Session Number: Parallel Session 2B
Time: Monday, August 23, PM

Paper Prepared for the 31st General Conference of
The International Association for Research in Income and Wealth

St. Gallen, Switzerland, August 22-28, 2010

Combining Expenditure Surveys and Tax Registers: the Case of Milan Municipality

Lisa Crosato, Paolo Mariani, Mauro Mussini and Biancamaria Zavanella

For additional information please contact:

Name: Lisa Crosato
Affiliation: Università di Milano-Bicocca

Email Address: lisa.crosato@unimib.it

Combining expenditure surveys and tax registers: the case of Milan municipality.

Lisa Crosato, Paolo Mariani, Mauro Mussini and Biancamaria Zavanella[1]

Preliminary version, please do not cite.

Abstract

In this paper we combine two data sources: the sample survey on family expenditure conducted by the Milan Municipality and the Chamber of Commerce of Milan (wave 2007-2008) and the tax register matched to the local population and family register in the data-warehouse AMeRIcA, regarding residents in Milan in 2007.[2] We mainly focus on data combination and run a few matching exercises following Jenkins et al (2008) approach. Using both information on individuals and households we select four variables (Census Section, Type of Family and two compounded variables for age and sex) and combine them into four matching criteria applied in a first stage independently and in a second one hierarchically. Results are encouraging since we match more than 70% of the survey households through the automated matching exercise.

Keywords: Exact matching, administrative data, income, tax records.

[1] Corresponding author: Lisa Crosato, Statistics Department Università di Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, 20126 Milano. email: lisa.crosato@unimib.it.

[2] We thank Milan Municipality (Statistics Office) and the Chamber of Commerce of Milan for providing us with the data. We are particularly grateful to Lorena Scarcello and Flavio Necchi for their support to our project.

1 Introduction

Administrative data have been lately exploited not only for the study of income distribution (Atkinson, 2007), but also for data validation and evaluation of error measure in sample survey (Jäckle et al. 2004, Dragoset and Fields 2006).

In this paper we combine two data sources containing, among others, the main monetary proxies of well-being: income and consumption. Both sources, a sample survey and an administrative register, regard the municipality of Milan (Italy). Although international literature clearly points to this direction (Jones & Elias 2006, Jenkins 2008), to our knowledge this is one of the first attempts of matching survey and administrative data in Italy (see ISTAT 2008).

The first dataset is the sample survey on family expenditure[3] (ICFM hereafter) conducted by the Milan Municipality and the Chamber of Commerce of Milan (wave 2007-2008). This survey is directed to a sample of about 800 families resident in the Milan Municipality and follows, both in sample selection and survey composition, ISTAT (Italian National Statistical Office) Household Expenditure Survey. The aim of the survey is therefore the estimation of the expenditures of Milan families in real goods and services. The second dataset collects administrative records regarding the entire population resident in Milan and covers a 8 year period, from 2000 to 2007. The data derive from tax records matched to the local population and family register in the data-warehouse AMeRIcA[4] (Mariani, Mezzanzanica e Zavanella, 2006 a and b), which contains more than 800,000 individual records per year. Each individual is characterized by yearly income as appearing in income tax return, divided according to income source (salary income, including pensions and subsides, private income, estate income and income from other sources such as capital gains or rental income.) The integration with the population register represents one of the main advantages of AMeRIcA, since it allows reconstruction of families and implies the total coverage of the population either considering individuals or families. Furthermore, several socio-demographic characteristics are recorded for tax-payers and not. This makes it possible also a thorough characterization of people in principle exposed to poverty, each year and over time (Crosato e Zavanella, 2010; Minotti, Mussini e Zavanella, 2010).

This paper concerns the linkage between the above sample survey and the administrative register. We mainly focus on data combination, referring to the literature which offers several tools (see Belin and Rubin, 1995, Gill, 2001, Ridder e Moffit, 2007, Jenkins et al. 2008). Not disposing of a unique identifier for either households or individuals, we apply a relaxed version of the exact matching method (defined almost exact matching by Gill, 2001) considering as linked two entities showing the same value for a given number of variables. In principle we could try a match both within individuals and within households, but after a few unsuccessful attempts we resolve to integrate information on both units. Disposing of variables describing both entities in either dataset, we perform our matching between the survey households with "administrative" families also recurring to artificial variables for households constructed combining individual characteristics of their components, such as age and sex. Following Jenkins et al. (2008), two compounded variables, Age Code and Sex Code together with the Census Section and the Type of Family are combined in four criteria

[3] "Indagine sui Consumi delle Famiglie nel comune di Milano"

[4] AMeRIcA ("Anagrafe Milanese e Redditi Individuali con Archivio") is managed by the Statistics Department of Milan-Bicocca University on the behalf of Milan Municipality.

in order to test their discriminating power and reliability through several matching exercises.

The paper is structured as follows. Section 2 describes the data sources, in section 3 we discuss possible matching methods and delineate our strategy, we list and discuss matching variables (3.1) and finally perform the matching exercises (3.2 and 3.3). Section 4 concludes, summarizing main results, and try to give a few insights on the research issues raised by the new dataset characterizing families and individuals on both the expenditure and the earning side.

2 Data description

Since the two data sources to be combined are of a local character and not well known, in the following we try a brief but thorough description of both of them to outline their main features and highlight their potentiality for the linkage project.

The ICFM survey

The survey on consumptions of households living in Milan carried out in the 2007-2008 period represents the survey data source that we used in the linkage experiment. This survey, started in 2005, it is carried on with annual periodicity. The ICFM survey is based on a stratified sampling in which the households are the sampling units and the stratification variables are the municipality sub-areas, the number of household members, and the age of the householder. The questionnaire is very detailed, asking respondent to provide information on various forms of non-durable and durable consumption goods. In addition, the ICFM survey contains information on the properties lived in or owned by the household and the characteristics of each family member. The questionnaire includes also a question about the household net income level asking respondent to state their income class.

The collection period of the ICFM survey in 2007-2008 took place from March 2007 till February 2008. The households interviewed were 808, including 2,403 individuals. A unique identifier was assigned to each household. However, it cannot be used for linkage purposes as it is a survey specific coding which is unrelated to any coding defined in other survey or administrative data (for example, the unique identifier assigned to each household in AMeRIcA). Table 1 classifies the sample of respondents in 2007-2008 according to type of family.

Table 1: Sample of households in ICFM survey divided by Type of Family.

Type of Family	frequencies	Percentage
Single	123	15.2
Single parent with children	68	8.4
Couple with no children	183	22.6
Couple with children	434	53.7
All	808	100

The AMeRIcA data-warehouse

Administrative records used for the linkage derive from AMeRIcA project which provides demographic and income information concerning individuals and households resident in the Milanese area. The structure of AMeRIcA is based on a Data Warehouse that combines administrative data records from the tax register of the Milan Revenue Agency with the Milan Population Register. Therefore, AMeRIcA is a combined administrative data source and it represents the first experience in Italy of linkage of administrative records belonging to different administrative data sources.

The Revenue Agency registers the Italian Personal Income Tax (IRPEF) in categories: pre-tax income by source (pre-tax income is defined as the sum of incomes from dependent employment, self-employment, company income, rental income, agricultural income, real estate income, and other incomes), taxable income, gross IRPEF, net IRPEF, tax allowance, detraction for dependent family members and other deductible costs. The archives of the Registry provide information about marital status, citizenship, address, family composition, residence and other personal data (gender, date of birth) for each registered individual. The combination of the two administrative sources enables AMeRIcA to render information on pre-tax income, income tax paid, tax allowance and detractions plus family size and composition of residents in the Milanese area.

The AMeRIcA pilot project begun in 2000 on the basis of the rising interest for the use of administrative data for statistical purposes, and now this Data Warehouse covers the 2000-2007 period. In our study, we focus on household records available for 2007 and their components. We refer to a population composed of 653,686 households.

Table 2: Population of households resident in Milan divided by household type.

Type of Family	Frequencies	Percentage
Single	317,880	48.63
Single parent with children	63,201	9.67
Couple with no children	101,472	15.52
Couple with children	136,098	20.82
other	35,034	5.36
Total	653,685	100

3 Data Linkage

Literature on data combination offers mainly three tools for linkage: statistical matching, probabilistic matching and exact (or deterministic) matching (see Jabine and Scheuren, 1986; Gill, 2001; Ridder e Moffit, 2007). Statistical matching is advisable if the fraction of units that are in both data sources is small, so that we can treat the two samples as independent with negligible intersection. In our case exact or probabilistic matching are the most suitable methodologies, since not only a large part of the units are common to both data sources, but households belonging to the survey are sampled from the administrative source. Therefore the first dataset is, on principle, included in the second. A possible source of violation of the above statement could be mainly information recorded at different times, such as births or deaths after the interview for the survey has been made or residential mobility acknowledged only by the population register. We will go back to this point later on, assuming in the following all units in the ICFM survey to be included in AMeRIcA.

Now, technically speaking, exact matching is feasible when both datasets contain the same variable or characteristic available for all units, fixed, easily recordable, verifiable and unique to that individual (Gill, 2001). When available, this variable is usually identified with some unique identification number assigned to individuals at birth such as the National Health Service Number or, in Italy, the fiscal code, an alphanumerical code composed by letters and numbers corresponding to name, sex, date and place of birth. An important (and isolated) precedent of data matching regarding income and consumption surveys was conducted by ISTAT (2009) using exact matching with fiscal code. ISTAT (the Italian National Institute of Statistics) matches several Eu-Silc waves with the Italian tax register in order to assess measurement error in Eu-Silc regarding Italy and to retrieve missing information from the tax register. ISTAT had the opportunity to access microdata on tax records for the whole Country and to assign fiscal codes to individuals present in both datasets. On the contrary, for privacy reasons, we do not possess either the fiscal code or the information useful to retrieve it in neither data source.

When the data sources lack a unique identification variable, the alternative is to create artificially a compound key starting from a few characteristics which, jointly considered, form a sort of identifying code of each unit (Gill, 2001). In other words, the records under scrutiny are compared simultaneously on several variables, their different combinations constituting as many criteria to proceed with (Jenkins et al. 2008). Due to this highly characterizing matching key, exact matching should in principle return one-or-none type links, leaving no space for the so called possible links.

On the contrary, probabilistic matching is based on the premise that the datasets involved in the matching are characterized by errors and omissions and there is no common unique identifier. This method explicitly provides for possible (or one-to-many) links, as well as for positive links and negative links[5], established after comparison of several variables and related probability scores (Fellegi and Sunter, 1968). The tricky side of probabilistic record linkage is related mainly to the estimation of probability scores and thresholds in order to determine whether a given couple of records would reasonably belong to the same unit. This, in turn, requires prior information on both the distribution of the matching variables in the original population and on the probability of errors, such as coding and transcription errors or omissions characterizing the generating process of the records under comparison.

In our case, strictly speaking, exact matching would not be feasible due to the absence of a unique identifier. In fact, the ICFM survey households are characterized by an internal identifying code and individuals only by their family code, whilst in AMeRIcA individuals have a secreted fiscal code and households an identifying number which remain the same over years. On the other hand, turning to probabilistic matching would complicate matters, even if we possess all the information needed on the distribution of matching variables in the population (America) because we lack information on errors.

Since we could not identify a set of variables that, compounded, substitute the unique identifier, we resort to almost-exact matching as defined by Gill (2001) by relaxing the exact match criterion a little. We use then the number of variables that agree (at least three) to establish if the record pair should be linked, controlling for disagreement within the remaining variables by clerical inspection. In practice, we closely follow Jenkins et al. (2008) procedure also to have a basis for comparison. It should be stressed, however, that differently from the above authors, we maintain throughout the paper the "possible links" category among our outcomes.

[5] In the following, we use indifferently the terms positive, negative and possible links or, respectively, match, non-match and possible match.

3.1 Selecting matching variables and matching criteria.

Before proceeding further with the linkage it is worthwhile to devote a few lines to the choice of matching variables and matching units in order to clarify the pattern we followed.

The variables both (potentially) common and unique to the ICFM survey and the data-warehouse AMeRIcA are reported in Table 3. The first kind of variables represent candidate matching variables, where candidate refers to the way the same variables are defined and/or recorded. Among these, we left apart those related to income since they originate from completely different processes: income, as well as the number of income receivers, reported in the survey derives from a specific question whereas in AMeRIcA the same variables derive from tax records. These and the second kind of variables represent the informative added value which either dataset brings through the matching process, the first ones for comparison and possible correction of item non response and the second ones for enlarging the informative set characterizing households and individuals. Other variables potentially useful in data matching, but differently recorded are the characteristics of householders, which in the survey are collected or double-checked during the interview and in AMeRIcA are registered through administrative procedures.

Table 3 is also divided in two parts according to the candidate matching units, since either dataset supplies a number of characteristics for both households and individuals. As a consequence, out first approach was to process separately individuals and households matching, proceeding to the linkage within individuals (or households) on the basis of their proper characteristics, but this unfortunately resulted in not a single positive link or negative link, only possible links. Starting from households, we tried a matching using Census Section, Number of Components and Number of Children, after blocking according to the Type of Family. Of course, the Number of Children was of no help in identifying families with no children and singles, but in any case, also for large (with 6 or 7 components) families, results were very poor. It went even worse with our attempt to link individuals according to Sex, Age and Census Section retrieved from the family. We then realized that with no names and addresses our matching would be a mission impossible.

Therefore, in order to augment the quality of our matching variables, we have decided to fully exploit our data sources and in particular to integrate matching units constructing two artificial variables relating households and individuals. Each household was assigned two vectors:

- an Age Code reporting ages of all family components in increasing order;
- a Sex Code given by the sex of all family components ranked according to increasing age.

Of course, these artificial variables share potential reporting errors in the number of components and the family sex code is affected by errors in age recording when implying reverting components' age ranking. We devoted particular attention to these matters during the final stage of clerical inspection, but since adding these compounded variables led to enormously improved matching rates, we have kept them. Apparently, the Age Code and Sex Code possess a large discriminating power but may present some problems with reliability, since they can suffer from an accumulation of errors.

Table 3: comparison of variables

Matching units: households	
ICFM survey	**AMeRIcA**
Postal code (CAP: 38 codes) Census Section	Address Functional Area (180 areas) Census section (6,036 sections)
Number of components	Number of components
Number of children	Number of children
Type of family (1=single, 2=couple with children, 3=couple without children, 4=single parent)	Type of family (1=single, 2=couple with children, 3=couple without children, 4=single parent, 5=other)
Number of income receivers	Number of income receivers
Professional condition of the householder	
Number of pension recipients	
Monthly consumption	
Income class	Taxable Income Net Income Taxes Pensions and subsides
	Number of Italians
	Income source
Matching units: individuals	
Sex	Sex
Year of birth (age as 2007-year of birth)	Age
Reference person (Householder)	Householder
Relationship of each family component to the householder	
Education level	
Working position	
	Taxable Income Net Income Taxes Pensions and subsides
	Income source
plus all the variables relating the family they belong to.	

The other two variables available on households were Type of Family and Census Section. The first one might cause some mismatch due to the category “other”, which is present in AMeRIcA but not in the ICFM survey. On the contrary, Census Section should be measured in the same way in both datasets (being assigned and not requested in the survey) and subject, at most, to coding or reporting errors and to residential mobility discrepancies. In addition, Census Sections represent the finest territorial grid of Milan municipality with 6,036 sections. For these reasons we expect it to show both high reliability and discriminating power.

To conclude, we have singled out four variables suitable for our matching exercise on households and precisely Census Section, Type of family, Age Code and Sex Code. Following Jenkins et al. (2008), we organize the selected variables in four criteria, each of them excluding one variable at a time. The main advantage of this procedure is that, as we do not dispose of a unique identifying code, using the single variables rarely leads to match any record, while using compound keys (or criteria) allows for some success in matching and at the same time the exclusion of one variable in turn allows for assessment of their discriminating power.

The four criteria we have used to link[6] households from the ICFM survey to families from the data-warehouse AMeRIcA are precisely:

Criterion 1: Census Section, Age Code and Type of Family;
Criterion 2: Census Section, Sex Code and Type of Family;
Criterion 3: Type of Family, Sex Code and Age Code;
Criterion 4: Census Section, Sex Code and Age Code.

We first run the four linkage exercises independently and secondly combine the criteria in a hierarchical matching process. We decided to apply hierarchical matching separately for possible links and negative links since the size of "possible matches" grey zone varies a lot according to the selected criterion. This makes hierarchical matching even more relevant. The estimation of false-positive rates and false-negative rates along with clerical inspection of non-matches to assess accuracy conclude our linkage exercise.

3.2 Linkage rates

Linkage rates resulting from the four match criteria and hierarchical matching are reported in table 4. We do not expect to incur in true non-matching issues since the administrative register we use covers the whole population, so that all respondents participating in the ICFM survey should be present in AMeRIcA. The fact that part of the interviews in the ICFM survey were run in 2008 should not be a problem since households were selected among residents in Milan in 2007. There could be a few households present in the ICFM survey and not in AMeRIcA due to residential mobility, but this would happen only in the case that a household was interviewed early in 2007 and then moved to another municipality, disappearing from the Milan population register. Therefore in our study the expected true non-matching is near zero; talking about non matching, we mainly refer to false non matches due to errors, omissions or different dates in registering the same variable in the two datasets.

We first comment linkage rates as resulting from the automated matching, leaving clerical inspection on single cases to a second moment in order to separately highlight the discriminating power of the variables as such and their reliability and, therefore, the extent to which they lead to mismatches or false non-matches.

Table 4 reports matching results both in absolute values and as a fraction of the ICFM survey. The table is divided into three main panels. The top one refers to matching according to single criteria and to the pooled linkage obtained using at least one of the criteria. The central panel reports matching results obtained applying a given criterion to the possible matches singled out by another criterion and the bottom panel reports further matches resulting from application of a given criterion to the non-matches produced by a different criterion.

Among the four independent criteria, the first (Census Section, Age Code and Type of Family) and the fourth (Census Section, Sex Code and Age Code) return matching rates remarkably larger than the remaining two. In particular, the best combination of variables results Census Section, Sex Code and Age Code with a matching rate of 69.6% while Census Section, Age Code and Type of Family matches 63% of the

[6] We have implemented all linkage procedures through the R software for statistical computing (with particular reference to "merge" function).

records, both expressed as a fraction of the ICFM survey. On the one hand, this suggests that Census Section and Age Code are the outstanding variables in our matching exercise. Comparing the column of possible linkage rates of criteria 2 and 3 confirms the noticeable discriminating power of Census Section[7] versus Age Code since lack of the first leads to a largest ratio between the number of records in AMeRIcA to be possibly matched to ICFM survey records and the number of the corresponding ICFM records (more than 700 to 1 for criterion 3 vs almost 16 to 1 for criterion 2). On the other hand, comparison between possible match and non-match rates regarding criteria 1 and 4 highlights that Type of Family leads to more non-matches (208) with respect to Sex Code (189) when coupled to Census Section and Age Code. This could be attributed both to the lower reliability of Type of Family and to the dependence of Sex Code on Age Code which could invalidate the power of the former in correcting negative links identified by Age Code. We will further analyze this aspect through clerical inspection.

The pooled matching rate achieved applying at least one criterion (76.1%) improves the linking rate of the best criterion by 6.5 points and limits non matching cases to 44 (5.5%). Possible matches resulting from the pooling exercise are 149. We expect these households to be characterized by a small number of components (3 or smaller) which makes our compound variables, Age Code or Sex Code, less effective or with a very low discriminating power among Type of Family "single". For singles in fact both compounded variables collapse to Age and Sex. Note that applying the pooled matching exercise, in our case, is equivalent to count as a match any paired couple for which at least three variables coincide. Thus, pooling results from several criteria could in principle lead to the same record matched to different ones according to different criteria. This happened only for six records in the ICFM survey which were matched to two different records in AMeRIcA, but for all six of them the right match was easily determined. On the contrary, hierarchical matching does not run this risk since the second criterion is applied only to cases left unsolved by the first one (possible links) or classified by the first criterion as a non-match.

The central panel of table 4 reports results of hierarchical matching on possible matches combining two criteria in turn (for instance, C1+C2 means that criterion 2 was applied to possible matches outlined by criterion 1). Rates corresponding to application of any criterion after criterion 4 are not reported because there was no further match to be counted. On the contrary, we can see that in the remaining cases applying a second criterion solves many uncertainties, leading to better matching rates especially for criteria 2 and 3. The worst rates are obtained, as expected, from combination of criteria 2 and 3 independent of their application order, confirming their bad discriminating power. The best rate is achieved by running criteria 2 or 4 after criterion 1. Still, the 67% of matched records does not equal the 69.6% reached by criterion 4 alone and improves the previous matching rate of criterion 2 by 4% only (corresponding to about 35% of multiple matches transformed in matches). We can state that the contribution of

[7] It is important to stress the role of Census Section in the matching exercise since it is the only "address" variable common to both datasets. In the beginning we were not given any such variable for the ICFM survey. Earlier matching attempts were run without address information, then we obtain the postal code for the ICFM and derive it from address in America (very time consuming job) and the matching started working. Only in the end we have obtained census sections, so we thank the Chamber of commerce and the Municipality of Milan about this.

hierarchical matching greatly varies according to the goodness of the first criterion applied, and does not always help to solve uncertainties.

Table 4: matching rates for ICFM households

	matches		possible matches		non-matches		all	
Independent matching	n	%	n *(to N)*	%	n	%	n	%
C1: Census Section, Age Code and Type of Family	509	63.0	91 *(254)*	11.3	208	25.7	808	100
C2: Census Section, Sex Code and Type of Family	138	17.1	584 *(9,299)*	72.3	86	10.6	808	100
C3: Type of Family, Sex Code and Age Code	204	25.3	476 *(338,624)*	58.9	128	15.8	808	100
C4: Census Section, Sex Code and Age Code	562	69.6	57 *(143)*	7.1	189	23.3	808	100
Pooled matching: at least one of the above	615	76.1	149	18.4	44	5.5	808	100
*Hierarchical matching on possible matches**								
C1+C2 or C1+C4	541	67.0	57 *(143)*	7.1	210	26.0	808	100
C1+C3	526	65.1	73 *(307)*	9.0	209	25.9	808	100
C2+C1 or C2+C4	535	66.2	67 *(166)*	8.3	206	25.5	808	100
C2+C3	390	48.3	247 *(6,049)*	30.6	171	21.2	808	100
C3+C1	468	57.9	124 *(494)*	15.3	216	26.7	808	100
C3+C2	369	45.7	259 *(6,149)*	32.1	180	22.3	808	100
C3+C4	473	58.5	114 *(428)*	14.1	221	27.4	808	100
*Hierarchical matching on non matches***								
C1+C2	567	70.2	169 *(1,230)*	20.9	72	8.9	808	100
C1+C3	537	66.5	159 *(23,105)*	19.7	112	13.9	808	100
C1+C4	584	72.3	57 *(143)*	7.1	167	20.7	808	100
C2+C1	551	68.2	67 *(166)*	8.3	190	23.5	808	100
C2+C3	393	48.6	252 *(6,070)*	31.2	163	20.2	808	100
C2+C4	553	68.4	67 *(166)*	8.3	188	23.3	808	100
C3+C1	485	60.0	124 *(494)*	15.3	199	24.6	808	100
C3+C2	387	47.9	291 *(6,314)*	36.0	130	16.1	808	100
C3+C4	498	61.6	114 *(428)*	14.1	196	24.3	808	100
C4+C1	584	72.3	59 *(147)*	7.3	165	20.4	808	100
C4+C2	585	72.4	155 *(1,272)*	19.2	68	8.4	808	100
C4+C3	571	70.7	134 *(37,233)*	16.6	103	12.7	808	100

Notes: matching rates are calculated as a proportion of the ICFM survey (n=808). The column possible matches reports in brackets the number of records in AMeRIcA *(N)* possibly matching the n records in the ICFM survey. In hierarchical matching, criteria are applied in order of appearance.

* In case of possible matches the second step of the matching is confined to comparison, according to the second criterion, between the n records in the ICFM survey and the N ones in AMeRIcA, otherwise we could match the same record in AMeRIcA with two different records in the ICFM survey.

**results include the matches gained through hierarchical matching on possible matches.

The case of hierarchical matching applied to non-matches (table 4, bottom panel) points to slightly different conclusions, since all linking rates corresponding to the four independent criteria are improved. Matching rates are to be compared with the central panel ones since they are obtained by counting all matches achieved both through the single criteria and through hierarchical matching on possible links. Three combinations outperform the others and precisely criterion 1 followed by criterion 4, criterion 4 followed by criterion 1 and criterion 4 followed by criterion 2. All of them exceed 72% of record matches, but mixing C1 and C4 seems the best solution, since it minimizes the possible matches ratio to 2.5 records in AMeRIcA for 1 record in the ICFM survey while C4+C2 set it to over 8 to 1, due to the low discriminating power of criterion 2. The return to hierarchical matching with respect to criterion 4 alone consists in 2.7% of additional matches gained from the non-matches according to criterion 4. The number of possible matches in fact remains stuck to 57. This of course indicates the necessity to check the differences in the left-out variable (Age Code or Type of Family) between additional matched records.

We now move to evaluate the extent to which our criteria agree or disagree by calculating the size of overlapping between matched and non-matched records by the four independent exercises (table 5). Note that in the calculations reported in Table 5, possible matches are joint to non-matches, meaning that e.g. combination “1000” represents the number of records linked by criterion 1 and non-linked or possibly linked by the remaining three criteria.

Table 5: Overlapping of linkage outcomes among ICFM households

Linkage outcomes*	frequencies	percentage
0000	193	23.9%
1000	22	2.7%
0100	22	2.7%
0010	8	1.0%
0001	70	8.7%
1100	0	0.0%
1010	0	0.0%
1001	275	34.0%
0101	3	0.4%
0011	2	0.2%
0110	1	0.1%
1110	0	0.0%
1011	100	12.4%
1101	19	2.4%
0111	0	0.0%
1111	93	11.5%
all	808	100%

*possible matches are grouped with non-matches, so for instance case 0000 counts the records identified as possible links or negative links by all criteria. Viceversa, case 1111 groups the records identified as positive links by all criteria and so on.

Results in Table 5 can be summarized as follows: of the 808 ICFM households 11.5% were matched by all four criteria, 14,7 % by three criteria, 34.8% by two, 15.1% by 1 criterion and the remaining 23.9% were not matched at all. We can confirm Jenkins et al. (2008) results of a small degree of overlap between records matched by different criteria. The largest agreement is attained between the two best criteria, the first and the fourth ones, with 34 percent of the records matched by both but not by the second and third criteria. It is worth noting that only 11.5% of ICFM households are matched simultaneously by the four criteria, but this is to be ascribed to possible matches being considered as non-matches (as literature suggests, see Gill, 2001). One may have in fact the erroneous impression that overlapping links resulting from the four criteria is equivalent to a matching exercise using all variables. Instead, we calculated that the rate of linkage achieved by imposing that all four variables (Census Section, Age Code, Sex Code and Type of Family) coincide is 64%, corresponding to 519 records matched. On the other hand, as mentioned, working with a few variables at a time has the considerable advantage to separately evaluate their effect on the matching process.

3.3 Linkage accuracy and clerical inspection.

In order to assess the accuracy of our linkage exercise we follow again Jenkins et al (2008) suggestion to measure it along both dimensions of false-positive rate and false-negative rate.

The false-positive rate is calculated for criterion C*i* as the number of mismatches by C*i* over the total number of matches by C*i*, where mismatch is defined as a link classified as positive by C*i* and negative or possible by the other three criteria. Referring to table 5 and considering criterion 1, the false-positive rate is given by the frequency value corresponding to "1000" over the sum of all frequencies corresponding to matching patterns of type 1xxx (no matters if the other criteria agree or not with the match). This rate represents in principle the extent to which a given criterion is wrong in assessing positive links, but it is also likely to be the higher the larger is the discriminating power of the criterion or the lower is the reliability of other criteria. Therefore, a careful inspection of the matches performed by one criterion only should be practiced.

The false-negative rate, on the contrary, is here defined as the fraction of negative or possible links by C*i* that are judged positive by at least another criterion. In terms of table 5 entries and referring to criterion 1, the false-negative rate is given by the sum of the frequencies corresponding to match patterns of type 0xxx, with at least one x equal to 1, over the sum of frequencies corresponding to all match patterns of type 0xxx. Again, to correctly estimate false-negative rates clerical inspection of non-matches by any criterion is needed.

Table 6 reports false-positive and false-negative rate estimates. The ranges are derived adjusting the above calculations after checking single cases classified as a match by one criterion only. Through visual inspection of variables regarding match patterns 1000 and alike, we observed that:

- for pattern 1000, which counts 22 positive links by criterion 1 judged as negative by the other three criteria that share Sex Code variable, we can consider 14 cases as genuine matches since in 6 ones the survey does not report sex for at least one component (so our Sex Code variable ends up with an NA) and in 8 cases the high number of components suggests that correspondence between Age Codes is unlikely to be casual. These 14 cases can be attributed to measurement error in the sex variable;
- Pattern 0100 presented 22 households that differ substantially by Age Code (meaning all ages in the code differ), therefore we conclude they represent true mismatches;
- Pattern 0010 returned 70 records. Among these, 40 households were classified with Type of Family equal to 5 in AMeRIcA and therefore could not have a counterpart in the ICFM survey, while three of them were classified as single parents in ICFM and as couples with children in AMeRIcA. After checking Age Code we can state

they are correctly classified in the administrative source. So these 43 can be considered correctly classified as positive links by criterion 3 while the remaining 37 are possible matches;

- Pattern 0001 corresponds to 8 households with different Census Section. We cannot state with certainty if this is due to measurement error or residential mobility (but we are inclined to favor the second factor since the sections differ by all digits), therefore we have adopted a conservative approach and considered them as true mismatches.

Table 6: False-positive and false-negative rates

	false-positive rate		false-negative rate	
Matching criterion	percentage	(n*)	percentage	(n*)
C1: Census Section, Age Code and Type of Family	1.6%-4.3%	(509)	24.6%- 35.5%	(299)
C2: Census Section, Sex Code and Type of Family	15.9%	(138)	58.4%- 63.0%	(521)
C3: Type of Family, Sex Code and Age Code	3.9%	(204)	64.7%-68.0%	(604)
C4: Census Section, Sex Code and Age Code	5.2%-12.5%	(562)	16.8%-21.5%	(246)

*denominator of the rate calculation. See text for explanation on the calculation of ranges.

This analysis leads to modified false-positive rates for criteria C1 and C4 once corrected for false mismatches caused by the left-out variable, and to modified false-negative rates for all criteria. Criterion 1 further reduces both its false-positive rate, achieving the lowest one equal to 1.6% and negative rate (the second lowest). Criterion 3 was the one with the original smaller value of false-positive rate (3.9%), but from previous results we can say that it was not very useful during our matching process. The scarce validity of criteria 2 and 3 is confirmed by their large rate of false-negatives. Criterion 4, on the contrary, has the smaller false-negative rate and the third largest false-positive rate.

To sum up, criterion 4 seems to be the best one since it shows the largest linkage rate (both alone and combined with criterion 2 or 1), the lowest false-negative rate and a false-positive rate comparable with the lowest ones. The combination of Census Section, Age Code and Sex code could then lead to satisfactory matching rates. Type of Family showed a certain degree of subjectivity in the ICFM survey where the administrative category “other type of family” was forced mainly into “couple, with or without children”.

As to the other single variables, we can add that inspection of the records not matched by any criterion (44) revealed that 31 of them share the same Age Code and Sex Code with a record in AMeRIcA[8] but for a digit, which in most cases is missing in one or the other dataset. This in turn is due to a difference in the number of components and should be attributed most probably to different dates of registration of the same piece of information. Resolving the 57 possible matches (see Table 4, criterion 4 and C1+C4) instead would require additional information, since they agree on all the four variables. In particular, 84% of them are singles so Age Code and Sex Code lose their discriminating power, and the remaining 16% are 2-components households.

[8] In order to retrieve a correspondent record in AMeRIcA, we checked for households with plus/minus one component within the section reported in the ICFM survey.

4 Conclusions and further research on matched data.

In this paper we have combined two data sources: the sample survey on family expenditure conducted by the Milan Municipality and the Chamber of Commerce of Milan (wave 2007-2008) and the tax register matched to the local population and family register in the data-warehouse AMeRIcA, regarding residents in Milan in 2007.

Following Jenkins et al. (2008) procedure, we have proposed a few matching exercises that lead to the identification of the same households across the ICFM survey and AMeRIcA data warehouse.

First of all we confirm the positive results obtained by the above authors on the feasibility of this kind of data linkage. A second contribution of our paper is to suggest and test alternative variables to perform data combination, even in absence of unique identifiers or highly identifying characteristics. In fact, for the ICFM survey respondents, we were not allowed to access a few high quality variables, such as address, names or date of birth widely used –and thus widely examined- in previous studies. On the other hand we could exploit a distinctive feature of the datasets to be merged: to collect information on both households and individuals. We have then proposed to integrate information on individuals and households in order to construct artificial family variables on the basis of their components' age and sex. Another original aspect of this work consists in the analysis of possible links in order to assess discriminating power of involved variables.

Comparison between records was based on four variables characterizing households and precisely Census Section, Type of family, Age Code and Sex Code. The selected variables were combined in four criteria, each of them excluding one variable in turn. The linkage was then conducted through the almost-exact matching method (Gill, 2001), considering as positive links all paired records with at least three corresponding variables. The matching procedure was divided into a first stage characterised by application of single criteria, a second stage of hierarchical matching and a third one for assessing accuracy and clerical inspection.

The main, encouraging, result is that 76% of the records in the ICFM survey are successfully matched according to at least one criterion, only 5.5% do not find a correspondence in AMeRIcA and the remaining 18.4% remain possible links. More specifically, results can be summarized as follows.

The first (Census Section, Age Code and Type of Family) and the fourth (Census Section, Sex Code and Age Code) criteria return matching rates remarkably larger than the remaining two. In particular, the best combination of variables is given by Census Section, Sex Code and Age Code with a matching rate of 69.6% while Census Section, Age Code and Type of Family matches 63% of the records. The outcomes of hierarchical matching combining two criteria in turn again outline Criterion 1 and criterion 4 as the most appropriate, in that not only they noticeably increase linking rates with criteria 2 and 3 but improve also each other performances. The assessment of linkage accuracy does not completely help to favour criterion 1 on criterion 4 or viceversa, since the former shows a lower false-positive rate while the latter a lower false-negative rate. Though, a careful comparison of the above rates points to criterion 4 as the best one, since its false-negative rate is almost 10 points lower while its false-positive rate less than 4 points higher than the corresponding rates of criterion 1.

Besides, criterion 4 helps in solving possible links generated by criterion 1 while the opposite is not true.

Analysis of discriminating power and reliability of single variables was also possible thank to the leave-one-out variable procedure in constructing matching criteria along with clerical inspection of non-matches. Turning to single variables, in fact, we can state that Census Section and Age Code are the outstanding ones in our matching exercise. A comparison of possible linkage rates with criteria excluding them (criterion 2 and criterion 3) confirms the noticeable discriminating power of Census Section versus Age Code since lack of the first leads to a largest ratio between the number of records in AMeRIcA to be possibly matched to ICFM survey records and the number of the corresponding ICFM records. As for reliability, the two artificial variables share potential reporting errors in the number of components. More, Sex Code is affected by errors in age recording when this implies reverting components' age ranking. In fact, visual inspection of non-matches revealed that for 31 households a missing component in one or the other dataset implied failure of matching by both Age Code and Sex Code. Furthermore, in some cases reporting errors in Age Code turn to errors in Sex Code.

Census Section, instead, is a highly reliable variable since also in the survey it is assigned and not asked for. In principle it is subject to reporting errors and discrepancies due to residential mobility, but we ascertained that only 8 households were classified as non-matches according to different Sections. Finally, Type of Family is for sure the least discriminating variable since it seems of help only in criterion 1, mixed to Census Section and Age Code, while criteria 2 and 3 lead to quite poor results. Furthermore, comparison between possible links and negative links regarding criteria 1 and 4 highlighted that Type of Family classifies more non-matches (208) with respect to Sex Code (189) when coupled to Census Section and Age Code. The reliability of this last variable is questionable since, in the ICFM survey, interviewers categorize households only in four categories instead of the five present in the administrative file. Once checked for this discrepancies (we find 40 households misclassified for this reason) the Type of Family might be rather used as a blocking variable.

We think that, in order to improve matching rates, given the considerable number of possible matches left unsolved also by the pooled matching, a probabilistic matching approach might be applied with success. This would require careful consideration of two critical aspects. The first is the correlation between Age Code and Sex Code, which violate the assumption of independence among variables necessary to construct matching probability ratios (Fellegi and Sunter, 1968); the second one is to derive estimates for errors in the generating process of matching variables. As to this second point, we think that repeating the present exercise with the next wave of the survey and 2008 tax records could serve the purpose, using one of the two matches as a training set to estimate errors and thresholds and the other as a validation set for checking results. Besides, repeating the analysis in subsequent years might help for controlling measurement errors and omissions and also to confirm differences due to temporal lags between the interviews and the administrative updating process.

In this paper we have devoted our attention on the methodological and practical sides of data linkage. Matched households are now endowed with information collected in both datasets ready to be used. For instance, this would allow us to perform comparisons based on the income information collected in the two data sources. Recent studies (Fiorio and D'Amuri, 2005; Marino and Zizza, 2008) show that the combined used of fiscal income and income derived from survey is a reliable way for estimating tax evasion. This approach meets with the difficulty of obtaining fiscal data that may be compared with survey income data (Baldini et al., 2009). Our linkage exercise represents an initial attempt to overcome this issue, although on a local basis, thereby providing new evidence in favour of the successful exploiting of both data sources. However, this represents only a small part of the possible uses of such a combined dataset.

To conclude, the combined data set derived from our linkage exercise might become a fertile field for both the construction and the study of univariate and multivariate indicators of well-being. In fact, when dealing with the measurement of economic well-being, its several dimensions should be taken into account. One of the most recent and widely-used approaches (Berloffa and Modena, 2010, Stiglitz Commission 2008) leads to the definition of composite indicators to shrink the several dimensions of well-being into a synthetic measure. We might also contribute to the debate on definition and measurement of well-being offering a systematic comparison of different levels of well-being as measured by income (retrieved from tax records) and consumption (obtained from the sample survey). In our opinion this is a very important matter in Italy considering that, although consumption is generally considered as a better indicator of well-being (Meyer e Sullivan, 2003), policy interventions on poverty are almost exclusively based on tax record income. This becomes even more important since combining the two data sources we can evaluate expenditures of non tax-payers.

Notwithstanding the fact that the information contained in both datasets is confined to a single Italian municipality, we think results may have broader significance, especially on the methodological side, considering that all municipalities have access to the tax records of their residents and many of them conduct their own expenditure surveys.

References

Atkinson, A.B., 2007, Measuring Top Incomes: Methodological Issues, in A. Atkinson and T.Piketty (eds) Top Incomes over the Twentieth Century: A contrast Between European and English Speaking Countries, Oxford, Oxford University Press.

Baldini, M., Bosi, P., and Lalla, M. (2009) Tax Evasion and Misreporting in Income Tax Returns and Household Income Surveys. Politica Economica, 3, pp 333-348.

Belin, TR e Rubin DB, 1995, A Method for Calibrating False-Match Rates in Record Linkage, JASA, 90.

Berloffa, G., and Modena, F., 2010, Economic well-being in Italy: The role of income insecurity and intergenerational inequality. ECINEQ Working Paper 2010 – 168.

Copas, JB e Hilton, FJ, 1990, Record linkage: statistical models for matching computer records, J. of the Royal Statistical Society (A), pp.287-320.

Crosato, L e Zavanella B, 2010, L'evoluzione del reddito dei cittadini milanesi (2000-2004) sulla base di archivi amministrativi" in M. Mezzanzanica e B. Zavanella (a cura di) "I numeri della città: un quadro socio-economico del comune di Milano sulla base di fonti amministrative", FrancoAngeli, Milano.

Dragoset L.M. and Fields G.S., 2006, U.S. Earnings Mobility: Comparing Survey-Based and Administrative-Based Estimates, ECINEQ working paper 2006-55.

Fiorio, C.V., and D'Amuri, F. (2005) Worker's tax evasion in Italy. Giornale degli economisti e Annali di economia. 64, pp 247-270.

Gill, L., 2001, Methods for automatic record matching and linking and their use in National Statistics. National Statistics Methodological Series No. 25. London: Office for National Statistics.

ISTAT, 2008, Combining survey and administrative data in the Italian EU-SILC experience: positive and critical aspects, UNECE- Work session on statistical data editing, WP-14, retrieved in August 2009 from http://www.unece.org/stats/documents/2008/04/sde/wp.14.e.pdf.

Jabine TB and Scheuren FJ, 1986, Record Linkages for Statistical Purposes: Methodological Issues, Journal of Official Statistics, 2(3), pp. 255-277.

Jäckle A., Sala E. Jenkins S.P. e Lynn P., 2004, Validation of Survey Data on Income and Employment: the ISMIE Experience, ISER Working Papers, N. 2004-14.

Jenkins SP, Lynn P, Jäckle A., Sala E, 2008, The Feasibility of Linking Household Survey and Administrative Record Data: New Evidence for Britain, Int. J. of Social Research Methodology, 11, pp 29-43.

Jones, P. and Elias, P., 2006, Administrative data as a research resource: A selected audit. A report to the ESCR Research Resources Board. Retrieved in August 2009 from http://www.esrc.ac.uk/ESRCInfoCentre/Images/Administrative_Data_a_selected_audit_tcm6-25869.pdf

Mariani, P., Mezzanzanica M. e Zavanella, B., 2006a, "Statistical Information Systems and Data Warehouses for Job Marketplaces" *Atti della XLIII Riunione Scientifica della Società Italiana di Statistica*, CLEUP, Torino 14-16 giugno 2006, pp.289-292;

Mariani, P., Mezzanzanica, M. e Zavanella B., 2006b, Sistema informativo statistico per il supporto decisionale: il progetto AMERICA - Anagrafe Milanese E Redditi Individuali Con Archivio, in *Metodi Modelli e Tecnologie dell'informazione a supporto delle Decisioni*. A cura di P. Amenta, L. D'ambra, M. Squillante, e A. Ventre, Pubblicazioni DASES (Dipartimento di Analisi dei sistemi economici e sociali Università degli Studi del Sannio). Franco Angeli.

Marino, R. and Zizza, R. (2008) L'evasione dell'IRPEF: una stima per tipologia di contribuente. Paper presented at the XX Conference SIEP, September 2008.

Meyer BD e Sullivan JX, 2003, Measuring the well-being of the poor using income and consumption, Journal of Human Resources pp. 1180-1220.

Minotti, SC, Mussini M and Zavanella, B, 2010, Simulazione di alcuni sistemi fiscali europei sui redditi delle famiglie milanesi, in M. Mezzanzanica e B. Zavanella (a cura di) "I numeri della città: un quadro socio-economico del comune di Milano sulla base di fonti amministrative", FrancoAngeli, Milano.

Ridder, G e Moffit, R, 2007, Econometric Methods for Data Combination, Handbook of Econometrics.

Stiglitz Commission, 2008, Survey of Existing Approaches to Measuring Socio-Economic Progress. Joint Insee-OECD document prepared for the plenary meeting of CMPSP by (at) Insee).