Session Number: Session 2B
Session Title: Improving Estimates from Survey Data
Session Organizer(s): Stephen Jenkins, University of Essex, Colchester, UK, and Holly Sutherland, University of Essex, Colchester, UK
Session Chair: Holly Sutherland, University of Essex, Colchester, UK

Paper Prepared for the 29th General Conference of
The International Association for Research in Income and Wealth

Joensuu, Finland, August 20 – 26, 2006

ADMINISTRATIVE AND SURVEY MICRODATA ON SELF-EMPLOYMENT: THE ITALIAN EXPERIENCE WITH THE EU SILC PROJECT

Paolo Consolini, Marco Di Marco, Roberta Ricci, Silvano Vitaletti

For additional information please contact:

Author Name(s) : Marco Di Marco
Author Address(es) : ISTAT - via Ravà 150, 00142 Roma, Italy
Author E-Mail(s) : marco.dimarco@istat.it
Author Telephone(s) : (+39) 6 4673.4744

This paper is posted on the following websites: http://www.iariw.org

IARIW 29th General Conference
Parallel session 2b 'Improving estimates from survey data'
Joensuu, Finland 20-26 August, 2006

Administrative and Survey Microdata on Self-Employment: the Italian Experience with the EU SILC project

Paolo Consolini, Marco Di Marco, Roberta Ricci, Silvano Vitaletti

Istat
Italian National Statistical Institute

preliminary draft, please do not quote

contact person:

Marco Di Marco
ISTAT - via Ravà 150 00142 Roma (ITALY)
phone (+39) 6 4673.4744
e-mail marco.dimarco@istat.it

ABSTRACT
The Italian experience with the EU SILC project may provide some insights on the definition and measurement of self-employment incomes when both survey and administrative data are available. To cope with the demanding aim of the project, the Italian national statistical institute set up a mixed data collection strategy, based on a *paper and pencil* face-to-face interview and on the *linkage* of administrative with survey data. A first conceptual issue concerns the definition of self-employment incomes. Economic, accounting and administrative definitions of self-employment incomes do not necessarily match and could raise problems of reliability and comparability. Moreover, the different definitions have an influence on the subjective understanding of the term 'income' by the respondents. In the Italian EU SILC, disposable self-employment income is set as the maximum value between the net income reported in the survey questionnaire and the net *taxable* income in the tax return. Under the assumption that no individual over-reports her/his income, the rule minimises under-estimation either in the administrative or in the survey data, depending on which of the two is larger. The paper summarises the data production process of the Italian EU SILC, focusing on the collection, editing and imputation of survey incomes, on the record linkage between survey and administrative data and on the empirical results obtained. With respect to the exclusive use of survey data, the linkage with administrative data has increased substantially the number of percipients (+15,6 %) and the average self-employment income (+11,9 %). Among the individuals for which both sources contain self-employment incomes, the record linkage reveals that under-estimation is more frequently observed in the tax data than in the survey data. It turns out, moreover, that self-employment income in the integrated dataset is more unequally distributed than in the survey .

CONTENTS [1]

1. Self-employment incomes in the Italian EU SILC: an overview p. 2
2. Taxable self-employment incomes p. 4
3. Data collection, editing and imputation of self-employment incomes p. 5
4. Combining survey and administrative data on self-employment incomes p. 10
5. The distribution of self-employment incomes in the Italian EU SILC p. 15
Appendix: 'The survey questionnaire on self-employment incomes' p. 17
References p. 19

[1] Sections 1 and 5 have been drafted by Marco Di Marco, section 2 by Roberta Ricci, section 3 by Silvano Vitaletti and section 4 by Paolo Consolini. The opinions are those of the authors and do not necessarily involve the responsibility of the Istat.

1. SELF-EMPLOYMENT INCOMES IN THE ITALIAN EU SILC: AN OVERVIEW

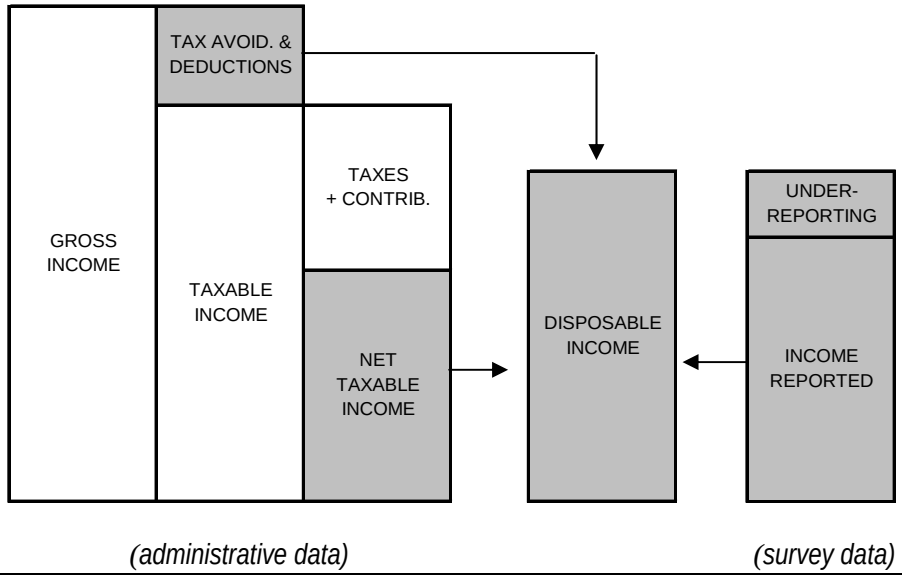
Two clear-cut statements, taken from the "Canberra Handbook", depict the state of the art for what concerns the measurement of self-employment incomes in household surveys:

> "*Income data for the self-employed are also generally regarded as unreliable as a guide to living standards*" (Canberra Group, 2001, p.54).

> "*Household surveys are notoriously bad at measuring income from capital and self-employment income*" (Canberra Group, 2001, p.62).

Figure 1 below shows, in a simplified sketch, the problem of collecting self-employment incomes when either survey or administrative data are available: the shaded areas correspond to the income available to an individual for his/her personal use.

Figure 1 Personal gross, taxable, reported and disposable income



The alternative sources of microdata on earnings from self-employment may not contain the item 'disposable income' as such. Survey data may be affected by *under-reporting*. Administrative data gathering the individual tax returns do not take account, of course, of illegal tax evasion and may not display all the authorized deductions allowed in the calculation of taxable income (tax avoidance). The accounting books, on their turn, usually report about the taxes paid by the company as a juridical entity and do not contain information on the personal taxes levied from the owners' profits. However, ignoring tax evasion, the accounting profits, net of company taxes, can be viewed as a measure of gross personal income. Nonetheless, they could still be different from personal taxable income. Indeed, the tax authorities may allow special deductions for the profits retained and invested in the business, stipulate departures from accounting rules for depreciation costs *etc.* Some categories of taxpayers (*e.g.* small family business, farmers, starting-up companies…) may be subject to a preferential tax regime that grants them special benefits.

In the EU SILC project, the standard procedure to measure net self-employment incomes requires to collect the amount of money drawn out of self-employment business only when the profit/loss from accounting books or the taxable self-employment income (net of corresponding taxes) are not available. For the Italian EU SILC, both tax and survey microdata are available, through an exact matching of administrative and survey records (Istat 2005, 2006). However, both sources may be affected by under-estimation of self-employment incomes. Moreover, some individuals report self-employment incomes in only one datasource. This is the case of some individuals whose professional status at the time of the interview is different from

that of the income reference period and of many percipients of small and/or secondary self-employment incomes.[2]

For what concerns the definition and the collection of self-employment incomes in the Italian EU SILC, the choice between the available sources has taken into account the following considerations:

- Tax returns usually contain exact information on taxable incomes and tax liabilities. They also provide information on social security contributions. Therefore, tax records can be used to measure the net *taxable* income. In general, neither taxable income is identical to gross income, nor net taxable income is identical to disposable income (see figure 1). In principle, if the deductions from profits are available to the company owners for their personal use, then they should be considered as components of both the gross and the disposable personal incomes. However, not all the tax abatments allowed are explicitly shown in the tax returns. By definition, tax evasion is also not available in the tax files.

- Survey data on self-employment income may be affected by *under-reporting* to an unknown extent. Moreover, gross income is usually unknown by the interviewees and the collection of the additional information needed to compute it (taxes and social security contributions) puts an excessive response burden on the respondents. Survey data should therefore be integrated by external sources and/or microsimulations. In both cases, the amount of taxes and contributions could be added to the net income reported in the survey to get a 'survey based' measure of gross income. If taxes and contributions are correctly measured, the result of such an addition gives gross income net of survey *under-reporting* (this should be kept in mind when assessing the international comparability of the data). Therefore, the 'survey based' measure of gross income would be equal to taxable income only if survey under-reporting equals the sum of tax avoidance and of tax evasion.

In the Italian EU SILC, when both the administrative and the survey datasources report it, income from self-employment is set equal to the maximum value between: (i) the (net) self-employment income resulting from the tax return and: (ii) the (net) self-employment income reported by the interviewee. This departure from the standard definition is adopted in order to minimise either under-estimation due to tax avoidance/evasion in the administrative data or under-reporting in the survey data, depending on which of the two is larger. The procedure increases the degree of international comparability, under the assumption that self-employment income in the benchmark country is not under-estimated. Under the alternative assumption that self-employment incomes are downward biased in the other country, international comparability depends on the degree of under-estimation affecting the two national data sources. Besides, provided that accuracy is at least as important as international comparability, it seems advisable to regard the minimisation of under-estimation as a prerequisite for the measurement of incomes in any case.

For what concerns the possible *trade-offs* between accuracy and international comparability, it is important to note that comparability 'within-country' ("any couple of households living in country A can be compared") could be regarded as a necessary, though not sufficient, condition for 'across-country' comparability ("any household living in country A can be compared with any household in country B"). Since the under-estimation of some income components decreases 'within country' comparability, it follows that accuracy and international comparability should be simultaneously pursued by setting in each country the true national income as the 'comparable benchmark'.

From the point of view of economic theory, a controversial issue concerns the allotment of self-employment earnings between the categories of labour and capital incomes. At this regard, the naming and accounting conventions encompassed in the tax laws are not necessarily the most suitable for economic analysis and, moreover, may also hamper international comparability. The System of National Accounts opportunely sums up both components in the concept of 'mixed income', a convention that permits to analyse them as rewards for independent labour, often assisted by the worker's capital[3]. The Canberra Group (2001, p. 118) and the

[2] The survey data include as self-employment incomes those small compensations for minor and informal services that are frequently unnoticed for tax purposes. For example, the earnings of baby-sitters. On the other hand, some minor self-employment incomes shown in the tax returns may be disregarded during the interview to ease the response burden.
[3] Some self-employeds (*e.g.* subcontractors) do *not* use their own capital in production.

ILO resolution on income surveys (2003, p.2) recommend to exclude from self-employment income the profits of unincorporated businesses distributed to 'sleeping partners', an advice that clearly attaches more weight to the 'labour' component. Given the ambiguity of the definition of self-employment incomes in the tax laws, for the Italian EU SILC the tax source has been used with caution (substantially, to check and replace the underreported survey incomes). In fact, to avoid errors due to legal definitions, the earnings of the self-employeds that have been reported in the tax data exclusively under the 'capital incomes' heading have been ignored (*i.e.* they have not been compared with the survey incomes, nor have they been loaded in the final dataset)[4].

2. TAXABLE SELF-EMPLOYMENT INCOMES

The Italian Civil Code and the TUIR[5] classify the earnings from self-employment under different names. In particular, in the Italian Personal Tax Code the incomes of the self-employeds are scattered across an assortment of different items:

(i) *labour incomes*
- the compensation for the effort spent by an independent worker in her/his job;
- the reward for the entrepreneurial ability to organise the inputs in best way;
- the remuneration of a temporary worker, formally hired as an independent collaborator or subcontractor, whose role is in fact not too different from dependent employment (Co.Co.Co.);
- the remuneration of a cooperative stakeholder;
- the fees earned by members of advisory committees, consultants etc.;

(ii) *capital incomes*
- the profits of unincorporated businesses run by independent workers (additional to the compensation for their work);
- royalties on patents, writings, artworks etc…[6];

For some juridical persons, the remuneration for the business risk of the non-working stockholders (which is *not* self-employment income, according to the Canberra Handbook) goes under the same name given to the capital incomes of the (working) owners.[7] Another confusing stipulation is the inclusion of cooperative stakeholders and of 'Co.Co.Co.' subcontractors among the self-employeds, on the one hand, and in the group of employees and pensioners whose earnings are taxed under the name of 'Dependent employment and similar incomes', on the other hand.

The methods prescribed to measure gross and taxable self-employment incomes vary accordingly. For each of the preceding items, both the Civil Code and the TUIR contain detailed accounting rules, also depending on the legal status of the business. The definition of *taxable* self-employment income is, consequently, quite complex. There is not a set of simple, wide-ranging rules to convert the accounting profits, the compensation for independent work etc. into a measure of the tax base of the IRE[8]. Furthermore, to the regulations of the Civil Code relating to the measurement of self-employment incomes as such, the Tax Code adds a large list of additional instructions, addressed to particular categories of taxpayers and/or related to specific accounting items, in order to differentiate the tax burden among the self-employeds.

The three main broad categories of special accounting rules for tax purposes are:
- departures from the standard accounting rules that indirectly imply tax benefits or surcharges;

[4] In these cases, the survey income is retained in the final dataset as it is.
[5] TUIR (Testo Unico delle Imposte sui Redditi) is the Italian Personal Tax Code.
[6] These could generally be seen as (intellectual) property incomes. However, they could sometimes represent the normal remuneration for the regular commitment of an independent worker. For example, the rewards for the weekly articles of a free-lance journalist could be reported as author's royalties.
[7] In some cases, the stockholders' work effort may be just pretended, as it permits tax splitting. For example, in the small, family-run, businesses the profits could be shared with sham co-helpers in order to take benefit of lower marginal rates.
[8] IRE is the Italian personal income tax.

- tax benefits (surcharges) in the form of explicit reductions (increases) of the tax base;
- administrative estimates of the 'normal' amounts earned from specific self-employment activities.

Examples of the first category are the favourable rules concerning the evaluation of capital depreciation (accelerated amortization) and the DIT (Dual Income Tax). Under the DIT, the stockholders' profits are splitted into two separate components: 'capital' and 'earned' income. 'Capital' income is computed by applying a standard rental rate, set by the tax authorities, to the value of assets and then taxed at a favourable flat rate. The residual 'earned' income is taxed under the normal progressive personal tax schedule.
On the other hand, some costs displayed in the accounting books are not wholly deductible from the tax base. For example, only 50 % of the expenditures on goods and services used for professional and personal uses as well (*e.g.* a car) can be accounted for as negative components of the tax base. Similarly, some 'public relations' expenses (*e.g.* presents to customers, hotel accomodations for visitors etc.) can be deducted only up to a given percentage of total revenues. Another departure from the standard accounting rules concerns small family businesses, allowed to compute the tax base under a simplifyed set of accounting rules. Also, preferential tax regimes are allowed for starting-up businesses, including those emerged after years of tax avoidance in the 'informal' economy.

An important example of an explicit tax benefit is the Law n. 383/2001 (the so-called 'Tremonti bis'), that provides incentives to investments. The rule allows a deduction from taxable income that is related to the profits retained and reinvested in the business. The scope of the incentive is to promote technological change through the renewal of the stock of capital: half of the increase in the expenses on new capital equipment may be excluded from taxable income. Moreover, the increase is evaluated with respect to the average expenditure on capital goods of the preceding five years (except for the highest value).

The 'Sectoral Studies' (*Studi di settore*), introduced by the Law n.427/1993, are the most important example of the use of administrative values for the assessment of taxable self-employment incomes. Most professionals, retailers, artisans etc… may report a 'normal' amount of taxable income, estimated by the Tax Administration on the basis of a cluster analysis of a large sample of businesses and professionals. By accepting the administrative estimates, the self-employeds avoid official auditing by the *Guardia di Finanza* and, therefore, any possible surcharge and fine for tax evasion. Thus, provided that their true gross income is greater than the administrative threshold level, these taxpayers benefit from a deduction. Another favourable tax regime, based on administrative estimates, is allowed to minimal taxpayers, engaged in marginal self-employment activities (secondary or part-time independent jobs, occasional free-lance services etc…): whatever the actual costs incurred, taxable profits can be estimated as a given percentage of total revenues.

3. DATA COLLECTION, EDITING AND IMPUTATION OF SELF-EMPLOYMENT INCOMES

In the survey questionnaire for the Italian EU SILC, the amount of self-employment income is asked after a reminder question, requesting YES/NO replies to a list of possible personal uses of earnings (consumption and saving)[9]. This sequence has been devised in order to suggest to the interviewee an interpretation as close as possible to the 'money drawn out' concept. In effect, for some self-employeds the literal translation into Italian of the question "Have you drawn out money from business for your personal use?" may evoke the idea of a deceitful behaviour like, for example, to withdraw money from the cash account without taking note in the books of the corresponding revenues (tax evasion).

For what concerns the amount of self-employment incomes, the instructions to the interviewers advise them to explain that "*self-employment activity has led to*:

- ***earnings*** *if the individual or her/his family has got from it an amount of money that has been used for personal/household expenses, saved, invested in the business or in financial activities, dwellings and other real estates;*
- ***a loss*** *if he/she has not obtained from it any money to pay for personal/household expenses or to save/invest and, also, has used incomes from other sources, borrowed money or sold assets to pay for the costs of the self-employment activity.*"

[9] See the questions in the Appendix

The reason for such a definition is quite simple: if positive earnings are 'money drawn out' from business, then losses should be understood as 'money put into' it. It is important to highlight that respondents are prompted to distinguish between money invested in the business and other expenses related to their self-employment activity. The former may be considered as an increase in the value of the business (*i.e.* of respondents' wealth), while the latter amount to a genuine loss, when not offset by sufficient revenues. During the pilot tests of the EU SILC questionnaire, most self-employeds have proved to be much more confident with the simple logic of the preceding definition than with the concept of income entailed by the accounting rules (to say nothing of the complex computation of taxable income). After all, households living on self-employment incomes should wish to estimate from time to time how much money can be spent or saved by sketching an off the record summary of their operations. Whilst these home-made unofficial budgets may well be sufficient to manage the households economic resources, the exhausting fulfilment of accounting books and Tax Reports is usually left to tax consultants[10].

The Italian administrative registers, based on the self-reporting of income for tax purposes, are also available for the EU SILC project. Nevertheless, it is not advisable to rely exclusively upon this source of information for the measurement of self-employment incomes, for the reasons already explained in the previous section. In fact, in order to measure self-employment income as precisely as possible, the administrative data has been used, when opportune, to replace the survey data. It was expected that, though the interviewees may show a certain degree of reticence, in the Italian context survey under-reporting should have a more limited extent with respect to tax avoidance and evasion, as the answers to the survey questionnaire do not entail tax consequences[11]. Moreover, to minimise the percentage of missing answers to the income question, for those respondents who do not remember the exact amount of their self-employment income, a supplementary question asks for an approximate amount, to be selected out of a predetermined list.

Nevertheless, a minority of interviewees refuses to answer about their self-employment incomes, it does not matter how careful is the interviewer. It is also important to highlight that the interviewers were repeatedly advised not to compel persons visibly embarrassed or bothered, as they could provide false answers. As a general principle, missing answers were always preferred to false ones. In addition, interviewers were also asked to directly provide their own assessment, after the interview, of the reliability of the reported incomes. The latter information has been used, in the editing phase, to detect and eliminate some blatant errors. Other answers have been eliminated as they appeared as *outliers* according to the Hidiroglou-Berthelot method, modified in order to be applied to the case of an univariate distribution[12]. The whole approach to the collection of self-employment incomes through personal interviews aims at minimising reporting errors and,

[10] This conclusion is supported by many anedoctal reports from the interviewers. Also, during an in-depth interview, made to test the questionnaire for the EU SILC pilot surveys, a shopkeeper said that in 2002 his *gross* taxable self-employment income 'could have been' about 12,000 euros, while his *net* earnings had amounted to 24,000 euros, approximately. He explained that the latter figure was an estimate of his, based on personal documentation. In his words, the difference with respect to taxable income was due to legal deductions made by the tax consultant. Interestingly, he had no precise idea about the 'cooking-up' of official accounting books and tax returns made by his tax consultant.

[11] Needless to say, a special effort has been made to persuade the interviewees that, according to the Italian laws, their answers are collected solely for statistical purposes and will never be transmitted to the tax authorities.

[12] The Hidiroglou-Berthelot method first transforms the variable, in order to make its distribution symmetrical:

$$HB_Y = \begin{cases} \dfrac{Y - Me_Y}{Y} & 0 < Y < Me_Y \\ \dfrac{Me_Y - Y}{Me_Y} & Y \geq Me_Y \end{cases}$$

Then, to account for density, the minimum and maximum thresholds are determined according to the distance between the quartiles and the median:

$$HB_{\min}(K) = Me_{HB_Y} - K(Me_{HB_Y} - Q1_{HB_Y}) \quad HB_Y < Me_{HB_Y}$$

$$HB_{\max}(K) = Me_{HB_Y} + K(Q3_{HB_Y} - Me_{HB_Y}) \quad HB_Y \geq Me_{HB_Y}$$

The observations found outside this range are considered as *outliers*. The lower (upper) threshold has been set equal to 40 times the distance between the median and the first (third) quartile of the transformed variable (*i.e.* K=40).

at the same time, at devising suitable imputation procedures for the missing values. The setup of the imputation procedures has been eased, on the one hand, by the rich qualitative information available in the survey and, on the other hand, by the reduction of the bias due to the unreliable answers retained among the valid cases. These latter have been minimised by the systematic preference for missing with respect to false answers and by the removal of the unreliable amounts.

Table 1 Valid, missing and not reliable self-employment incomes in the survey data, by individual characteristics at the time of the interview

	Valid		Missing		Not reliable		Outliers		Total	
	N	%	N	%	N	%	N	%	N	%
GEOGRAPHICAL AREA										
North	2775	69.3	763	19.1	463	11.6	4	0.1	4005	100
Centre	1395	75.1	292	15.7	169	9.1	1	0.1	1857	100
South & Islands	1186	66.8	373	21	215	12.1	1	0.1	1775	100
SEX										
Males	3650	72.9	770	15.4	578	11.6	6	0.1	5004	100
Females	1706	64.8	658	25	269	10.2	-	-	2633	100
AGE										
15-34	1420	64.1	599	27	196	8.8	1	0	2216	100
35-44	1612	72.8	363	16.4	239	10.8	-	-	2214	100
45-64	2120	73.2	405	14	367	12.7	4	0.1	2896	100
65 and over	204	65.6	61	19.6	45	14.5	1	0.3	311	100
EDUCATION										
Primary school	816	69.3	214	18.2	145	12.3	3	0.3	1178	100
Intermediate school	1534	67.6	441	19.4	294	13	1	0	2270	100
High school	2133	71.3	553	18.5	305	10.2	1	0	2992	100
Univ. degree, post-graduated, etc.	873	72.9	220	18.4	103	8.6	1	0.1	1197	100
YEARS OF ACTIVITY										
1-5	327	58.4	190	33.9	42	7.5	1	0.2	560	100
6-10	926	66.9	341	24.6	117	8.5	-	-	1384	100
11-14	1064	70.4	308	20.4	140	9.3	-	-	1512	100
15-24	1378	72.8	279	14.7	236	12.5	-	-	1893	100
25 and over	1661	72.6	310	13.5	312	13.6	5	0.2	2288	100
PROFESSIONAL STATUS (a)										
No more self-employed	496	65.1	206	27	60	7.9	-	-	762	100
Co.Co.Co.	494	50.6	431	44.2	50	5.1	1	0.1	976	100
Entrepreneurs	476	80.3	38	6.4	79	13.3	-	-	593	100
Professionals	999	79.8	128	10.2	125	10	-	-	1252	100
Artisans, shopkeepers etc.	2493	76.2	317	9.7	457	14	5	0.2	3272	100
Co-helpers	104	54.5	70	36.6	17	8.9	-	-	191	100
Coop. stockholders	294	49.7	238	40.3	59	10	-	-	591	100
HOURS WORKED PER WEEK										
No more self-employed	496	65.1	206	27	60	7.9	-	-	762	100
Up to 19	153	59.3	85	32.9	19	7.4	1	0.4	258	100
20-29	380	61.3	181	29.2	58	9.4	1	0.2	620	100
30-39	554	64.1	221	25.6	89	10.3	-	-	864	100
40-49	1952	71.3	472	17.2	311	11.4	2	0.1	2737	100
50 and over	1821	76	263	11	310	12.9	2	0.1	2396	100
TOTAL	5356	70.1	1428	18.7	847	11.1	6	0.1	7637	100

(a) At the moment of interview

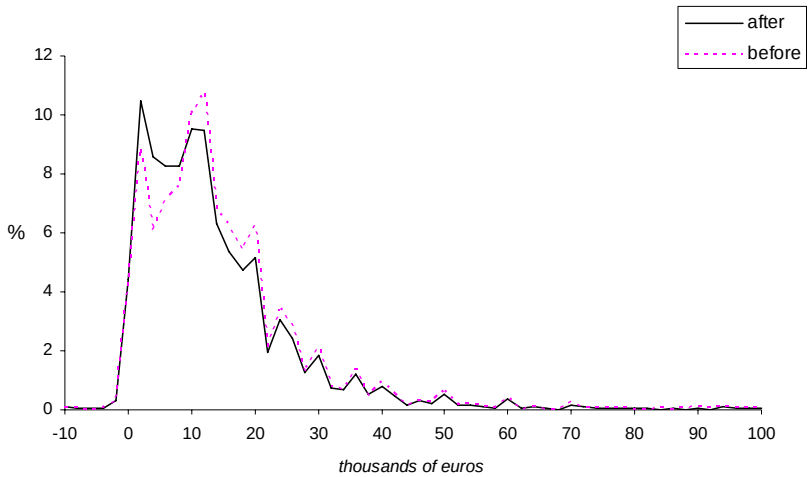
In the 2004 Italian EU SILC, the non-response rate to the question on self-employment income is on average 18.7% and mainly regards *border-line* situations with more volatile incomes (Table 1). Indeed, non-response affects significantly those marginal forms of self-employment such as the “Co.Co.Co.” temporary

subcontractors[13] (44.2%), the members of cooperatives (40.3%) and the co-helpers in small family businesses (36.6%). Higher than average non-response rates are also observed in the case of the self-employeds with a shorter work experience (33.9% for the subgroup with 1 to 5 years of seniority) and of people engaged in secondary or less regular activities (self-employment income is missing for the 32.9% of percipients who work less than 20 hours a week). Finally, non-response is more frequent among the young people (27.0% of the 15-34 age class) and the women (25.0%).

Information considered as not reliable by the interviewers, on its turn, is on average equal to 11.1% and is more frequent among the more experienced and/or regular workers. Indeed, according to the interviewers, income is unreliable for the 12.8% of the self-employeds aged 45 and over, for the 13.1% of those who have been working more than 15 years, the 12.9% of those who work 50 hours a week (or more). As regards the professional categories, entrepreneurs (13.3%) and shopkeepers/artisans (14.0%) stand out. On the whole, about 70% of the interviewees with self-employment incomes reported reliable amounts of income[14].

The missing and unreliable values of survey incomes have been imputed by means of the multiple regression models available in the IVEware software of the University of Michigan. Programmed as an add-on of the SAS System, IVEware enables to process with relative simplicity a high number of variables correlated among them without explicitly modelling a complete set of multivariate relations. In brief, for each missing value of each variable, IVEware generates a prediction conditional to the values of all the other variables. In a first preliminary step, the model imputes in sequence the missing values of the variables, starting from the one that presents the lowest non-response rate. Once these initial values have been imputed, each variable enters as a covariate in the imputation process of all the other ones. From the second step up to the convergence, the model is made up of a number of equations corresponding to the number of variables to be imputed. When the convergence of the estimates has been reached, each equation predicts the missing values of the corresponding dependent variable. The main advantage of IVEware consists in its flexibility: for each variable, it is possible to select the relevant covariates and to define the appropriate regression model (linear model for the continuous variables, logit for the dichotomous variables, log-linear for the discrete variables). Moreover, it is possible to restrict the imputations taking account of consistency constraints (for example, imputing only a particular subset of cases) and to assign upper and lower bounds to the imputed values. The latter characteristic is useful for transforming in precise values the approximate figures indicated by those who do not remember the exact amount of their income. In these cases, the precise values are imputed by IVEware within a band centred on the approximate figures given by the interviewee. In fact, the smoothing of the approximate amounts has been accomplished in a separate step, preliminar to the ‘true’ imputation process of the missing and unreliable answers.

Figure 2 Survey self-employment incomes before and after the imputation process



[13] Co.Co.Co., literally coordinated and permanent collaboration contracts, are legally framed as self-employment (and so registered as independent work) but very often have the same characteristic of temporary dependent work.
[14] Only 6 cases were identified as outliers.

The variables taken into consideration as covariates have been:

- *Territorial characteristics*: geographical area, municipality (type and size);
- *Individual characteristics*: gender, age, marital status, education, health condition, profession (first digit of the Istat CP2001-code), sector of activity (ATECO-code), kind of self-employment, availability of dependent workers, hours worked weekly, multiple jobs, age at which person started to work, change of employment in the past 12 months, past profession, sector of past activity, kind of past self-employment;
- *Characteristics of the household and of the dwelling*: household size, type of dwelling (surface, rooms, possession title), availability of durable goods (washing machine, dishwasher, car, satellite TV, personal computer and/or access to Internet), net monthly household income, subjective appraisal of the ability to make ends meet, utility bills (gas, phone and electricity), condominium fees, municipal street cleaning and sewage removal charges, rent, mortgage repaiments (principal and interest).

Table 2 Mean self-employment incomes in the survey dataset (valid and imputed), by individual characteristics at the time of the interview

	Valid	Imputed (missing)	Imputed (unreliable or outliers)		Total	
			before	*after*	*before*	*after*
GEOGRAPHICAL AREA						
North	15835	12979	5466	16046	14342	15316
Centre	14848	11492	5629	12559	13847	14111
South & Islands	11277	8219	5241	12146	10347	10740
SEX						
Males	16315	12747	6020	15670	14895	15691
Females	10833	9892	4187	11526	9928	10669
AGE						
15-34	11158	10742	4363	11965	10330	11117
35-44	15393	11223	6208	13850	14207	14543
45-64	16093	12755	5555	16176	14524	15637
65 and over	15955	10652	5164	12684	13969	14431
EDUCATION						
Primary school	11495	10212	4713	10591	10454	11148
Intermediate school	12922	9325	4909	13585	11629	12309
High school	15197	12005	5815	15169	14020	14604
Univ. degree, post-graduated, etc.	18800	15400	6891	19570	17533	18242
YEARS OF ACTIVITY						
1-5	9574	9847	2453	9437	8746	9656
6-10	11760	10475	4671	12055	10964	11468
11-14	14726	11749	6672	14295	13789	14079
15-24	16223	12482	5946	14342	14721	15437
25 and over	15645	12194	5213	15930	13973	15217
PROFESSIONAL STATUS (a)						
No more self-employed	7671	5868	1852	12235	7043	7543
Co.Co.Co.	10250	9870	4606	9095	9722	10022
Entrepreneurs	22622	22231	7309	18190	20442	22006
Professionals	21146	21370	7696	20971	19650	21151
Artisans, shopkeepers etc.	13195	12695	5075	13146	11925	13140
Co-helpers	13436	12161	6176	10025	12416	12665
Coop. stockholders	10127	10107	5202	12741	9303	10380
HOURS WORKED PER WEEK						
No more self-employed	7671	5868	1852	12235	7043	7543
Up to 19	9305	6533	-1701	7868	8033	8280
20-29	9593	7702	3965	8869	8837	8972
30-39	12014	11193	6317	11683	11226	11770
40-49	15400	14629	5736	15183	14065	15242
50 and over	17814	14400	6325	16170	16134	17225
TOTAL	14569	11431	5442	14363	13315	13959

(a) At the moment of interview

The effects of imputation have been controlled comparing the distribution of self-employment incomes before and after the process (that is, all the percentiles, by comparing the cumulated frequency distributions). Also, the imputation is evaluated comparing the main distribution parameters (average, median, first and third quartiles, minimum, maximum, standard error) broken down by gender, age, geographical area, education, type of self-employment, profession and other important covariates. For the 2004 Italian edition of EU SILC, by comparing the frequency distributions, it turns out that the imputation process has led to a substantial change in the lower tail, while leaving almost unchanged the rest of the distributive profile (Figure 2). In fact, after the imputation, the increase in the density of incomes lower than 10,000 euros per year is evident, whilst the rest of the curve shows minor 'local' changes, as the other unrealiable or missing incomes were repositioned in a rather homogeneous way in the central part of the distribution.

The imputations are fully consistent with the previous findings about the distribution of the missing and unreliable answers (Table 2). Because of the higher non-response rates in marginal (*i.e.* low-income) self-employment activities, the imputation of the missing values increases the density of the left tail and, therefore, results in a lower mean income (11,431 euro per year compared to the 14,569 euros of the valid, non-missing data). At the same time, the imputation of the unreliable answers (and of the few *outliers*) has almost tripled their mean income (from 5,442 euro per year to 14,363), being these cases mainly related to more solid employments. Overall, the composition of both effects has led to a slight increase in the average income from self-employment: from 13,315 euros a year before the imputation to 13,959.

It is interesting to note that the variability of income is substantially preserved after the imputation: the coefficient of variation grows slightly, from 104.0% to 105.3%.

4. COMBINING SURVEY AND ADMINISTRATIVE DATA ON SELF-EMPLOYMENT INCOMES

The availability of tax data on self-employment income permits an innovation in the production of statistics on this subject, based upon an integrated use of the information derived from both administrative and survey data. Obviously, by combining data from different sources more information becomes available than that provided by each source separately. From a technical point of view, the integration of the administrative source with the EU SILC 2004 survey data has required the implementation of the following steps:

Key (individual identifier)

Each sample person has been identified with her/his tax code (*i.e.* the personal identification number assigned to each individual by the Italian tax authorities). The tax codes have been primarily retrieved out of the Population Registers by the Statistical Offices of the Municipalities who partecipated in the survey. As the information released by the local statistical offices may be missing or inaccurate, the Istat has also requested them to collect auxiliary data on the individuals to be interviewed. Indeed, the personal tax code is univocally determined by the values of selected individual characteristics (name, surname, sex, birth date and birth place). Thus, the collected tax codes were compared with those resulting from the computation based on the avaliable individual characteristics and, when opportune, corrected. In the 2004 Italian EU SILC, the total number of sample persons aged fifteen and over is 67,743. Of these, 64,175 have a tax code associated in the Population Registers or correctly computed from the other available information (a coverage of 94.7%). The missing tax codes correspond to cases when both the collected tax code and the detailed personal information are incomplete, missing or inconsistent between them.

Linkage of survey and tax records

In a second step, the tax codes of the previous phase were matched to those contained in the Personal Tax Annual Register, embracing all the Italian tax codes[15]. The procedure searched for the tax codes of the 67,743 sample persons among the ones encompassed in the tax files. In particular, the fundamental aim of the linkage concerned the 52,509 adults (15 years and over) that actually participated in the survey. Among

[15] Each resident in Italy has her/his own tax code, attributed in the first months after birth. Therefore, the actual taxpayers constitute a subset of the population provided of a tax code.

these latter, the number of successfully matched records was 49,202 (Table 3). In other words, the tax source covers the 93.7% of the adults interviewed for the 2004 Italian EU SILC survey. The 3,307 unmatched units (6.3%) are either individuals with no tax code available in the Population Registers (4.4%) or persons not included in the initial survey frame but later registered as additional household's members by the interviewers (1.8%).

Table 3 Main results of the linkage between tax and survey records

	SURVEY DATA			
	Sampled	%	Interviewed	%
Linked with tax records	64,175	94.7	49,202	93.7
Not linked	3,568	5.3	3,307	6.3
TOTAL	67,743	100	52,509	100

Loading the tax data

The third step consisted in reading and checking the information on self-employment incomes included in the tax records. At this stage, two relevant sources of microdata have been uploaded: (i) the "*UNICO persone fisiche*" and: (ii) the "*730*" tax returns[16]. Once implemented, the reading procedures lead to a suitable database of tax records that has been used to build the net (taxable) self-employment income. The Italian tax system distinguishes between two broadly defined categories of self-employment incomes: the '*redditi da libera professione*' (earnings from free profession) and the '*redditi d'impresa'* (business incomes). The latter may also include incomes attributed to sham co-helpers for tax splitting purposes. Income splitting within a small family business occurs when there is a transfer of taxable income from a person in a higher income bracket to a person in a lower income bracket. Because of the progressive tax schedule, it is thus possible for some self-employeds to lower total household income liability. We define as 'sham co-helpers' the persons who appear as percipients of self-employment income in the tax returns but, at the same time, convincingly report themselves in the survey as inactive, non-working persons during the income reference period (students, housewives etc.). The income received by sham co-helpers has been assigned to the active self-employed household members.

Identifying the percipients of self-employment incomes

Of the 7,637 individuals identified as percipients of self-employment incomes in the survey, 7,168 were successfully linked with the corresponding tax records (93.9%). Among these matching records, 5,219 correspond to persons who filed a tax return with taxable self-employment incomes, whilst 2,071 individuals who are percipients of self-employment incomes in the survey did not file a tax statement (Table 4).

Table 4 Self-employment incomes in the tax and survey sources, by result of the linkage and content of the two datasets

[16] With few exceptions, the '*UNICO persone fisiche*' form must be filled by the generality of percipients of self-employment incomes. In particular, by any person who is the sole or joint owner of an unicorporated business for which he/she works and by those taxpayers who perceive incomes from unincorporated businesses. The '*730*' form must be filled by the percipients of secondary and/or occasional self-employment incomes (for example, an employee who adds to the wage a self-employment income from a secondary job as a free-lancer).

		SURVEY DATA Did you earn Self-employment income in 2003?				
	TAX DATA	YES	%	NO	%	TOTAL
LINKED	*Self-empl. income reported*	4,024	52.7	1,195	2.7	5,219
	Self-empl. income not reported	1,073	14.1	4,821	10.7	5,894
	No tax return	2,071	27.1	36,018	80.3	38,089
NOT LINKED		469	6.1	2,838	6.3	3,307
TOTAL		7,637	100.0	44,872	100.0	52,509

Another group of 1,073 matched records contain self-employment earnings in the survey but only capital income and/or rents in the tax files. For these cases, it has been assumed that the professional status as a self-employed during the income reference period and the related survey income are reliable. In fact, it has proved impossible to ensure that the capital incomes of these 1,073 individuals in the tax dataset were comparable with the self-employment incomes reported in the survey. Therefore, to avoid errors due to the ambigous definitions of the Tax Code, for this subset of records the two sources have not been compared and the survey incomes have been retained in the final dataset. On the other hand, the tax source also includes 1,195 percipients of (labour-type) self-employment incomes that have not been reported in the survey. In these cases, the definition underlying the tax data is comparable with the one adopted for the EU SILC questionnaire and the tax records have thus been retained in the final dataset. Interestingly, almost all the individuals reporting self-employment incomes only in the tax returns report themselves in the survey as persons no more engaged in self-employment activities at the time of the interview[17]. These interviewees (notably, the pensioners) are especially inclined to disregard their past incomes.

In sum, whilst only 4,024 records display self-employment incomes in both datasources, the integrated data contains 8,832 percipients of self-employment incomes. Indeed, 3,613 percipients are present only in the survey dataset (469 of which because of failures of the record linkage) and 1,195 only in the tax records.

Self-employment incomes in the integrated dataset

The assumption underlying the fourth step has been that true disposable self-employment income may be under-reported by both sources. In order to minimise under-estimation, self-employment income has been set to the maximum value between the net income resulting from the tax source and the net income reported in the survey. In most cases, the comparisons of the self-employment incomes reported in the two sources has been made at the individual level. However, for the small family businesses, the comparison has been made at the household level, that is by comparing the sums of the self-employment incomes received by all household members in the two sources. The 8,832 percipients of self-employment incomes considered here are the ones for which the choice of the maximum value between tax and survey incomes at the individual level is consistent with the analogous comparison made at the household level. There are, however, also 80 small family businesses that have undergone a special treatment because of the presence of pretended co-helpers. For these latter households, in fact, the comparison between tax and survey self-employment incomes has been made solely at the household level and, subsequently, the allotment of total household income to the members has ignored the sham co-helpers. In such circumstances, the comparison of the two datasources at the individual level may not be fully consistent with the rule adopted for the generality of cases, as the sham co-helpers have been ignored[18].

Table 5 presents the main results of the linkage. The final (merged) data can be divided according to the original source in four distinct groups: 1) incomes present solely in the tax records; 2) incomes present only in the survey data; 3) incomes coming from tax records, being greater than the corresponding survey incomes; 4) incomes corresponding to cases when the survey amounts are greater than the corresponding net income in the administrative source[19]. The results for the integrated data are reported in the last four columns. For each group, the table displays mean income and the number of percipients broken down by

[17] This subgroup also includes employees who have earned self-employment earnings in the income reference period from temporary/secondary jobs.

[18] For the sake of simplicity, these particular cases have been excluded from the tables and will not be further considered in what follows.

[19] The third panel of the table also reports about the main characteristics of the two datasets before the linkage (*i.e.* considering all their records).

individual characteristics. The 1,195 records coming exclusively from the tax file reduce the overall mean income. In effect, the mean of the self-employment incomes present solely in the tax records is lower (9,113 euros) than the average of all survey incomes (13,959 euros).

For 3,613 percipients, self-employment income is reported solely in the survey. The mean income of this group is 11,627 euros, again a value lower than the average income computed on the whole of the survey data. As already noticed, the majority of this group is made of taxpayers who filed a tax return without reporting self-employment incomes or did not file a tax return. The presence of 841 subcontractors (Co.Co.Co) in this group is not surprising, as most of them can fill the same tax form used by the employees (the '*CUD*'). About 500 percipients are inactive individuals (at the time of the interview), who earned on average 524 euros per month in the income reference period. For the artisans, the shopkeepers and the other self-employeds included in this group, it is impossible to distinguish, on the basis of the available information, between the percipients who totally avoid taxes and those who are missing for some other reason. In fact, for the 2004 edition of the Italian EU SILC, only the tax forms for self-employment incomes (the '*730*' and the '*UNICO*') have been made available. The results of the record linkage indicate that some self-employment incomes should be found in the '*CUD'* form, together with (dependent) employment incomes and pensions. For example, the fees earned by the members of advisory committees, in addition to a wage or a pension, could be reported in the *'CUD'* tax form, even though they are self-employment incomes. Another issue that deserves further analysis concerns some percipients who report themselves in the survey as professionals (*e.g.* doctors, engineers…), even though they are dependent workers (*e.g.* in a hospital, in the public administration…). The secondary self-employment incomes of these individuals can sometimes be found in the '*CUD'* form. This form will be used for the 2005 edition of the Italian EU SILC.

Among the group of percipients who have reported self-employment incomes in both sources, 1,612 display an higher amount in the tax data, whilst 2,412 persons reported a larger income in the survey data. For what concerns the former group, mean income of the (selected) tax data (25,425 euro) is approximately twice that of the (discarded) survey incomes of the same group (12,763 euro). Similarly, among the group of percipients with higher survey incomes, the mean of the (selected) survey income (18,253 euros) is nearly twice as great as the mean income of the (discarded) net taxable incomes (8,765 euros).

The last panel of Table 5 shows the change in the number of earners and in mean self-employment income due to the integration of datasets. With respect to the exclusive use of survey data, the linkage with tax records has increased the earners by 15.6% and the average income by 11.9%. A remarkable result concerns the large number of elderly who report self-employment incomes in the tax data but do not in the survey. The addition of the incomes of these elderly earners who are present solely in the tax data lowers the average self-employment income for this sub-group of individuals.

Table 5 Self-employment incomes in the survey, tax and final datasources, by individual characteristics at the time of the interview

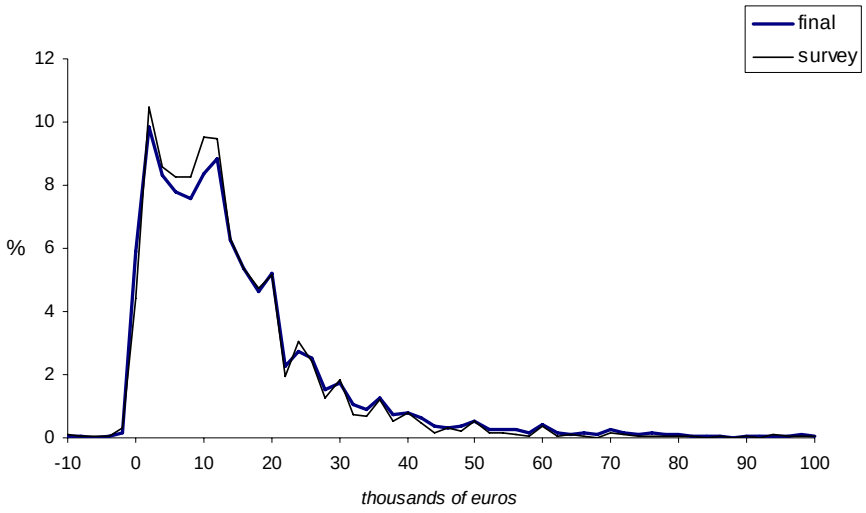
INDIVIDUAL CHARACTERISTICS:	UNITS WITH S.E. INCOMES IN ONLY ONE SOURCE				UNITS WITH S.E. INCOMES IN BOTH SOURCES						TOTALS FOR EACH SOURCE (before merging)				TOTALS FOR THE INTEGRATED DATA (after merging)			
	Only tax data		Only Survey data		Tax data > Survey data			Survey data ≥ Tax data			Survey data		Tax data					
	Mean of tax incomes	Number of cases	Mean of survey incomes	Number of cases	Mean of survey incomes	Mean of tax incomes	Number of cases	Mean of survey incomes	Mean of tax incomes	Number of cases	Mean of survey incomes	Number of survey records	Mean of tax incomes	Number of tax records	**Mean of final* incomes**	**Number of final* records**	*% (C-A) /A*	*% (D-B) /B*
											(A)	**(B)**			**(C)**	**(D)**		
GEOGRAPHICAL AREA																		
North	8415	576	12921	1776	13882	27727	907	19516	9789	1322	15316	4005	15307	2805	**17189**	**4581**	*12.2*	*14.4*
Center	9897	337	11973	873	12908	24823	397	18104	8020	587	14111	1857	13548	1321	**15620**	**2194**	*10.7*	*18.1*
South and Islands	9602	282	8928	964	9280	19423	308	15107	6945	503	10740	1775	11147	1093	**12103**	**2057**	*12.7*	*15.9*
SEX																		
Male	9311	677	13323	2171	14245	28012	1123	19647	9597	1710	15691	5004	15433	3510	**17652**	**5681**	*12.5*	*13.5*
Female	8855	518	9073	1442	9359	19484	489	14859	6740	702	10669	2633	11028	1709	**11942**	**3151**	*11.9*	*19.7*
AGE																		
15-34	7987	362	10127	1255	9342	20145	378	14400	6702	583	11117	2216	10894	1323	**12262**	**2578**	*10.3*	*16.3*
35-44	9904	258	12139	946	13137	25108	493	18372	9032	775	14543	2214	14373	1526	**16446**	**2472**	*13.1*	*11.7*
45-64	10113	382	12702	1253	14173	28092	685	20521	9673	958	15637	2896	15987	2025	**17902**	**3278**	*14.5*	*13.2*
65 and over	8189	193	11934	159	15300	31241	56	18060	10088	96	14431	311	12459	345	**13812**	**504**	*-4.3*	*62.1*
EDUCATION																		
Primary school	9899	259	8892	633	10444	20550	200	15695	7357	345	11148	1178	11457	804	**12329**	**1437**	*10.6*	*22.0*
Intermediate school	9916	260	10384	1091	10185	21279	496	16927	8025	683	12309	2270	12935	1439	**14238**	**2530**	*15.7*	*11.5*
High school	8675	425	13229	1369	12654	24615	645	17817	8472	978	14604	2992	13598	2048	**16125**	**3417**	*10.4*	*14.2*
Univ. Degree, post-graduated, etc.	8213	251	13344	520	19450	38541	271	23709	11915	406	18242	1197	18689	928	**20077**	**1448**	*10.1*	*21.0*
YEARS OF ACTIVITY																		
1-5	8077	133	9030	383	6693	16696	56	13010	5537	121	9656	560	8643	310	**10161**	**693**	*5.2*	*23.8*
6-10	8382	287	10130	767	10441	20752	256	15038	7088	361	11468	1384	11368	904	**12518**	**1671**	*9.2*	*20.7*
11-15	9578	189	11863	712	12993	27873	296	17848	8790	504	14079	1512	14652	989	**16169**	**1701**	*14.8*	*12.5*
15-24	11817	208	13152	756	13256	24982	460	19471	9280	677	15437	1893	15043	1345	**17646**	**2101**	*14.3*	*11.0*
25 and over	8313	378	12451	995	13937	27566	544	19821	9613	749	15217	2288	15164	1671	**17019**	**2666**	*11.8*	*16.5*
PROFESSIONAL STATUS																		
No more self-employed	9133	1190	6292	499	7205	19474	108	11806	2370	155	7543	762	9180	1453	**9191**	**1952**	*21.8*	*156.2*
Co.Co.Co.	4298	3	9684	841	8881	21088	53	14224	4791	82	10022	976	11040	138	**10665**	**979**	*6.4*	*0.3*
Entrepreneurs	.	.	22823	279	17702	42122	123	23584	9124	191	22006	593	22050	314	**27071**	**593**	*23.0*	*0.0*
Professionals	.	.	18045	337	19991	35885	360	23791	12826	555	21151	1252	21899	915	**25722**	**1252**	*21.6*	*0.0*
Artisans, shopkeepers etc.	4357	2	11379	1168	10444	20435	854	16626	8241	1250	13140	3272	13182	2106	**15740**	**3274**	*19.8*	*0.1*
Co-helpers	.	.	12090	136	10766	20266	23	16475	6531	32	12665	191	12274	55	**13809**	**191**	*9.0*	*0.0*
Coop. stockholders	.	.	9460	353	8608	19201	91	13686	6874	147	10380	591	11587	238	**12011**	**591**	*15.7*	*0.0*
HOURS WORKED PER WEEK																		
No more self-employed	9133	1190	6292	499	7205	19474	108	11806	2370	155	7543	762	9180	1453	**9191**	**1952**	*21.8*	*156.2*
Up to 19	.	.	6769	173	7505	17455	45	15689	8049	40	8280	258	13028	85	**10016**	**258**	*21.0*	*0.0*
20-29	2000	1	7305	386	11152	19660	83	12036	5459	151	8972	620	10460	235	**10098**	**621**	*12.6*	*0.2*
30-39	.	.	10118	491	11493	20851	158	15747	7856	215	11770	864	13361	373	**13482**	**864**	*14.5*	*0.0*
40-49	5315	3	13420	1178	12647	25486	640	19386	9446	919	15242	2737	16010	1562	**18230**	**2740**	*19.6*	*0.1*
50 and over	3663	1	15914	886	14916	29170	578	19904	9934	932	17225	2396	17288	1511	**20657**	**2397**	*19.9*	*0.0*
TOTAL	**9113**	**1195**	**11627**	**3613**	**12763**	**25425**	**1612**	**18253**	**8765**	**2412**	**13959**	**7637**	**13991**	**5219**	**15615**	**8832**	***11.9***	***15.6***

Note (*) : The term "final" refers to the results of the integration between survey and tax data-sources.

5. THE DISTRIBUTION OF SELF-EMPLOYMENT INCOMES IN THE ITALIAN EU SILC

The use of administrative data has changed the tails of the distribution of self-employment incomes (Figure 3). Indeed, with respect to the survey data, the final (*i.e.* integrated) dataset contains a lower percentage of self-employment incomes in the range 2,000 - 12,000 euros per year and an higher proportion of percipients with incomes greater than 20,000 euros.

Figure 3 Distribution of self-employment incomes in the survey and in the final dataset *(all percipients)*



The concentration of self-employment income is different, too (Table 6). The Gini index shows that survey data are characterised by much less inequality (0.48) than the tax data (0.59). In the final dataset the Gini measure of inequality of self-employment incomes amounts to 0.50. The decomposition of the Gini index by sub-groups of percipients, precisely by their professional status at the time of the interview, reveals that the final data encompass an higher degree of inequality 'between groups' than the two datasources taken separately. Moreover, after the integration, the 'between groups' component has an higher importance in explaining overall inequality.

Table 6 Gini index decomposed by subgroups of percipients of self-employment income
(all the self-employment incomes available in each source)

	SURVEY DATA			TAX DATA			FINAL DATA		
Overall Gini	0.48		*100%*	0.59		*100%*	0.50		*100%*
- between groups	0.17		*36%*	0.18		*30%*	0.21		*41%*
- within groups	0.10		*22%*	0.12		*21%*	0.09		*18%*
- crossover	0.20		*42%*	0.28		*49%*	0.21		*41%*
	group specific Gini	share of population	share of income	group specific Gini	share of population	share of income	group specific Gini	share of population	share of income
Employees	0.59	4.6%	2.4%	0.69	11.6%	7.7%	0.61	9.6%	5.6%
Enterpreneurs	0.45	7.8%	12.2%	0.60	6.0%	9.5%	0.46	6.7%	11.6%
Professionals	0.43	16.4%	24.8%	0.54	17.5%	27.4%	0.45	14.2%	23.4%
Artisans/shopkeepers...	0.43	42.8%	40.3%	0.52	40.4%	38.0%	0.41	37.1%	37.4%
Co-helpers	0.49	7.7%	5.8%	0.52	4.6%	3.8%	0.48	6.7%	5.1%
Coop. stockholders	0.41	2.5%	2.3%	0.53	1.1%	0.9%	0.42	2.2%	1.9%
Co.co.co.	0.47	12.8%	9.2%	0.61	2.6%	2.1%	0.47	11.1%	7.6%
Unemployeds	0.64	1.4%	0.7%	0.67	2.5%	1.0%	0.57	2.2%	1.0%
Other inactive	0.59	4.0%	2.4%	0.63	13.8%	9.6%	0.60	10.2%	6.4%

In fact, with respect to the survey data, the tax file includes an higher proportion of percipients of secondary ('employees') and of marginal/temporary ('unemployeds', 'other inactive') self-employment incomes[20] as well as larger shares of the corresponding incomes. Furthermore, in both sources (and in the final data as

[20] In the previous sections, these percipients have been collectively referred to as 'No more self-employeds' (at the time of the interview). Actually, there are also persons who were not full-time self-employeds in the income reference period, too.

well), these sub-groups are the ones with the highest degree of inequality. More generally, the self-employment incomes of all the sub-groups of percipients are more unequally distributed in the tax datasource than in the survey.

In the final data, the majority of retained records for the sub-groups made of 'employees', 'unemployeds' and 'other inactive' come from the tax datasource, while the opposite is true for all the other categories of percipients, namely for those who are self-employed at the time of the interview (Table 7).

Table 7 Sources of self-employment incomes in the final dataset, by subgroups of percipients
(all the percipients of self-employment incomes in the final dataset)

	survey	tax	all
Employees	36.4	63.6	100.0
Enterpreneurs	79.3	20.7	100.0
Professionals	71.2	28.8	100.0
Artisans/shopkeepers...	73.9	26.1	100.0
Co-helpers	84.6	15.4	100.0
Coop. stockholders	88.0	12.0	100.0
Co.co.co.	94.3	5.7	100.0
Unemployeds	44.7	55.3	100.0
Other inactive	28.3	71.7	100.0
All	**68.2**	**31.8**	**100.0**

A closer look at the results of the linkage permits to conclude that both datasource miss a substantial amount of information. More precisely, of all the percipients of self-employment incomes in the integrated dataset, the 40.9% would have been ignored by using exclusively the available tax records, whilst the 13.5% do not reveal themselves as percipients of self-employment incomes in the survey (Table 8).

Table 8 Percipients of self-employment incomes in the integrated dataset, by content of the two datasources
(All the percipients in the integrated dataset)

	TAX DATA		**SURVEY DATA**			**FINAL DATA**
	reported	not reported	observed	missing (imputed)	NO to S.E. question	
Employees	71.0	29.0	26.9	14.5	58.6	100.0
Enterpreneurs	53.0	47.0	79.8	20.2	none	100.0
Professionals	73.1	26.9	80.0	20.0	none	100.0
Artisans/shopkeepers...	64.3	35.7	76.1	23.9	0.1	100.0
Co-helpers	40.3	59.7	49.6	50.4	none	100.0
Coop. stockholders	28.8	71.2	54.5	45.5	none	100.0
Co.co.co.	14.1	85.9	50.3	49.4	0.3	100.0
Unemployeds	66.0	34.0	32.5	21.3	46.2	100.0
Other inactive	79.5	20.5	21.8	11.8	66.4	100.0
All	**59.1**	**40.9**	**60.5**	**26.0**	**13.5**	**100.0**

APPENDIX: 'THE SURVEY QUESTIONNAIRE ON SELF-EMPLOYMENT INCOMES'

Q1 In < *year* >, did you earn or lose any money from self-employment?

NO…… ☐ → *Interviewer: skip to next section*
YES..... ☐

Q2 Thanks to the money earned from self-employment, in < *year* > have you been able to...
Interviewer: read all answers, tick a box per each row

	NO	YES
pay, either partly or totally, for personal and household expenses of any kind…….	☐	☐
save, that is to put money aside (in a bank, purchasing financial assets, invested in the business, in the stock market etc.) …….	☐	☐
buy or renovate a dwelling, a building, a land for you or your household ……..	☐	☐
repay, either partly or totally, a personal or family debt ……	☐	☐
pay alimonies for a former spouse and/or for separated children ……..	☐	☐
give financial support or lend money to relatives, friends, other people……	☐	☐
Other ways to use the money (*specify)*:		
_____	☐	☐

Q3 In < *year* >, what has been your income from self-employment, net of all taxes and social security contributions?

€ |__|__|__|,|__|__|__| /00 → *Interviewer: go to next section*

I don't know…….. ☐ → *Interviewer: go to next question (Q4)*
It has been a loss……☐ → *Interviewer: go to Q5*

Q4 Can you please tell me which of the following amounts is closer to your self-employment income in < *year* >?

More than 90,000 euros ☐
80,000 euros ☐
70,000 euros ☐
60,000 euros ☐
50,000 euros ☐
40,000 euros ☐
35,000 euros ☐
30,000 euros ☐
28,000 euros ☐
26,000 euros ☐
24,000 euros ☐
22,000 euros ☐
20,000 euros ☐
18,000 euros ☐
16,000 euros ☐
14,000 euros ☐
12,000 euros ☐
10,000 euros ☐

8,000 euros ☐
6,000 euros ☐
4,000 euros ☐
3,000 euros ☐
2,000 euros ☐
1,000 euros ☐
500 euros ☐
Less than 500 euros ☐

Q5 In < *year* >, what has been your loss?

€ |__|__|__|,|__|__|__| /00 → *Interviewer: go to next section*

I don't know…….. ☐ → *Interviewer: go to next question (Q6)*

Q6 Can you please tell me which of the following amounts is closer to the loss?

More than 90,000 euros ☐
80,000 euros ☐
70,000 euros ☐
60,000 euros ☐
50,000 euros ☐
40,000 euros ☐
… … … … …

REFERENCES

Canberra Group (2001)
Final Report and Recommendations, Ottawa

ILO (2003)
Resolution Concerning Household Income and Expenditure Statistics, Geneva

ISTAT (2005)
Seminario: 'Dalla progettazione alla realizzazione della nuova indagine campionaria sui redditi e le condizioni di vita delle famiglie', Rome

ISTAT (2006)
Seminario: 'Strategie metodologiche dell'indagine campionaria sui redditi e le condizioni di vita delle famiglie', Rome