestito nella ristrutturazione dei debiti delle imprese in crisi*, in A. Zazzaro (a cura di), Le banche e il credito alle imprese durante la crisi, Bologna, Il Mulino, **TD No. 763 (June 2010).**

F. D'AMURI, *Gli effetti della legge 133/2008 sulle assenze per malattia nel settore pubblico,* Rivista di politica economica, v. 105, 1, pp. 301-321, **TD No. 787 (January 2011).**

R. BRONZINI and E. IACHINI, *Are incentives for R&D effective? Evidence from a regression discontinuity approach,* American Economic Journal : Economic Policy, v. 6, 4, pp. 100-134, **TD No. 791 (February 2011).**

P. ANGELINI, S. NERI and F. PANETTA, *The interaction between capital requirements and monetary policy*, Journal of Money, Credit and Banking, v. 46, 6, pp. 1073-1112, **TD No. 801 (March 2011).**

M. BRAGA, M. PACCAGNELLA and M. PELLIZZARI, *Evaluating students' evaluations of professors,* Economics of Education Review, v. 41, pp. 71-88, **TD No. 825 (October 2011).**

M. FRANCESE and R. MARZIA, *Is there Room for containing healthcare costs? An analysis of regional spending differentials in Italy,* The European Journal of Health Economics, v. 15, 2, pp. 117-132, **TD No. 828 (October 2011).**

L. GAMBACORTA and P. E. MISTRULLI, *Bank heterogeneity and interest rate setting: what lessons have we learned since Lehman Brothers?,* Journal of Money, Credit and Banking, v. 46, 4, pp. 753-778, **TD No. 829 (October 2011).**

M. PERICOLI, *Real term structure and inflation compensation in the euro area*, International Journal of Central Banking, v. 10, 1, pp. 1-42, **TD No. 841 (January 2012).**

E. GENNARI and G. MESSINA, *How sticky are local expenditures in Italy? Assessing the relevance of the flypaper effect through municipal data,* International Tax and Public Finance, v. 21, 2, pp. 324-344, **TD No. 844 (January 2012).**

V. DI GACINTO, M. GOMELLINI, G. MICUCCI and M. PAGNINI, *Mapping local productivity advantages in Italy: industrial districts, cities or both?*, Journal of Economic Geography, v. 14, pp. 365–394, **TD No. 850 (January 2012).**

A. ACCETTURO, F. MANARESI, S. MOCETTI and E. OLIVIERI, *Don't Stand so close to me: the urban impact of immigration,* Regional Science and Urban Economics, v. 45, pp. 45-56, **TD No. 866 (April 2012).**

M. PORQUEDDU and F. VENDITTI, *Do food commodity prices have asymmetric effects on euro area inflation,* Studies in Nonlinear Dynamics and Econometrics, v. 18, 4, pp. 419-443, **TD No. 878 (September 2012).**

S. FEDERICO, *Industry dynamics and competition from low-wage countries: evidence on Italy*, Oxford Bulletin of Economics and Statistics, v. 76, 3, pp. 389-410, **TD No. 879 (September 2012).**

F. D'AMURI and G. PERI, *Immigration, jobs and employment protection: evidence from Europe before and during the Great Recession,* Journal of the European Economic Association, v. 12, 2, pp. 432-464, **TD No. 886 (October 2012).**

M. TABOGA, *What is a prime bank? A euribor-OIS spread perspective,* International Finance, v. 17, 1, pp. 51-75, **TD No. 895 (January 2013).**

L. GAMBACORTA and F. M. SIGNORETTI, *Should monetary policy lean against the wind? An analysis based on a DSGE model with banking,* Journal of Economic Dynamics and Control, v. 43, pp. 146-74, **TD No. 921 (July 2013).**

M. BARIGOZZI, CONTI A.M. and M. LUCIANI, *Do euro area countries respond asymmetrically to the common monetary policy?,* Oxford Bulletin of Economics and Statistics, v. 76, 5, pp. 693-714, **TD No. 923 (July 2013).**

U. ALBERTAZZI and M. BOTTERO, *Foreign bank lending: evidence from the global financial crisis,* Journal of International Economics, v. 92, 1, pp. 22-35, **TD No. 926 (July 2013).**

R. DE BONIS and A. SILVESTRINI, *The Italian financial cycle: 1861-2011,* Cliometrica, v.8, 3, pp. 301-334, **TD No. 936 (October 2013).**

D. PIANESELLI and A. ZAGHINI, *The cost of firms' debt financing and the global financial crisis,* Finance Research Letters, v. 11, 2, pp. 74-83, **TD No. 950 (February 2014).**

A. ZAGHINI, *Bank bonds: size, systemic relevance and the sovereign,* International Finance, v. 17, 2, pp. 161-183, **TD No. 966 (July 2014).**

S. MAGRI, *Does issuing equity help R&D activity? Evidence from unlisted Italian high-tech manufacturing firms,* Economics of Innovation and New Technology, v. 23, 8, pp. 825-854, **TD No. 978 (October 2014).**

G. BARONE and S. MOCETTI, *Natural disasters, growth and institutions: a tale of two earthquakes,* Journal of Urban Economics, v. 84, pp. 52-66, **TD No. 949 (January 2014).**

2015

G. BULLIGAN, M. MARCELLINO and F. VENDITTI, *Forecasting economic activity with targeted predictors,* International Journal of Forecasting, v. 31, 1, pp. 188-206, **TD No. 847 (February 2012).**

A. CIARLONE, *House price cycles in emerging economies,* Studies in Economics and Finance, v. 32, 1, **TD No. 863 (May 2012).**

G. BARONE and G. NARCISO, *Organized crime and business subsidies: Where does the money go?,* Journal of Urban Economics, v. 86, pp. 98-110, **TD No. 916 (June 2013).**

P. ALESSANDRI and B. NELSON, *Simple banking: profitability and the yield curve,* Journal of Money, Credit and Banking, v. 47, 1, pp. 143-175, **TD No. 945 (January 2014).**

R. AABERGE and A. BRANDOLINI, *Multidimensional poverty and inequality,* in A. B. Atkinson and F. Bourguignon (eds.), Handbook of Income Distribution, Volume 2A, Amsterdam, Elsevier, **TD No. 976 (October 2014).**

M. FRATZSCHER, D. RIMEC, L. SARNOB and G. ZINNA, *The scapegoat theory of exchange rates: the first tests,* Journal of Monetary Economics, v. 70, 1, pp. 1-21, **TD No. 991 (November 2014).**

FORTHCOMING

M. BUGAMELLI, S. FABIANI and E. SETTE, *The age of the dragon: the effect of imports from China on firm-level prices*, Journal of Money, Credit and Banking, **TD No. 737 (January 2010).**

G. DE BLASIO, D. FANTINO and G. PELLEGRINI, *Evaluating the impact of innovation incentives: evidence from an unexpected shortage of funds,* Industrial and Corporate Change, **TD No. 792 (February 2011).**

A. DI CESARE, A. P. STORK and C. DE VRIES, *Risk measures for autocorrelated hedge fund returns,* Journal of Financial Econometrics, **TD No. 831 (October 2011).**

D. FANTINO, A. MORI and D. SCALISE, *Collaboration between firms and universities in Italy: the role of a firm's proximity to top-rated departments,* Rivista Italiana degli economisti, **TD No. 884 (October 2012).**

M. MARCELLINO, M. PORQUEDDU and F. VENDITTI, *Short-Term GDP Forecasting with a mixed frequency dynamic factor model with stochastic volatility,* Journal of Business & Economic Statistics, **TD No. 896 (January 2013).**

M. ANDINI and G. DE BLASIO, *Local development that money cannot buy: Italy's Contratti di Programma,* Journal of Economic Geography, **TD No. 915 (June 2013).**

J. LI and G. ZINNA, *On bank credit risk: sytemic or bank-specific? Evidence from the US and UK,* Journal of Financial and Quantitative Analysis, **TD No. 951 (February 2015).**