Multi-source data collection strategy and microsimulation techniques for the Italian EU-SILC[1]

Paolo Consolini[2], Gabriella Donatiello[3]

Abstract

This chapter presents the multi-source data collection strategy that has been developed at the Italian National Institute of Statistics since 2004 for the EU-SILC project with a focus on the integration methodology that has been implemented to build net and gross income target variables. The first part of the paper describes the imputation and correction processes carried out by Istat to obtain the final income variables. The second part of the study explains the complex and innovative methodology devised to setup and use a microsimulation model when multiple integrated data sources are available, a task that goes far beyond the traditional "gross to net" (or "net to gross") conversion of survey incomes. The results show that combining microsimulations with integrated survey and administrative data definitely enhances data quality.

Keywords: Administrative Data, Survey Data, Data Integration, Microsimulation, Income, Multi-mode data collection, Record linkage.
C810 - Methodology for Collecting, Estimating, and Organizing Microeconomic Data

1. Introduction

The Italian SILC survey (EU-SILC) is based on the "Computer Assisted Personal Interviewing" method of collecting data and uses administrative microdata in order to reduce measurement errors. Many researchers, including statisticians, psychologists, sociologists and economists, share common concerns about the weakness of the measurement process in the survey method. As is well known, errors can be due to any of the many factors influencing the measurement process: the questionnaire, the respondent and the interviewer, as well as the data collection method. The structure and the wording of the questions affect the interpretation by the respondent. Even when the interviewee fully understands a question, he could still have memory problems in giving a reliable answer.

Understanding and memory problems often lead to measurement errors: omissions and recall errors being typical examples. In order to limit the possible bias on the income

[1] Paragraphs 2, 2.1, 2.2 have been drafted by Paolo Consolini; paragraphs 3, 3.1, 3.2 by Gabriella Donatiello and paragraphs 1 and 4 by both authors. The opinions are those of the authors and do not imply any responsibility for the Istat (Italian National Institute of Statistics).

[2] Senior Researcher (Istat), e-mail: consolin@istat.it
[3] Senior Researcher (Istat), e-mail: donatiel@istat.it

reported in the questionnaire by the interviewees and to improve the general data quality of the survey, a project of multi-source data collection has started at Istat since 2004. The integration technique used to combine survey and administrative microdata to produce the EU-SILC income target variables, can be viewed as a flow process, starting from the analysis of the income definitions adopted by the different data-sources, developing through the choice of the best matching key and the more effective record linkage methodology, followed by a consistent, problem-solving, approach for the harmonizing of units and variables, the handling of inconsistencies and of under/over coverage of the integrated data sources (survey, administrative, imputed, microsimulated) to end up with the reconciliation of values reported in the different sources with the final set of income target variables of the EU-SILC project. For the first Italian edition of the project (2004), the process involved only two 'problematic' income components: self-employment income and pensions. From the second edition (2005) onward it includes employment incomes, too.

In the following all the steps of the integration process will be analysed, focussing on the solutions adopted to handle the problems arising from the integration of different data sources (harmonization of units and definitions, incoherencies of income sources, reconciliation of inconsistent income amounts). At the same time, the impact that the data integration and editing process has on the final values of the income components will be provided and discussed. To sum up, the administrative data are used to support the editing and imputation processes and to ease the construction of gross incomes with microsimulation techniques.

According to EU Regulation, in Italy the estimation of gross income statistics became mandatory starting from survey year 2007. A microsimulation model which estimates taxes and social insurance contributions for the income reference year is one of the most traditional technique used for the net-gross conversion of income variables. However, Istat decided to setup a new methodology based on the contemporary use of the Siena microsimulation model (SM2-EuSilc) and of a record linkage between survey and administrative data from multiple sources. The available administrative data in terms of net incomes, tax credits and income deductions have been utilized together with survey data as inputs for the SM2-EuSilc model. Administrative data have also been used when appropriate as benchmarks for the microsimulation results. In fact, administrative data and microsimulation estimates are jointly considered for reciprocal comparison and validation and for the construction of the final data set of gross incomes at the individual and household levels. Some significant outputs are finally compared and validated with external sources, mainly taken from National Accounts.

2. The record linkage of administrative and survey data for the italian EU-SILC

The Italian SILC team has developed an innovative strategy in the measurement of self-employment incomes since 2004. This strategy consists in a multi-source data collection, based on personal interviews (PAPI-CAPI) and on the record linkage of administrative with survey data. The term record linkage has been used to indicate the bringing together of two

or more separately recorded pieces of information concerning a particular individual or family[4]. The commonly way to combine administrative and survey data is by selecting an individual matching-key able to link the same unit among different data-sources. In other words, the integration of administrative and survey data at micro level is performed by linking individuals through common key variables. The aim of combining administrative and survey data is to improve data quality on income components (target variables) and relative earners by means of imputation of item non-responses and reduction of measurement errors. In addiction matching tax returns records with survey data also provides information at micro level on social security contributions, taxable incomes and tax liabilities. This information is used to measure the gross/net taxable income and represent the input for the SM2 microsimulation model. For the first EU-SILC edition (2004), the integration process involved only two income components: self-employment income and pensions. The following editions also include an integration procedure of information on employment incomes in the tax and survey data sources.

The target population of the EU-SILC survey is the Italian resident population: all private households and their current members residing in Italy at time of data collection. Persons living in collective households and in institutions are excluded from the sample.

The analysis units are adult members (aged 16 and more) living in private households[5].

2.1 The measurement of income components

With regards to the measurement of self-employment incomes in household surveys there are two clear-cut statements, taken from the "Canberra Handbook", that depict the state of the art: "Income data for the self-employed are also generally regarded as unreliable as a guide to living standards"; "Household surveys are notoriously bad at measuring income from capital and self-employment income".

The alternative sources of microdata on earnings from self-employment may not contain the variable 'disposable income'. Survey data may be affected by under-reporting. On the other hand, administrative data gathering individual tax returns do not take account of illegal tax evasion and may not display all the authorized deductions allowed in the calculation of taxable income (tax avoidance). In general, neither taxable income is identical to gross income, nor net taxable income is identical to disposable income. 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 abatements allowed are explicitly shown in the tax returns. By definition, tax evasion is also not available in the tax files.

In the EU-SILC project, the standard procedure to measure net self-employment income 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. However, both sources may be affected by under-estimation of self-employment incomes.

[4] Newcombe 1995.

[5] Until 2010, in IT-SILC survey were also interviewed people aged fifteen-year-old.

Moreover, some individuals report self-employment incomes in only one data source. 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 recipients of small and/or secondary self-employment incomes[6].

Regarding the measurement of income from pensions it is assumed that the administrative data provide more accurate information respect to the survey data. The latter data source is used only if it is impossible to match the sample units to those contained in the Personal Tax Annual Register (unmatched units).

The integration of the administrative sources on pensions and pensioners needs an harmonization of units, definitions and variables and the reconciliation of the incoherencies between the income amounts reported in the different sources. Table 2.1 reports the most relevant meta-information on pension for each administrative data-source. It is noticeable that in most cases it is possible to estimate the final EU-SILC target variables only by bringing together two or more separate pieces of information recorded in different sources.

For example, in order to reckon the net amount received by the elderly separately for each different category of pensions included in the list of target variables, both the "yearly net taxable income from pensions" (Tax Register) and the "monthly gross payments" (Italian Social Security Agency) have to be broken down by kind of pension (employment, early retirement, survivors) and, moreover, to be consistent with the answers given by the respondents, when these are reliable.

The Pension Register collects information at the individual level on the relative beneficiaries, the monthly amount before tax, the classification according to EU-SILC target variables. On the other hand, the Tax Registers record the information on yearly gross/net incomes received by each pensioner without any distinction between the different categories of pensions and is not necessarily consistent with the target variables, namely when a particular kind of pension is tax exempt. In order to join the information of the Tax Registers with the Pension Register we need to define a "harmonized definition of pension income" that is comparable between all these data sources and the EU-SILC project. The common base for the comparison is represented by the "taxable income relating to pensions"[7].

The measurement of employee income is based on the comparison of administrative and survey data on wages and salaries after retention of taxes on labour and mandatory social security contributions at source. The main administrative data source for this income component is represented by the CUD/770 tax statements register. In Italy the employers, as withholding agents, are obliged to declare the net amounts of wages/salaries and of taxes and social contributions annually paid to and for their employees. However, the items included in employee income considered by the administrative source are not exactly the same of the target variable PY010N/G (employee cash or near cash income), therefore it is necessary to reallocate some of them in a proper way.

The administrative net income is obtained as net taxable employee income less taxes and social contributions retained at source. This aggregate is thus compared with the net employee income reported by the respondents in the questionnaire.

[6] For a more detailed analysis of this subject it is advised to see Consolini et al. (2006) and Di Marco M. (2006).
[7] See, for more details, Consolini P. (2008).

Table 2.1 – Meta information on pensions/pensioners by administrative sources

Data sources	Variables						Domains	Units
	Gross Income for pension		Net Income for pension		Number of payments	Pension type (Function)		
	Monthly	Yearly	Monthly	Yearly				
Pension Register (PR)	✓(a)	✓(c)	-	-	✓(c)	✓(a)	Census of pensioners of the Italian Social Security System	Pensioner and/or Pension
CUD/770 Tax Register	✓(b)	✓(a)	✓(b)	✓	✓(b)	-	All beneficiaries of taxable pensions	Pensioner
730 Tax Register	✓(b)	✓(a)	✓(b)	✓	✓(b)	-	All beneficiaries of taxable pensions (only.730 Tax Register)	Pensioner
Unico Tax Register	✓(b)	✓(a)	✓(b)	✓	✓(b)	-	All beneficiaries taxable pensions (only.Unico Tax Register)	Pensioner

(a): recorded data
(b): variables derived from the integration of data by different sources.
(c): partially estimated (new pensioners from Pension Registers 2003-2004). In the Pension Registers 2005 data are recorded.

EU-SILC also collects information on several "non-pension cash benefits" by using administrative data sources. In particular, unemployment benefits and family allowances are gathered - on a micro level - from the Inps (National Social Security Institute in Italy) database: "Employees' temporary benefits (GPT) of private sector". In order to improve the quality on non-pension cash benefits data (i.e. maternity leave, paid sick leave, etc.) new Social security's databases will be exploited in the next years.

The information on income from capital assets is collected by interviews and the final estimation of this component is typically underestimated as usually happens in income surveys. It is well-known that obtaining accurate and unbiased information on assets income or financial assets is problematic due to the reluctance of the extremely wealthy households to participate in social surveys at all and to respondents' reticence to declare the ownership of a specific asset. Currently, no administrative data are available to estimate income from capital assets or to adjust the underestimation on financial assets and related incomes.

2.2 The integration methodology

In order to carry out the integration of different databases, some basic requirements have to be satisfied by all sources involved. Namely, the statistical units must be uniformly defined in all sources (harmonisation of units), all sources should cover the same target population (completion of populations), all variables have to be defined and classified in the same way in the different sources (comparability of variables and classifications), all

data should refer to the same period or the same point in time[8]. In short, administrative data must be comparable with the EU-SILC survey data.

The technique used to link the administrative units to those in the survey sample is the exact record linkage. This results in combining information related to the same statistical unit by means of identifiers called "matching keys" to obtain an integrated set of information that is exact, in the sense that it actually refers to the matched individual, provided that each statistical unit is associated with a unique identifier not affected by errors. Different typologies of exact record linkage exist: we have used the simplest "one-to-one" relationship, where every statistical unit of a data source is associated with at most one record from the other data source[9]. Records in different data sources have been matched using the Personal Tax Number. Once that is accomplished, the identification numbers are dropped and replaced with an internal anonymous code, according to the policy of the Italian National Statistical Institute.

The integration process between survey and administrative data at the micro level can be summarized in the following three phases[10] (see also Figure 2.1):

a) Input data: the administrative archives

Tax Agency data and Social Security (Inps) data are the administrative data sources involved in the matching process. Personal tax numbers are checked and corrected and the information coming from multiple records relating to the same person is rearranged in order to avoid duplications. In practice this step consists in reading, checking and arranging the tax records' content on the three principal income components: employee income, self-employment income and pensions. At this stage, four relevant sources of microdata have been uploaded: 1) the "Pensions Register (PR)" from INPS (Italian National Social Security Agency); 2) the "CUD/770" tax statements register (of employees, temporary workers and pensioners) from National Tax Agency; 3) the "730" tax returns register from National Tax Agency (taxpayers with at least one CUD/770 tax statement), 4) the "Unico persone fisiche (UPF)" tax returns register of "self-employed" from the National Tax Agency.

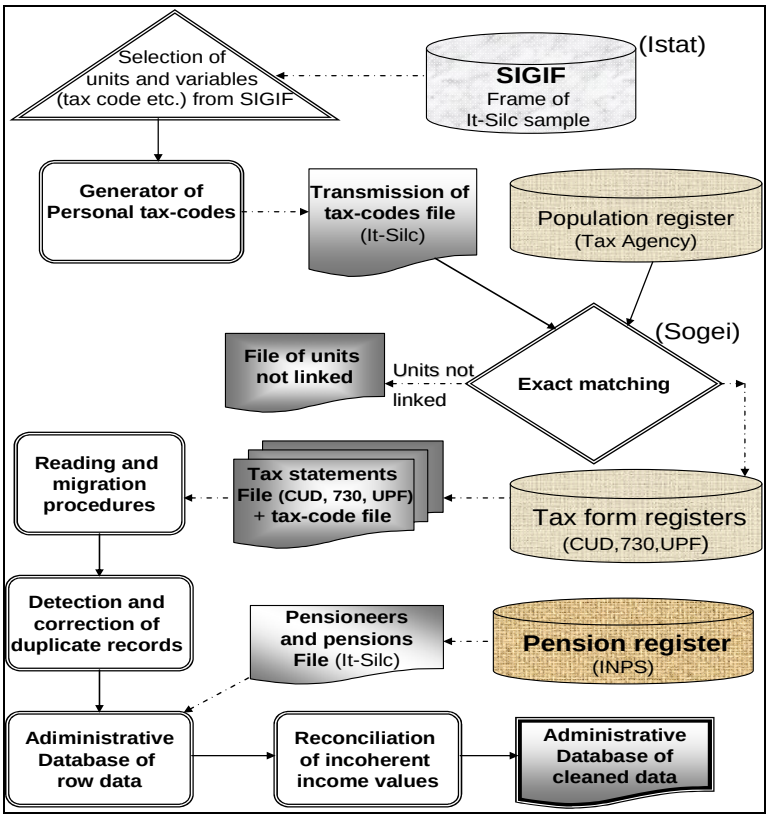
b) The exact matching procedure

At this step the survey and the administrative sources are matched using the Personal tax code number as the key variable. Each sample person is identified with her/his tax code (i.e. the personal identification number assigned to each individual by the Italian tax authorities). The output is a file (matched file) containing information on incomes both from the survey and the administrative archives. More precisely, linkage focuses mainly on adults (15 years and over) that actually participated in the survey. In 2008 the rate of successfully matched records was 96.4%. In other words, the tax source covers 96.4% of the adults interviewed for IT-SILC survey. The unmatched units (3.6%) are either individuals with no tax code available in the Population Registers (2.2%) or persons not included in the initial survey frame but registered later as additional household members by

[8] van der Laan, 2000.
[9] See Newcombe (1988), Herzog, Scheuren and Winkler (2007).
[10] See Consolini P. (2009).

the interviewers (1.4%).

Figure 2.1 - A simplified sketch of data integration process in IT-SILC



c) Detecting and solving incoherencies on income in the matched file

Sometimes the survey and the administrative data sources classify the same income of a recipient under different names. A complex system of editing rules has been established in order to choose which income component must be attributed. Similarly, analysing the coherence between administrative and survey data on the amounts of incomes that go under the same name has required a detailed editing procedure for reconciling monetary values.

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, comparisons of self-employment income reported in the two sources have been made at the individual level. However, for small family-run businesses, comparisons have been made at the household level, that is by comparing the sums of the self-employment incomes

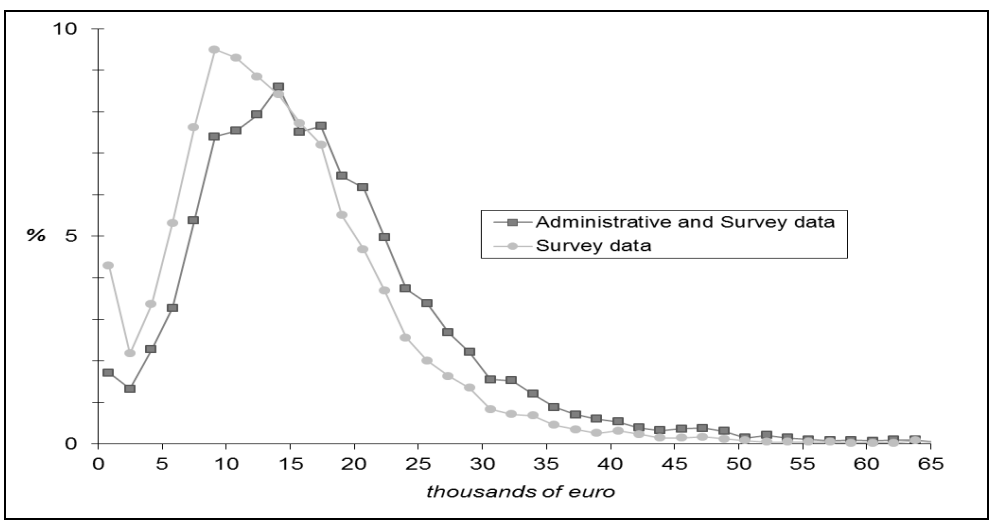
received by all household members in the two sources.

As regards the pensions, when the gross taxable pensions of the Pension Register is compared with the gross income pensions of the CUD/770 tax source, it turns out that for the 84.2% of the matched cases the relative difference between the two amounts is lower than 5%. The assumption underlying the building of the gross/net income from pensions is that the gross income reported in the Pension register is true and the proper information on the tax at source, as well as on tax credits, is included in the other administrative sources. Survey data on pensions (after retention at source) are taken in to account where it is impossible to link the administrative data.

With respect to employment income, we assume that true disposable employee income is included in the administrative source providing that employee does not receive exempt income items (like tips or bonuses) or is employed in sectors of hidden economy (like agricultural, private educational institutes, etc). The CUD/770 tax register includes 99.1% of employee income records reported in all administrative sources.

An assessment of the impact of the multi-source versus survey approach (income data collected by interview) on the equivalent income distribution has been carried out for IT-SILC 2011 edition. As displayed in Figure 2.2, the effect of the inclusion of administrative data involves a shift forward of the income curve. At first glance it seems that the adjustments produce a steady rise in the income levels across the whole survey distribution.

Figure 2.2 – Equivalised income (YEQ) distribution from survey and integrated database (Year 2011)



Merging administrative and survey data definitely brings about a rise of 28.3% in the number of recipients and an increase of 8.8% in the average of self-employment income compared to the exclusive use of survey data (Eusilc 2008 edition). When both sources report information on self-employment incomes, there is some evidence of a higher under-estimation rate on the tax data compared to the survey data. As results from data integration in Eusilc 2008, the number of employee income receivers increases of about 11% whereas

employee income increases by about 0.7%.

3. The building of EU-SILC gross income variables

For the estimation of EU-SILC gross incomes, Istat fitted the University of Siena's SM2 microsimulation model to the non-standard case of an integrated dataset from multiple sources. The Siena Micro-Simulation Model (SM2) has been adopted as a recommended procedure by the European Commission for the provision of EU-SILC gross income statistics. The first release of SM2 had been developed by Siena University team in 2003 and initially applied to the ECHP (*European Community Household Panel*) survey data. For the EU-SILC project, SM2 has been updated and extended to become a general and flexible tool for the "net to gross" and "gross to net" conversion of income variables that can simulate the functioning of the tax-benefit system of different countries[11]. In fact, the model can be applied to diverse input data collected in various forms across and within countries and it is able to generate variables in a comparable and standardised 'multi-country' format.

The model estimates income by component, breaking down gross amounts into taxes, social insurance contributions, social transfers, and net/disposable incomes. All the information on income components have to be collected, compiled or imputed in some form, and the model converts it, under a specified national tax system, to the standard form required by the EU-SILC project.

The SM2 consists of a standardised set of routines, which can deal with a great diversity of input data forms and national tax systems. Country-specific routines are required to convert the input data formats and to define the parameters of the national tax system in an appropriately standardised form. These routines are taken as inputs by the core of the model, that generates the required standardised outputs. The system maintains a clear distinction between the general and the country-specific parts, and it is developed to maximise the part which can be standardised in order to be easily applied for different tax benefit systems [12].

In 2004 Istat decided to test the application of the SM2 and use both administrative archives and sample survey data for the net-gross conversion of EU-SILC variables.

The data production process of the EU-SILC gross income variables can be summed up in three important steps: the first one is the adaptation of the model SM2-EuSilc from SM2; the second one is the integration of survey data and administrative data used jointly with microsimulations and the third one is the setup of the final dataset of individual and household income target variables.

The implementation by Istat of the SM2-EuSilc model has required the transition from the preliminary version applied to the ECHP data to the version applied to the EU-SILC data and the construction of the input and auxiliary variables on the basis of information

[11] The model was set up under the Eurostat project " Development of Appropriate Modelling or Imputation to Construct the EU-SILC Target Income Variables for Each EU Member States".

[12] For a detailed description of the model see: Betti et al, 2011.

collected by the new survey. Originally, the SM2 input file was based on the Eurostat releases of the ECHP User and Producer databases for three countries (Italy, France, Spain, see Eurostat, 2004). As regards to Italy, the income reference year was 1998 and the tax rules were those of the year 2003, in order to include the then recent tax reform.

The adaptation of the model to the new EU-SILC survey called for new procedures for the setup of the input file and implied the adjustment of some conversion routines of SM2.

The first step in the construction of the input file was a direct substitution, where possible, of the Echp variables with the new ones. The second step was the construction of the auxiliary variables based on the information available in the new survey. Several auxiliary variables were required for the input file of SM2 and particular attention was paid to the construction of the tax units. To identify the tax units at the household level, the "family procedure" used in Istat social surveys was applied. The procedure consents to classify the households on the basis of the couple and parental relationships, identifying the dependent persons and those entitled to the family tax credits.

Compared to ECHP, the new EU-SILC survey collects detailed information on sector of activity, work status, number of months in a given status and firm size: information that is useful for calculating the social security contributions for dependent and self-employed workers. Moreover, the breakdown of sickness and invalidity benefits is available in EU-SILC as well as data on pension contributions made to private insurance companies, which could be deducted from the tax base of the Italian personal income tax.

The transition to the new survey required also an adjustment of some conversion routines of SM2, in particular for the calculation of self-employment income and the estimation of the IRAP tax (regional tax on productive activities) paid by the self-employed, including the self-employed. In the ECHP survey, self-employment income was in fact collected as a gross amount, while in EU-SILC it is recorded as net income.

Additional modifications in SM2 conversion procedures were needed for the calculation of the income of the Co.Co.Co. (temporary subcontractors) which is nominally included in self-employment income, but in fact is treated as employment income. Data on this kind of workers were not available in the ECHP survey, and a variable defining the likelihood for an employee to work under a Co.Co.Co contract was estimated in SM2, using external sources. Extra amendments of SM2 procedures were needed also for the estimation of family deduction for dependent persons in order to include the second module of the IRPEF (personal income tax) reform of 2005 and for subsequent amendments to tax legislation occurred in 2007 and 2009.

3.1 A description of the Italian tax system as integrated in the model SM2-EU-SILC

The main components of the Italian tax system are summarised in table 3.1, which displays which income components are liable to social insurance contributions and income tax, respectively. Employment and self-employment incomes are subject to social insurance contributions, determined as a function of gross income (Gi), and to income taxation. Social insurance contributions are withdrawn from gross income to obtain the gross taxable income, as they are not subject to the main Italian personal income tax (Irpef).

Irpef is calculated by applying marginal progressive tax rates to increasing income

brackets and for this purpose the incomes subject to Irpef are pooled together. Specifically self-employment income is also subject to a special tax, Irap (Tax on income from production activities), determined as a function of value added, that includes gross taxable income from self-employment. This kind of 'double taxation' at a flat rate is handled in the model by simply treating it as a 'negative tax credit'.
In fact, a distinctive trait of the model is that properly defining certain ‘special deductions or tax credits’, in addition to the typical tax benefits, many complexities of a tax regimes can be incorporated into the standardised procedures. Income components which are not subject to the Irpef are automatically removed from the common pool by just specifying their 'component-specific deductions' as equal to the component's total gross taxable income (in order to exclude their contribution to net taxable income). This applies for example to tax-exempt benefits. Moreover if a component is taxed at a flat rate separately from the pool, it is possible to specify its ‘special deduction’ as equal to the component's total gross taxable income in order to remove it from the pool, and its ‘special tax credit’ as a negative quantity. In this way, the component taxed at a flat rate makes no contribution to the tax liability of the pool, but the final tax liability is automatically increased by the appropriate amount. Tax on property assets or financial capital income can be handled in a similar way.

Table 3.1 - Main components of income, tax and social insurance contributions in the Italian fiscal system (year 2011)

N	Income components	Social Insurance Contributions (Si)	Tax	Included in common pool	Component specific	
					Deduction (Di)	Tax Credits (Ci)
1	Employment income	Employer's $S_0(G_1)$ Employee's $S_1(G_1)$	IRPEF	X		$C_1(Y_1)$
2	Self-employment income	$S_2(G_2)$	IRPEF	X		$C_2(Y_2)$ $-f_2(H_2)$ "IRAP" (a)
3	Pensions		IRPEF	X		$C_3(Y_3)$
4	Property (rental and cadastral) income		IRPEF (b)	X		
5	Financial Capital income		Taxed at source (flat rate K_6)		H_6	$- K_6 \times H_6$
6	Education related benefits, Unemployment benefits		IRPEF	X		
7	Family benefits, Sickness invalidity benefits (c), Housing allowances, Any other personal benefits		Tax exempt		H_7	
8	Property value		IMU (on value of real estate)			-f8(value)

(a) Irap: Tax on income from production activities. f(..) stands for “a function of”.
(b) On total cadastral and on 85% of the rental income, if not subject to the new regime of rental income flat rate “cedolare secca”, launched in 2011.
(c) Part of the benefits are taxable.

In Italy, the incidence of social insurance contributions on income from work is different according to the source of income, occupational status and sector of activity. Employers' and employees' social insurance contributions are imposed on gross earnings

from wages. For dependent workers there are minimal and maximal amounts of contributions to be paid. These two limits depend on firm size (number of workers), sector of activity (based on the classification NACE Rev.2) and occupational status (workers, employees, executives).

Self-employed workers' social insurance contributions are divided into three main categories: for general self-employed (i.e. craftsmen or workers in commerce), agricultural self-employed, and professional persons. The social insurance contribution rates are different in these categories and apply to income brackets and they depend also on the age of the worker. There is also a common minimum and maximum base of contribution for general self-employed (in 2011, euro 14.552 and euro 71.737) and if the self-employment income is under the minimum, they have to pay as the minimum. The agricultural self-employed have to pay a fixed amount depending on their annual income brackets.The self-employed professional persons include partners in a company, and professional workers (entrepreneur or owner, assistant of a household firm) divided in two different categories: (a) professional persons not registered in any other compulsory social insurance institution[13] and (b) professional registered in any other compulsory social insurance institution who pay supplementary contributions, as well as the occasional self-employed workers, if their annual gross income is exceeding 5.000 euro. The first category also includes the PhDs or research grant recipients and the CoCoCo (temporary subcontractors) with a special status in employment that is essentially intermediate between dependent and independent employment. The CoCoCo are essentially considered self-employed, but they have particular treatment in the Italian fiscal system and their income is treated as employee's income and, for this reason, the social insurance contributions are also paid by the employer. These contributions are, however, lower than the normal ones. For the first category the social insurance contributions rate, in 2011, account for 26,72 per cent of the annual self-employment income. In the model this rate is applied for those professional persons who only have this type of income, without any other kind of incomes or pensions. For the other professional persons the 17 per cent of annual gross income is applied by the model.

3.2 The production of EU-SILC gross income statistics

The statistical production process of the EU-SILC income variables is made up of several complex phases that can be summarized in two broad sequential steps: first the construction of net income and then the production of the income before taxes and social insurance contributions. The availability of data from administrative sources, used from the stage of construction of net incomes, has enabled the joint, innovative use of the microsimulation model and administrative archives. The integration of survey data and register data in EU-SILC has the most important aim to reduce the under-estimation of incomes on the basis of available information (survey and registers). As is well known, data

[13] Professional persons registered in the compulsory social insurance institution "Gestione separata" of the Italian National Institute of Social Security.

from income tax returns could not contain information on a number of income components (untaxed incomes, incomes taxed separately or subject to withdrawal taxes) and may have problems of coverage in relation to the individuals included in a sample survey. The survey data, in turn, may be subject to withholding of information (reticence), under-reporting or inadequate representativeness of certain types of income or income recipients. The joint use of survey and administrative data enhances the advantages obtainable from the exclusive use of fiscal archives on the one hand and of microsimulation techniques on the other.

For the construction of the gross incomes variables, the "730 tax returns" and the "UPF tax returns" provide data on net and gross incomes, taxes at national and regional level, and data on tax credits[14], income deductions[15] of declarant and spouse[16]. It is worth noting that in any microsimulation model, as in the previous SM2, the income deductions and tax credits based on consumption expenditures usually needed to be estimated by regression technique based on external sources. Respect to what done in the phase of construction of net income a new integrated data set is then made with data on taxes, income deductions and tax credits. Before using all the available information (fiscal and survey data) as input file of the model SM2-EuSilc an additional procedure for checking the consistency and accuracy of the administrative data is applied. In this way a number of anomalies between withholding taxes, taxes paid, social security contributions and corresponding incomes are eliminated. Finally the SM2-EuSilc outputs are compared with the available administrative gross figures at the micro level in order to assess the quality of microsimulation estimates and for reciprocal validation.

The final database of individual and household incomes gross of tax and social security contributions is therefore constructed as the sum of net incomes, taxes paid and withholding taxes from administrative sources, if available, or as the sum of net incomes and microsimulated taxes[17]. It includes, additionally, social security contributions paid by workers and employers estimated by SM2-EuSilc. In fact, the available registers on compulsory social insurance contributions only cover data on employees of private sector (not employers) and on employees and employers of public sector and there is merely a partial coverage of the social insurance contributions of self-employed drawn on the UPF tax returns.

It should be noted that the implemented methodology is essentially based on a strategy of combining fiscal data and microsimulation estimates. In more details it can be said that when the net administrative incomes are higher than the survey incomes, the EU-SILC net and gross incomes completely arise from administrative data, while the social insurance

[14] Tax credits for expenditures as per Section I and III of Part RP of UPF 2012 (medical expenses, vehicle expenses and guide dogs for disabled people, mortgage interest, life assurance and accident insurance, tuition fees, funeral expenses, care expenses, expenses for children's sports activities, estate agents' fees, rent costs for students living away from home, other costs and expenses for building renovation work which are deductible at the rate of 41 per cent or 36 per cent) and other tax credits as per Section IV, Section V, Section VI and Section VII of Part RP.

[15] Deductions for principal dwelling and deduction for expenditures as per Section II of Part RP of UPF 2012 (social security and welfare contributions, regular maintenance allowances paid to spouse, social security contributions for home helps and cares, donations to religious institutions, medical and care expenses for disabled persons, supplementary health insurance and other expenditures).

[16] For taxpayers who filed both 730 and UPF returns, the UPF form was used as it generally contains additional, subsequent information compared with the 730 form.

[17] A stochastic component has been added to the withholding taxes and taxes paid from administrative sources to render the information used anonymous.

contributions are estimated by the model. Consequently the final EU-SILC gross variables do not differ from the fiscal ones. On the opposite, when the survey incomes are higher than the administrative data, the net incomes are those taken from the survey (collected or imputed), while the taxes derived from administrative sources, since these are taxes actually paid by the taxpayers. As always the social insurance contributions are estimated by SM2-EU-SILC. In such a case the final EU-SILC gross variables are essentially different from the fiscal gross data. It is worth mentioning that when the surveyed incomes are higher than the register data, the difference between the surveyed data and the tax data could not be considered as a direct measure of illegal tax evasion. As a matter of fact it is not possible for us to distinguish between the legal tax avoidance, allowed by the national fiscal system, and the real tax evasion[18].

As a direct result of the applied methodology the typical 'adjustment factors' used in any microsimulation model for correcting the disposable income and the gross income values in order to take into account the tax evasion are not applied. In effect EU-SILC disposable income partly includes income not reported to tax authorities, while the taxes for the most part are those derived from the income tax returns and do not require any adjustments.

Table 3.2 - IT-SILC target variables net-gross ratio and gross and net distribution by income components - Year 2011 (percentage values)

VARIABLES 2012 (INCOME REFERENCE YEAR 2011)	Ratio Net/Gross	Distribution	
		Gross	Net
Income from work	**71.4**	**66.5**	**63.1**
PY010 Employee cash or near cash income	72.1	47.6	45.6
PY050 Cash benefits or losses from self-employment	69.5	18.9	17.5
Property income	**73.0**	**4.0**	**3.9**
HY090 Interest, dividends, profit from capital investments in unincorporated business	79.4	1.1	1.2
HY040 Income from rental of a property or land	70.4	2.9	2.7
Taxable benefits	**83.4**	**28.2**	**31.2**
PY090 Unemployment benefits	86.0	2.1	2.5
PY100 Old-age benefits	82.5	24.1	26.4
PY110 Survivor' benefits	85.4	0.8	0.9
PY130 Disability benefits	96.4	1.1	1.4
Tax-exempt social transfers	**100.0**	**1.3**	**1.8**
PY140 Education-related allowances	100.0	0.1	0.2
HY050 Family related allowances	100.0	0.6	0.8
HY060 Social assistance	100.0	0.1	0.1
HY070 Housing allowances	100.0	0.0	0.0
HY080 Regular inter-household cash transfer received	100.0	0.4	0.6
Total	**75.2**	**100.0**	**100.0**

As shown in the table 3.2, the net/gross ratio varies by component for the differences in component-specific deductions and tax credits, and also in the social insurance contributions. The net-to-gross ratio is lower for income from work (71.4%) than for the other components, due to the social insurance contributions to which such income is

[18] In literature there are several works on tax evasion based on such differences. This subject actually goes beyond the present chapter focused on the Eu-Silc production process utilized as input file of the ISTAT microsimulation model.

subject. The ratio of net to gross taxable income of other incomes varies approximately from the low of 70.4% for property income, to 83.4% for various taxable benefits, to of course 100% for social assistance, housing and other tax-exempt benefits. The distribution of gross income shows clearly that the main income component is represented by income from work (66.5%), followed by old age benefits (24.1%). The differences in gross and net distribution proves that the tax burden is higher in income from work than the other components, like taxable benefits and property income.

In the following tables the comparison between EU-SILC and some appropriate external sources are also presented. Data from National Accounts, Labour Force Survey by Istat and data from Fiscal Agencies of the Ministry of the Economy and Finance and the Pensions Register by INPS (National Institute for Social Security) are used as external benchmarks.

The comparison of EU-SILC data with National Accounts figures are shown in table 3.3. The table reports the breakdown of total gross income into social insurance, tax and net components. EU-SILC estimates embrace all income components of target variables even those not included at present in the total gross household income (HY010) (i.e. imputed rent, all fringe benefits, own consumption, employers' social insurance contributions, interest paid on mortgage). On the average, net income, after tax and social insurance contributions including employers' contributions, accounts for 68.3% of total gross. The agreement of EU-SILC results and National Accounts figures is good and let the EU-SILC results satisfactory.

Nonetheless some aspects have to be considered when comparing EU-SILC with National Accounts. NA estimates generally use all the administrative data sources which are integrated in EU-SILC and as it is well known NA estimates are adjusted to account for the grey economy. However the grey economy is partially covered in EU-SILC given that some interviewees report income that are not enclosed in tax registers, including both tax avoidance/evasion and tax exempt. On the one hand EU-SILC integration methodology applies the rule of the maximum between survey and administrative income level, consequently the mean income of EU-SILC is usually higher than the administrative one (which is employed in NA estimates). Moreover EU-SILC survey, as well as other income surveys, typically under-estimates financial capital incomes, which are subject to tax withholding at source at some flat rate, whereas EU-SILC estimation method of imputed rent produces higher value than NA aggregate. Finally it is expected that the combined effect of the above mentioned features explains the closeness between the two data sources estimates.

Table 3.4 shows the comparison of EU-SILC income recipients and some external sources. The EU-SILC number of employees who have received wage or salary positively approximates the number of income earners from National Fiscal Agency data, which represents the universe of taxable employee income recipients. Differences in applied definitions (i.e. domestic vs resident employment), reference period and coverage of the two data sources can well explain the discrepancies in estimates. Furthermore the tax register does not report information on incomes and employees of the hidden economy that, as stated before, are partially included in the survey.

For lack of harmonization and divergence in definition of self-employment income, National Accounts are not directly comparable with EU-SILC estimates and other sources are employed. It should be noted that important differences in definition of self-employed make the agreement reasonable but not excellent. In fact in LFS a worker is classified as an

independent on the basis of his/her main activity and in NA the estimate of self-employed units is in term of full time equalised workers. On the other hand the EU-SILC estimate is referred to the number of people whose earnings from self-employment may have been temporary and/or from a secondary working activity. Data on beneficiaries for three kind of functions - old-age, survival and disability (according to ESSPROS classification) – are also reported and the comparison with external sources shows that EU-SILC estimates are quite close to the administrative data.

Table 3.3 - Distribution of total gross income – Year 2011 (euro per capita and percentage values)

	IT-SILC 2012 (income reference year 2011)		Istat N.A. 2011	Difference (% point)
Gross including SI	21697	100	100	
SI contributions	3917	18.1	17.3	0.8
- Employers' contribution	2718	12.5	12.2	0.3
- Employees' contribution	711	3.3	2.9	0.4
- Self -employment contribution	488	2.2	2.2	0.1
Gross taxable	17779	81.9	82.7	-0.8
Personal income tax and financial tax	2964	13.7	13.3	0.4
Net income	14814	68.3	69.4	-1.1

Sources: Istat (2012) and Istat (2013)

Table 3.4 - Comparison of It-Silc income recipients and some external sources - Year 2011 (Thousands of units)

	It-Silc	Fiscal Agencies	National Accounts (Ula)(a)	Labour Force Survey (LFS)	Pension Register of INPS (b)	Differen-ce (%)	Differen-ce (%)
Persons who have received wage and salary (cash or near cash)	21,459	20,951				2.4	
Persons who have received cash benefit or losses from self-employment	7,812		6,712	5,727		16.4	36.4
Beneficiaries of Old-age-Survival-Disability pensions	16,268				15,998	1.7	

(a) Full time equivalent unit of workers.

(b) Severance recipients and persons aged under 15 and/or residing abroad are not included.

Sources: Istat (2012) and MEF (2013)

4. Concluding remarks

This paper provides a thorough review of the methods used for the estimation of net and gross income variables of EU-SILC survey and for enhancing the quality of income statistics that are produced. The methodology is essentially based on the integration of survey data and administrative datasets at a micro level. The Italian mix-mode collection of data on income (EU-SILC) represents the first example in the field of integration of survey and administrative data among European NSIs. Producing statistics from administrative sources means that data collection, editing and other kinds of data processing are done by

methods different from the traditional ones. Instead of making quality controls of data received from the individuals, in this new perspective it is necessary to modify administrative data on the basis of our knowledge of differences in definitions and coverage between the administrative sources and the statistical needs. The administrative data are essentially exploited in order to fill in the survey missing values, correct outliers or unreliable values and produce the income variables, improving the quality of the final estimates. The joint use of survey data and administrative archives also allowed to use a more innovative methodology than the traditional microsimulation model for producing incomes before taxes and social insurance contributions. The double use of the tax sources on one side and the microsimulation estimates on the other enhances the advantages obtainable with the exclusive use of one of the two instruments. Administrative data especially in terms of income deductions and tax credits are used in the input file of the model instead of estimation by regression technique based on external sources. Moreover the linkage with administrative data has the advantage of reciprocal comparison and validation of the final estimates.

The achieved results of our process of integration allow to conclude that the best way for collecting household's income in Italy is to combine administrative data and survey income data. In effect by means of integration is definitely possible to improve the coverage and the quality of income data. Several projects for further improving data quality are conducting at ISTAT aiming at extending the administrative sources used and the timeliness of data. In a short time it will be a more massive use of administrative data in order to replace information from survey questions (i.e. pensions), while there is a transition towards a multi-mode data collection based on computer assisted techniques such as CATI (Computer Assisted Telephone Interviewing) or even CAWI (Computer Assisted Web Interviewing). The aim is finally to have shorter questionnaires and decrease data collection costs, reduce response burden, and achieve a better measurement of income data.

References

Betti G., G. Donatiello and V. Verma. 2011. The Siena Microsimulation Model (SM2) For net-gross conversion of Eu-Silc income variables. *The International Journal of Microsimulation.* 4(1): 35-53. http://www.microsimulation.org/IJM/IJM_V4_1.htm.

Burricand C. 2013. "Transition from survey data to registers in the French SILC survey", Eurostat. *The use of registers in the context of EU–SILC: challenges and opportunities*. Statistical Working Paper. Luxembourg. Publications Office of the European Union 2013. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-TC-13-004/EN/KS-TC-13-004-EN.PDF

Consolini, P., M. Di Marco, R. Ricci, R. and S. Vitaletti. 2006. "Administrative and Survey Microdata on Self-Employment: the Italian Experience with the Eu-Silc project". Iariw 29th General Conference. Joensuu. Finland, 20-26 August.

Consolini P. and G. Donatiello. 2013. "Improvements of data quality through the combined use of survey and administrative sources and micro simulation model". Eurostat. *The use of registers in the context of EU–SILC: challenges and opportunities*. Statistical Working Paper. Luxembourg. Publications Office of the European Union 2013. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-TC-13-004/EN/KS-TC-13-004-EN.PDF

Consolini, P. 2007. "Experiences on the harmonization of the definitions, the variables and the units for the Eu-Silc project in Italy". Working package WP2: Recommendations on the use of methodologies for the integration of surveys and administrative data. Final report of "ESSnet Statistical Methodology Project on the Area: Integration of survey and administrative data". pp.58-67. http://cenex-isad.istat.it/dokeos/document/document.php

Consolini P. 2009. *Integrazione dei dati campionari Eu-Silc con dati di fonte amministrativa*. Collana Istat Metodi e Norme vol. 39/2009. Roma. http://www3.istat.it/dati/catalogo/20090318_00/

Di Marco M. 2006. "International Comparability of Microdata on Incomes: Lessons From the Eu-Silc Project". VIII International Meeting on Quantitative Methods for Applied Sciences. Certosa di Pontignano (Siena). 11-13 September.

Donatiello G., G. Betti G., P. Consolini. 2012. *The Construction of Gross Income Variables of Eusilc (Eu Statistics on Income and Living Conditions) in Italy: A Mixed Strategy Using Microsimulation and Administrative Data*. Università degli Studi di Siena. *Quaderni del Dipartimento di Economia Politica e Statistica*. n. 652 – Settembre. http://www.econ-pol.unisi.it/quaderni/652.pdf

Donatiello G. 2011. *La metodologia di stima dei redditi lordi nell'indagine Eu-Silc - Indagine europea sui redditi e le condizioni di vita delle famiglie*. Collana Istat Metodi e Norme vol. 49/2011. Roma.

http://www3.istat.it/dati/catalogo/20110726_00/metodologia_stima_redditi_lordi_indagine_eu_Silc.pdf

Eurostat. 2004. *Income in EU-SILC: Net/Gross/Net conversion. Report on common structure of the model; model description and application to the ECHP data for France, Italy and Spain*, prepared by V. Verma, G. Betti and co-researcher. EU-SILC 133/04, Luxembourg, 2004.

Heuberger R., T. Glaser T., and E. Kafka. 2013. "The use of register data in Austrian SILC survey". *The use of registers in the context of EU–SILC: challenges and opportunities*. Statistical Working Paper. Luxembourg. Publications Office of the European Union 2013. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-TC-13-004/EN/KS-TC-13-004-EN.PDF.

Herzog, T.N., F.J. Scheuren, , W.E. Winkler. 2007. *Data quality and Record Linkage Techniques*. New York: Springer ed.

Kapteyn A., J. Y. Ypma. 2007. Measurement Error and Misclassification: A Comparison of Survey and Administrative Data. *Journal of Labor Economics*, 25 (3): 513-551. http://www.jstor.org/stable/10.1086/513298

Istat. 2012. "*Condizioni di vita (UDB It-Silc)"*. Roma: Istat. (microdata) http://www.istat.it/it/archivio/4152.

Istat. 2013. "*National Accounts Years 2011-2012"*. Roma. Istat.

MEF, Ministero dell'Economia e delle Finanze (2011). "Dichiarazioni fiscali". http://www.finanze.gov.it/stat_dbNew2011/index.php

Méndez Martín J.M. 2013. "Reconciliation of income data from survey and from administrative sources". *The use of registers in the context of EU–SILC: challenges and opportunities*. Statistical Working Paper. Luxembourg. Publications Office of the European Union 2013. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-TC-13-004/EN/KS-TC-13-004-EN.PDF.

Newcombe H.B. 1988. *Handbook of Record Linkage: Methods for Health and Statistical Studies. Administration, and Business*. Oxford (UK): Oxford University Press.

Nordberg L. 2003. "An Analysis of the Effects of Using Interview versus Register Data" in Income Distribution Analysis Based on the Finnish ECHP-surveys in 1996 and 2000, Chintex Working Paper #15, Work Package 5, December 22 2003.

UNECE. 2011. *Canberra Group Handbook on Household Income Statistics.* Second Edition. United Nations. New York and Geneva.

van der Laan, P. 2000. *Integrating administrative registers and household surveys*. Netherlands Official Statistics. 15. Summer 2000.

Wallgren, A. and B. Wallgren. 2007. *Register-based Statistics: Administrative Data for Statistical Purposes*. New York: John Wiley and Sons.