

THE CONSTRUCTION OF GROSS INCOME VARIABLES OF EU-SILC (EU STATISTICS ON INCOME AND LIVING CONDITIONS) IN ITALY: A MIXED STRATEGY USING MICROSIMULATION AND ADMINISTRATIVE DATA

Gabriella Donatiello

ISTAT, Via Ravà, 150, 00142 Roma, Italy; email: donatiel@istat.it

Gianni Betti

University of Siena, P.zza S. Francesco, 7, 53100 Siena, Italy; email: betti2@unisi.it

Paolo Consolini

ISTAT, Via Ravà, 150, 00142 Roma, Italy; email: consolin@istat.it

ABSTRACT: According to the EU Regulation on European Statistics on Income and Living Conditions (EU-SILC), Italy will provide household gross income statistics starting from survey year 2007.

Since in Italy both survey and fiscal data are used for the construction of the EU-SILC target variables, for the net-gross conversion of income variables, Istat has experimented a new methodology using in conjunction a microsimulation model (Siena Micro-Simulation Model SM2-EU-SILC) and an exact record linkage between survey and fiscal data at micro level.

The integration of microsimulation with register data has the advantage of using administrative data for the validation of microsimulation results. Since tax data have an incomplete coverage in respects of all surveyed individuals or in respects of some kind of social insurance contributions (i.e. employers' contribution), SM2-EU-SILC could estimate those taxes and social insurance contributions not covered by register data. Finally, the use of microsimulation and administrative data improves the quality and the amount of information on gross income.

This paper summarises the data production process and its main results, focussing on the joint use of SM2-EU-SILC and on the records linkage between survey and administrative data as well.

1. INTRODUCTION

Italy has provided gross income statistics, according to the EU Regulation on European Statistics on Income and Living Conditions (EU-SILC), starting from survey year 2007. The production of net and gross income microdata derived from the same sample survey is an important improvement for Italy.

For the net-gross conversion of income variables, ISTAT has experimented a new methodology using in conjunction a microsimulation model (Siena Micro-Simulation Model SM2-EU-SILC) and an exact record linkage between survey and fiscal data at micro level.

The 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.

For the construction of EU-SILC gross income variables, the Siena Micro-Simulation Model (SM2) has been adopted as recommended procedure by the European Commission. SM2 has been

developed for calendar year 2003 and applied to the ECHP (*European Community Household Panel*) survey data.

The reasons in developing SM2 are somewhat different from existing micro-simulation models. At the outset, SM2 is designed for *multi-country application*, as a *flexible tool* which is *portable* to the maximum extent possible across (at least the European) countries despite great differences in fiscal systems. The immediate context for the development of SM2 has been certain specific requirements of EU-SILC (EU Statistics on Income and Living Conditions). EU-SILC is a statistical source, developed by European Commission (Eurostat) and implemented by all EU and also many other European countries, for the generation of comparable and detailed information on living conditions and income of households and persons. The central issue to be addressed is that, while the source, type and form of input (collected) information varies across and even within countries, the output required at the European level has to be comparable and standardised (Eurostat, 2002). Furthermore, while the information which can be collected is limited to particular forms because of limitations of the sources providing it, it is required in both net and gross forms for diverse academic and policy research. We see SM2 as a *tool*, under continuing development, for meeting these objectives in the international, comparative context. Starting from data on household and personal income given in different forms (including some missing data), and on the basis of the prevailing fiscal system in a country, the model estimates full information on income by component, with breakdown of gross amounts into taxes, social insurance contributions, social transfers, and net and disposable income. Therefore it can be applied to diverse data sets to generate variables (such as the EU-SILC Target Variables) in a standard form. Furthermore, it is designed to be *flexible* to deal with an annual flux of data in different forms across and within countries and also with periodic changes in the national tax systems, which a longitudinal data source such as EU-SILC must deal with.

Thus an outstanding and unique feature of the SM2 system is that its core consists of a *standardised set of routines* which can handle a great diversity of input data forms and national tax systems. *Country-specific routines* are required to convert the input data into standardised forms, and also to specify parameters of the national tax system in an appropriately standardised form. These, then, form inputs to the central core of the system designed to generate the required standardised outputs. The system has been developed to maintain a clear distinction between the common and the country-specific parts, and even more importantly, to maximise the part which can be standardised. This feature makes the system an appropriate and convenient tool for multi-country application.

Given the specific context and objectives of its development, hitherto SM2 is fully 'data based' and does not incorporate simulation of benefits or any other income components. It is taken as given that information on all income components has been collected, compiled or imputed in some form, and that the objective is to convert it, under a specified national tax system applicable at the time, to the standard form (specifically that required by EU-SILC). It incorporates generally the same or similar level of detail as other major micro-simulation models - a little less detailed on some points but also more complete on some others. The main difference is that presently SM2 apart from so far being data-based rather than simulation-based concerning benefits and similar transfers. For this reason SM2 cannot be considered as an alternative to existing micro-simulation models. Rather it is a good complementary model capable of constructing the very detailed and standardised input data sets which can be used by European tax-benefit models such as Euromod.

For the net-gross conversion of EU-SILC variables, ISTAT decided to test the application of the SM2 using the new survey data and to experiment some methodological improvements based on the ISTAT experience in using both administrative data and sample survey data.

The Italian EU-SILC survey is based on the "face to face interview" method of collecting data and uses administrative micro-data in order to reduce measurement errors. Many topics, including statistics, psychology, sociology and economics, share a common concern with the weakness of the measurement process in the survey method. Errors can arise from any of the factors engaged in the

measurement process, like the questionnaire, the respondent and the interviewer, as well as the methods of collecting data. The questionnaire includes several sources that affect the interpretation of the question by the respondent, like the question wording and the structure of questionnaire. Even if the respondent fully understands the question, there can be memory problems. Memory problems are the cause of errors: omissions and telescoping errors are just two typical examples. In order to limit the impact of measuring errors on the income reported in the questionnaire by the interviewees, and generally to improve the data quality in the survey, a project of multi-source data collection has been started up in the Italian National Institute of Statistics (Istat). This project employs administrative data to support the editing and imputation processes in the survey on one hand, and on the other hand to provide a basis to build up the gross income in conjunction with microsimulation. The integration technique, used to combine survey and administrative data at micro level in Eu-Silc, can be viewed as a flow process, starting from the analysis of the phenomenon and data-sources, developing through the choice of the best matching-key and the more effective record linkage method, the problem solving approach for harmonizing units and variables, treatments to handle incoherence and under/over coverage of data-sources, and finishing with the reconciliation of values reported in the different sources. In the first EU-SILC edition (survey 2004), this process has involved only two income components which are self-employment income and pensions. Nevertheless for the second edition (survey 2005) it has included a third one: the employment incomes.

As well known, some problems could be met by the using of tax data for income data production. The definition of taxable income or tax units could be different in tax data and in income statistics, with problems of reliability and comparability. In addition, the tax data may have an incomplete coverage in respects of all surveyed individuals or in respects of some kind of income or social insurance contributions (i.e. employers' contributions), therefore a microsimulation model could estimate those taxes and social insurance contributions not covered by register data.

The ISTAT gross income data production process can be summed up in three important steps: the first one is the development of the model SM2-EU-SILC; the second one is the integration of survey data and administrative data used in conjunction with microsimulation and the third one is the construction of the final data set of individual and household gross income target variables.

The development of the model SM2-EU-SILC required a preliminary transition from ECHP to EU-SILC data. The introduction of the model to the new survey called for new procedures for the construction of the input file and implied the adjustment of some conversion routines of SM2. In particular for the estimation of self employed income and the CoCoCo (temporary subcontractors) income and for the calculation of IRAP tax (regional tax on productive activities). Additional adjustments in SM2 procedures were needed also for including the tax reform of year 2005.

The availability of administrative data for the Italian EU-SILC has consented to use both microsimulation and administrative archives in an innovative way.

The administrative data in terms of net incomes, tax credits and income deductions are utilized with survey data as input file of SM2-EU-SILC and as benchmark for microsimulation results.

Hence fiscal data and microsimulation estimates are both applied for reciprocal comparison and validation and for the construction of the final data set of gross incomes at individual and household level.

The final data set of individual and household gross income variables is definitely the result of the mixed strategy of using in conjunction microsimulation and administrative data.

For what concerns the gross income data production process, SM2-EU-SILC integrated with survey and register data has estimated taxes and social insurance contributions for those individuals not present in register data due to errors in the record linkage procedure (errors in the identification numbers of individuals). Furthermore the model has simulated the employers' social insurance contributions of the private sector and also the self employed' social insurance contributions not fully covered by the available administrative registers.

The final EU-SILC individual and household gross income variables are computed as net amounts plus taxes and social insurance contributions derived from register data, if available, or estimated by SM2-EU-SILC. In order to anonymize the administrative data used, a stochastic component has been added to the withholding taxes and to the taxes paid derived from registers.

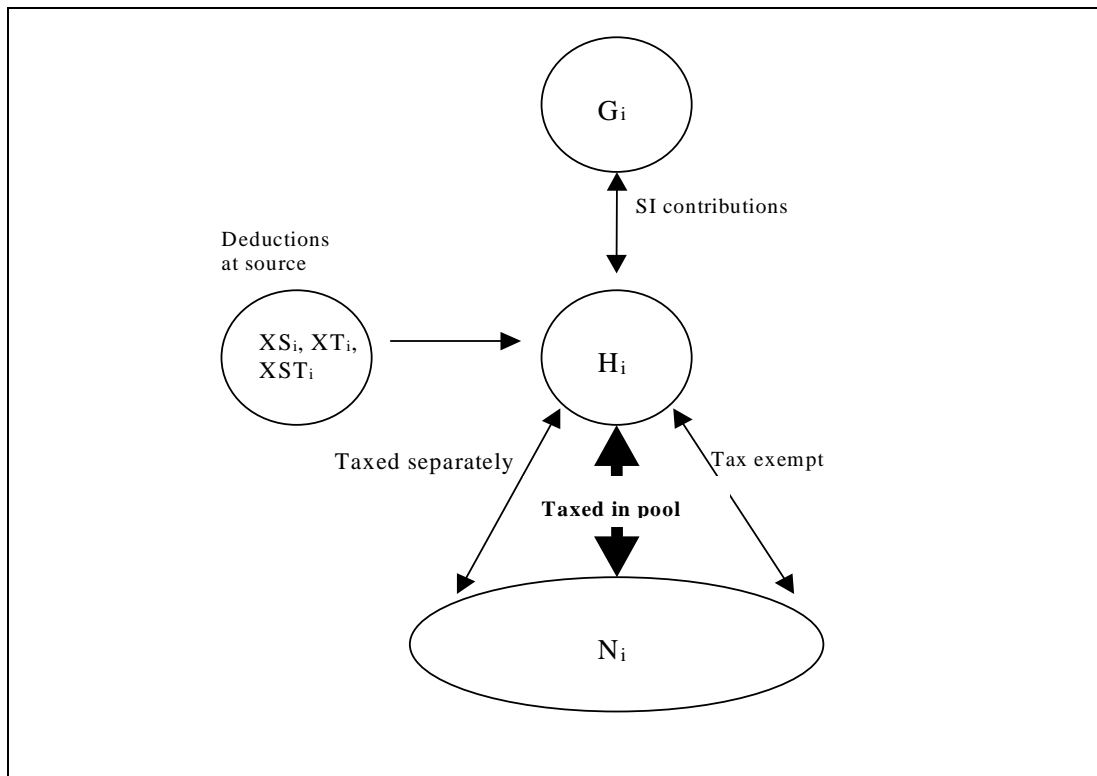
The paper reviews the EU-SILC data production process in Italy for survey year 2007, focussing on the joint use of SM2-EU-SILC and the integration of survey data and administrative data for the construction of gross income target variables.

Section 2 describes the main features of SM2, while **Section 3** explains the setting up of the integrated data set for the construction of EU-SILC net income variables. **Section 4** focuses on the ISTAT methodology of using in conjunction a microsimulation model and an exact record linkage between survey and fiscal data at micro level for the construction of EU-SILC gross income variables. Finally, in **Section 5** some outputs are reported and compared with National Accounts figures.

2. THE SIENA MICROSIMULATION MODEL (SM2)

Figure 1 shows the basic relationship between gross and net forms of income when more than one income components and possibly more than one individual in the tax unit are involved. The relationships between gross taxable income for a particular component, H_i , and quantities like gross income G_i and income after retentions at source XST_i are usually simple, dependent only on i , the income component concerned and determined independently of other components and other persons in the tax unit. The same applies to the relationship between H_i and net N_i for components which are taxed separately at a flat rate or a rate determined by the level of income from that component alone, and of course also for tax exempt components. Sometimes, dependence of the relationship on other sources of income may also be involved, but mostly these are simply in the form of upper limits which may apply to certain quantities pooled over more than one component.

Figure 1 Basic relationship between net and gross amounts



Generally, however, all or most taxable income is pooled together over components and over persons in the tax units for the purpose of determining the amount of tax due. The relationship between H_i and N_i for components in the pool is more complex than above. Going from known H_i to N_i is less problematic since the relationships (the tax rules) are a function of the former. These relationships are specified in more detail in Table 1.

Table 1 Gross-to-Net conversion algorithm

	Income measure	Total	by component⁽¹⁾
1	GROSS(2)	$G = \sum G_i \leftarrow$	G_i
2	Social Insurance contribution		$S_i = S_i(G_i)$
3	GROSS TAXABLE	$H = \sum H_i \leftarrow$	$H_i = G_i - S_i$
4	Component-specific deductions		$D_i = D_i(H_i)$
Aggregation over components and individuals in tax unit			
5	TAXABLE INCOME	$Y = \sum Y_i \leftarrow$	$Y_i = H_i - D_i$
6	Common deductions	$D_0 = D_0(H)$	
7	Taxable income(0)	$Y_0 = Y - D_0$	
8	Tax due(0)	$W_0 = W_0(Y_0)$	
9	Common tax credits	$C_0 = C_0(Y_0)$	
10	TAX DUE	$W = W_0 - C_0$	
11	Component-specific tax credits	$C = \sum C_i \leftarrow$	$C_i = C_i(Y_i)$
12	TAX PAID	$X = W - C$	
13	TOTAL NET	$N = H - X$	
14	Tax rate(0)	$R_0 = X/H$	
15	TAX RATE = TAX DUE/ TAXABLE INCOME	$R = W/Y$	
Disaggregation – personal income by component			
16	Proportionate tax by component		$X_i = R * Y_i - C_i$
17	NET BY COMPONENT		$N_i = H_i - X_i$

⁽¹⁾ The functional relationships in this column may be somewhat more complex or varied.

⁽²⁾ Gross including employers' social insurance contribution (SS) is: $GG = G + SS(G_1)$

The form in which data on income by component are available may vary from one country (tax regime) to another, and also among individuals and households within the same country. There are two dimensions of this variation:

A. Whether or not a particular component is subject to social insurance contributions and to income tax. Income tax may apply in various forms. (i) Some components may be pooled together, across components and also across individuals in some appropriately defined tax unit. (ii) Some may be subject to tax separately, each at a certain flat rate. (iii) Some components in the 'pool' may be tax-exempt up to a certain flat rate but taxed beyond that if a higher rate applies. (iv) Some may be subject to double taxation, perhaps representing some combination of the other forms.ⁱ (v) And of course, many types of incomes, in particular social transfers, may be tax

exempt. Mostly, the form applicable to each type of income is determined by the national tax regime, normally uniform for all respondents in a country. Hence this information can be compiled at the aggregate level and need not be collected at the micro level. There can be exceptions, however, for persons in special circumstances. There can also be other complications, such as more than one components, otherwise treated separately, being subject to common ceilings. In some systems, individuals have a choice among the various options.

B. The form in which the information has been collected. This may generally vary from one individual to another in the same survey, though a uniform reporting form may prevail for some components. In any case, the information on the form in which the data are available is required at the micro-level. The amount may for instance be reported as gross, or net of social insurance contributions and/or tax; and in the case of tax retentions, whether they are 'retentions at source' according to some rules or individual arrangements, or are the 'final retentions' of the tax actually due, in the sense explained below. Table 2 lists the various reporting forms.

Table 2 Forms of reporting of an income component

Income component (i) subject to tax and social insurance contributions	
Form (X_i) in which data on the income component have been collected:	
G_i	gross income (before tax and SI contributions, if applicable)
H_i	gross taxable (before tax, but after SI contributions, if any)
N_i	net income (after deducing 'final' tax and SI contributions, i.e., as the final amount actually received)
Income received after retentions at source:	
XT_i	taxed at source (but no SI contribution); tax at source T_i
XS_i	SI contributions (but not tax) at source; SI contributions at source S_i
XTS_i	both tax and SI contributions at source, tax and SI at source T_i+S_i

Here we describe the standardised 'core' of the SM2 system, taking account of complexities B, but assuming for the moment that through the information may be reported in diverse forms, all income components over individuals in the tax unit are pooled together and subject to a common tax schedule.

Conversion routines. Table 3 shows the procedure for converting the reported amount with any combinations of the above dimensions of variation into a standard form. As noted at the bottom of the table, the income components may be divided into two sets, say 'N' and 'H', depending on whether the amount reported is 'final net' (N_i), or is in some other form ($G_i, XS_i, XT_i, XTS_i, H_i$) more directly convertible to the 'gross taxable' form H_i . For all forms other than 'final net' N_i , it is convenient to take 'gross taxable income' H_i as the standard target of the conversion:

$$[G_i, H_i, XS_i, XTS_i, XT_i] \Rightarrow H_i.$$

This conversion involves the component and country-specific functional relationships or schedules, namely $S_i = S_i(G_i)$, for social insurance contributions, and $T_i = T_i(H_i)$, for tax retention at source.

As noted, tax retentions at source may be according to fixed schedules, or according to arrangements determined at the individual (micro) level.

In a majority of the cases, H_i can be determined directly from the collected amount, for instance from gross amount (G_i) reported for an income component i subject to social insurance contributions, we have: $H_i = G_i - S_i(G_i)$.

In other cases, an iterative procedure may be required. However, generally the iteration is very simple and converges quickly. This is because by and large component-specific schedules apply to each component separately. There are no other parameters to be estimated. The need for numerical

iteration arises simply from the fact that the unknown quantity to be determined (H_i) appears in an implicit equation.

Table 3 Calculation of H_i according to the form in which the component is specified

Set H

given value	XS_i	$H_i = XS_i$	
	G_i	$H_i = G_i - S_i(G_i)$	
$P_i =$	XT_i	$H_i = G_i - S_i(G_i)$ where $G_i = XT_i + T_i(H_i)$	Simple iteration, generally separately for each component
	XTS_i	$H_i = XTS_i + T_i(H_i)$	

Set N

given value $P_i =$	N_i	$H_i = Y_i + D_i(H_i)$ where $Y_i = [H_i - N_i + C_i(Y_i)] / R$	Double iteration (i) with assumed R, for each component in turn (ii) for determining R, common to all pooled components
------------------------	-------	---	---

Set of variables N: set of income components which are subject to income tax (irrespective of whether the component is also subject to social insurance contributions), and for which the 'final net' amount ($X_i=N_i$) has been specified in the data collected.

Set of variables H: all other income component (not subject to tax, or for which the data has been collected in a form other than the 'final net' amount).

The second panel of Table 3 shows the relationship between H_i and the reported amount in the form 'final net' N_i . Going from N_i to H_i in fact involves a double iterative loop. The inner loop of iteration is applied with an assumed value of the parameter "tax rate" (R , as defined in Table 1). Once this has been done for every income component in the group (including over all individuals in the same tax unit), an outer iterative loop obtains a convergent value of this parameter which is common to all those components. The N_i to H_i conversion process is therefore considerably more complex. Furthermore, this complexity is substantially increased in the presence of missing data, where the modelling and imputation procedures will have to be applied interactively.

Iterative procedure. Table 4 demonstrates the common structure of the iterative procedure. The procedure distinguishes between sets H and N as defined in Table 3, and may be applied as follows. The required H_i quantities for set H are computed (only once) using Table 3, and form an input into the iterative cycle for parameter R required for set N. The parameter is best estimated by using information on all income components from both the sets.

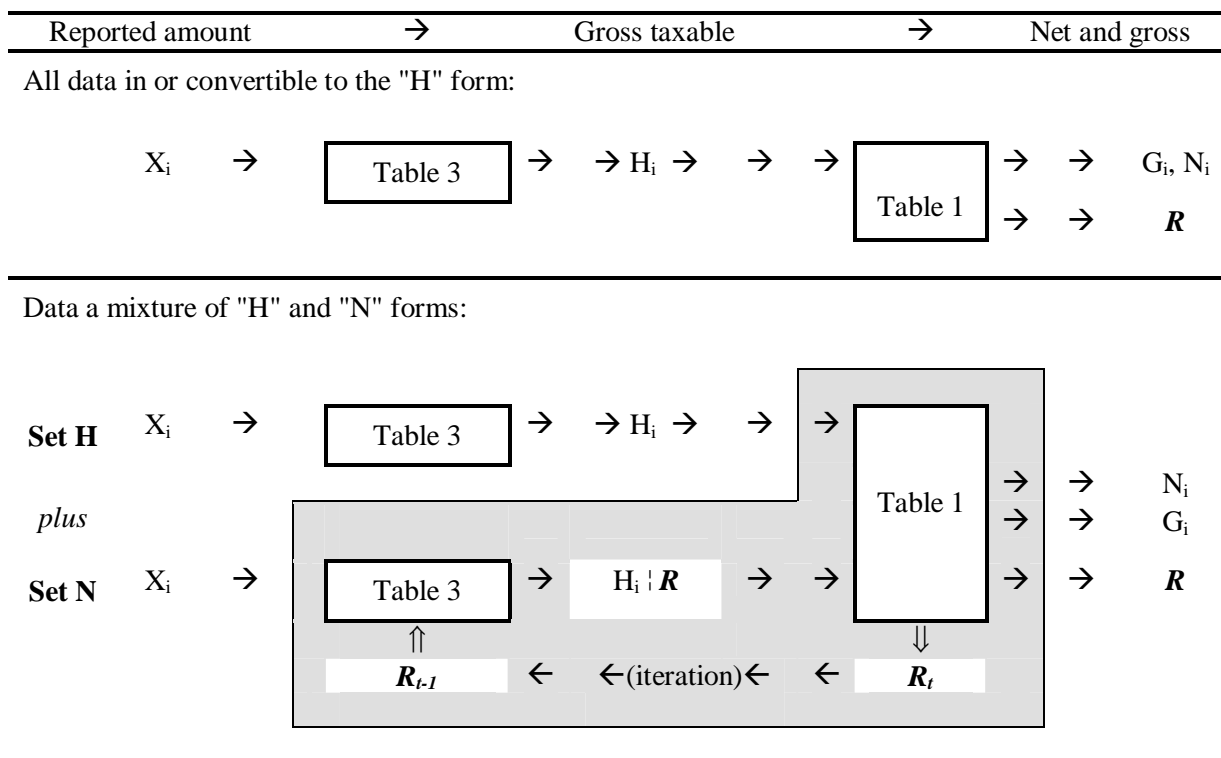
The net-to-gross iterative procedure could be affected by two common problems in microsimulation modelling: non-convergence and multiple-convergence.

By non-convergence we mean that starting from a net value, the procedure is not able to find any gross value. This may be because no gross value exists as a result of some peculiarities of the data, tax-benefit rules, or hypotheses made concerning deductions or tax credits which cannot be calculated from available data or rules. Alternatively, this may happen when in principle a solution exists but the SAS routine does not converge to the solution in an "acceptable" number of iterations. In SM2 SAS routines, this problem is dealt with as follows. The system finds a gross value the net corresponding to which is the closest to the given net value. Then the ratio (given / computed) net for each component is used to adjust its computed gross proportionately. The adjusted gross value can be taken to correspond to the given net amount. The adjustment required is usually very small.

For identifying the problem of multiple convergence, the SAS routine introduces small random perturbations in the computed "tax rate" R in order to identify whether it is a "local convergence", i.e. whether there exist multiple values of gross which correspond to exactly the same given net

value. If the problem of multiple convergence is identified, some judgemental (“reasonable”) criteria have to be used to select a particular solution.

Table 4 Common structure of the iterative model



3. THE RECORD LINKAGE OF ADMINISTRATIVE AND SURVEY DATA FOR THE ITALIAN EU-SILC

The EU-SILC (European Union Statistics on Income and Living Conditions) Italian team has developed a pioneer strategy in the measurement of self-employment income since 2004. This strategy consists in a multi-source data collection, based on a paper and pencil face-to-face interview (PAPI) 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 family¹. 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 addition matching tax returns records with survey data also provides information at micro level on social security contributions, taxable incomes and tax liabilities. All these information are used to measure the gross/net taxable income and represent the input of SM2 micro-simulation model.

In the first edition (survey 2004), this process has involved only two income components which are self-employment income and pensions. Nevertheless for the second edition (survey 2005) it has included a third one: the employment incomes. The target population is represented by the Italian reference population of EU-SILC: all private households and their current members residing in Italy

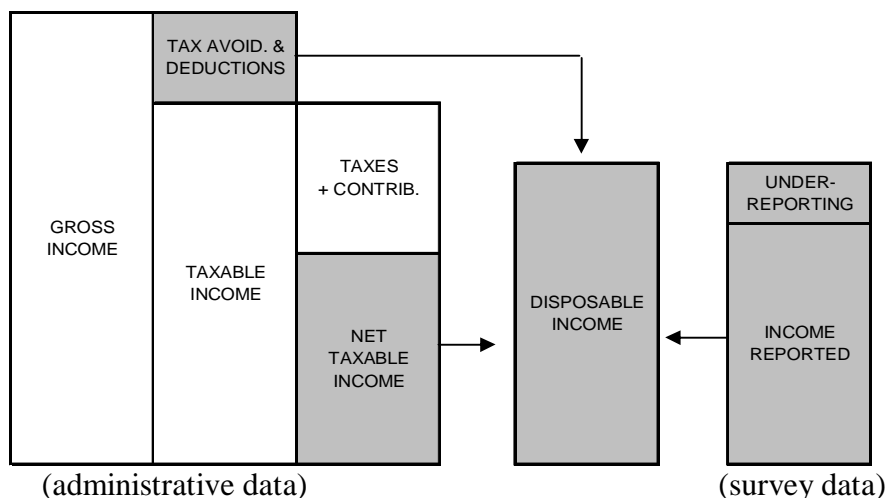
¹ Newcombe (1995).

at time of data collection. Persons living in collective households and in institutions are excluded from target population. The analysis units are adult members (15+ aged) who live in private households. The EU-SILC 2007 survey includes approximately 45.1 thousand interviewees aged 15 years or over. Among these, about 42.5 thousand units have a tax statement or a declaration in the administrative data-sources. As the Personal Tax Annual Register (including all the Italian tax codes) cannot be used directly from Istat, the record linkage has been performed by the Tax Agency on Istat's behalf. The exact linkage performed by Tax Agency on EU-SILC 2007's survey frame (individuals) produces 3.3% unmatched records that are partially retrieved by means of auxiliary information (0.9%). Transfer of know-how in performing record linkage from Istat to Tax Agency could further improve the effectiveness of the matching procedure in the coming years.

3.1 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”². Figure 2 below shows, in a simplified sketch, the problem of collecting self-employment incomes when either survey or administrative data are available, and the objective is to obtain disposable income: the shaded areas correspond to the income available to an individual for his/her personal use.

Figure 2 – Personal gross, taxable, reported and disposable 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

² Canberra Group (2001, p. 54 and p. 62).

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. 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 percipients of small and/or secondary self-employment incomes³.

Regarding the measurement of income from pensions it is assumed that the administrative data provides 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 defining the solutions to the problem of the harmonization of units, definitions and variables and the reconciliation of the incoherencies in income values between the sources involved. Table 5 reports the most relevant meta-information on pension of each administrative data-source. It is noticeable that only bringing together two or more separate pieces of information recorded in different sources it is possible to estimate the EU-SILC target variables. As a result, in order to reckon the net pension income distinctly for each function (target variables), the “yearly net tax income of the pensions” (National Tax Register) and the “monthly gross payments on pensions” broken down into functions and types (Italian Social Security System) have to be combined.

Table 5 – Meta information on pensions/pensioners by administrative sources

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

The Pension Register collects a set of information at 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 functions or the target variables. In order to join the information of the Tax Registers with the Pension Register we need to define a “harmonized

³ For a more detailed analysis of this subject it is advised to see Consolini *et al.* (2006) and Di Marco (2006).

definition of pension income” that is comparable between these data-sources. The common base of the comparison is represented by the “taxable income connected to the pensions”⁴.

The measurement of employment income is based on comparison of administrative and survey data on wages and salaries after retention at source. The main administrative source for this income component is represented by CUD/770 tax statements register. In Italy the employers, as withholding agents, are obliged to declare the amounts of wages/salaries and social benefits annually paid to their employees. As the employed income’s items covered by administrative source are not perfectly comparable with the target variable *PY010 (employee cash or near cash income)* it is necessary to reallocate some of them in a proper way.

The administrative net income is obtained as net taxable employed income less retention at source of tax and social contributions. This aggregate is thus compared with the net employment income retrieved from the EU-SILC questionnaire.

3.2 The integration methodology

In order to carry out the integration of alternative databasis, some basic requirements have to be satisfied by all sources involved. Therefore, the statistical units are to be defined uniformly 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 among the data-sources considered (harmonisation of the variables and classifications), all data should refer to the same period or the same point in time⁵. In other terms, administrative data need to 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 technique allows to combine information related to the same statistical units by means of a collection of identifiers called “match keys” provided that each unit is associated with a unique identifier not affected by errors. Different typologies of exact record linkage exist: in this case we refer to 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 source⁶. Records in different data sources are matched using the Personal Tax Number. Once the integration task is completed, the identification numbers are dropped and replaced with an internal system code according to the Italian rules and regulations that protect people confidentiality. The integration process between survey and administrative data at micro level, can be summarized in the following 4 phases⁷:

Selection of the matching-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);

Linkage of survey and tax records: tax codes of the previous phase were matched to those in the Personal Tax Annual Register, consisting of all the Italian tax codes. The procedure searched for the tax codes of the persons in the EU-SILC sample among the ones in the tax files. More precisely, linkage focuses mainly on adults (15 years and over) that actually participated in the survey. In the last year the rate of successfully matched records was 96.4%. In other words, the tax source covers 96.4% of the adults interviewed for the 2007 Italian EU-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 later registered as additional household’s members by the interviewers (1.4%).

Loading tax data: this step consisted in reading and checking information on the three principal income components (employment income, self-employment income, pensions) included in the tax records. At this stage, four relevant sources of microdata have been uploaded: i) the “Pensions

⁴ See, for more details, Consolini (2008).

⁵ van der Laan (2000).

⁶ See Newcombe (1988), Herzog, Scheuren and Winkler (2007).

⁷ See Consolini (2009).

Register (PR)” of INPS (Italian National Social Security Board), ii) the “CUD/770” tax statements register” of the National Tax Agency, iii) the “730” tax returns register of the National Tax Agency, iii) the “Unico persone fisiche” tax returns register” of the National Tax Agency .

Imputation through integration: 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 has been made at the individual level. However, for small family businesses, comparisons have been made at the household level, that is by comparing the sums of the self-employment income received by all household members in the two sources.

As regards the pensions, it is possible to put in relation each pensioner with the data source that records the payments transferred to him/her. The tools of relative differences, in terms of income values observed on the same statistical units across different data sources (administrative and tax), represent the core of the decisional structure used when defining the pension levels and, generally, when attributing the income components in presence of information from both the Pension Register and Tax sources. When we compare at the individual level the gross taxable pensions of the Pension Register with the gross income pensions of the CUD/770 tax source, we find out that 84.14% of the matched cases shows relative differences in absolute values under the 5% threshold. Such a result demonstrates the high coherence of information on pension levels between administrative and tax sources.

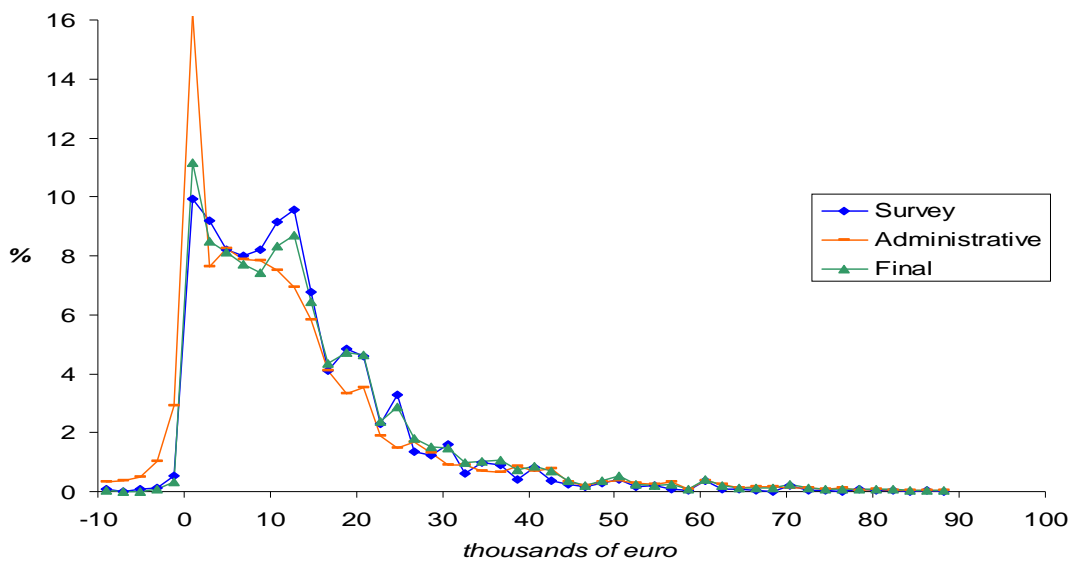
The assumption underlying the building of the gross/net income deriving from pensions is that the true gross income is reported in the Pension register and the proper information on the tax at source, as well as on tax credits, is included in the fiscal sources. Survey data on pensions (after retention 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 employment 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 employment income records reported in all administrative sources.

Finally, the use of administrative data has changed the tails of the distribution of self-employment incomes in Eu-Silc 2004 edition (Figure 3). Indeed, with respect to 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 a higher proportion of percipients with incomes greater than 20,000 Euros.

As a result, combining administrative and survey data brings about a rise of 15.6 % in the number of percipients and an increase of 11.9 % in the average of self-employment income compared to the exclusive use of survey data. 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.

Figure 3 - Distributions of self-employment incomes drawn by: survey, administrative and final dataset (all percipients)



4. MICROSIMULATION AND ADMINISTRATIVE DATA: A MIXED STRATEGY

According to the EU Regulation on European Statistics on Income and Living Conditions (EU-SILC), Italy has provided gross income statistics starting from survey year 2007. It is the first time that both net and gross variables are produced from the same survey in Italy.

For the net-gross conversion of income variables, ISTAT has experimented a new methodology using in conjunction a microsimulation model (Siena Micro-Simulation Model SM2-EU-SILC) and an exact record linkage between survey and fiscal data at micro level.

The 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.

For the construction of EU-SILC gross income variables, the Siena Micro-Simulation Model (SM2) has been adopted as recommended procedure by the European Commission. SM2 has been developed for calendar year 2003 and applied to the ECHP (*European Community Household Panel*) survey data.

For the net-gross conversion of EU-SILC variables, ISTAT decided to test the application of the SM2 using the new survey data and to experiment some methodological improvements based on the ISTAT experience in using both administrative data and sample survey data.

The ISTAT gross income data production process can be summed up in three important steps: the first one is the development of the model SM2-EU-SILC; the second one is the integration of survey data and administrative data used in conjunction with microsimulation and the third one is the construction of the final data set of individual and household gross income target variables.

4.1 The gross income data production process

The development of the model SM2-EU-SILC required a preliminary transition from the ECHP data to EU-SILC ones. In fact, the SM2 input file was based on ECHP Udb and Pdb data and applied for three countries (Italy, France, Spain). 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 latest tax reform in Italy.

The introduction of the model to the new survey called for new procedures for the construction 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 in the input file of SM2 and particular attention was paid in 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 reconstruct the family relationships in the households and allows the dependent persons to be singled out.

Compared to ECHP, the new survey collects detailed information on sector of activity, work status, number of months in a given status, or firm size useful for calculating the social security contributions for dependent workers and also for self employed. Moreover, the breakdown of sickness and invalidity benefits is available in EU-SILC as well as data on pension contributions made to private entities, which give rise to a tax deduction in the Italian system.

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. 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 CoCoCo. (temporary subcontractors) which is nominally included in self employment income, but in fact is treated as employee income. Data on this kind of workers were not available in the ECHP survey, and a variable defining the propensity of an employee to be a CoCoCo 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 tax reform of year 2005.

The availability of administrative data for the Italian EU-SILC has consented to use both microsimulation and administrative archives in an innovative way.

The integration of survey data and register data in the Italian EU-SILC has the most important aim to reduce the under estimation of income variables on the basis of available information (survey and registers).

As well known, tax data may have an incomplete coverage in respects of all surveyed individuals or in respects of some kind of income or social insurance contributions (i.e. employers’ contributions) and a microsimulation model could estimate those taxes and social insurance contributions not covered by register data.

Five archives related to employee income, self employed income, old age benefits and unemployment benefits are used. Through an exact matching of administrative and survey records, the tax data are integrated with survey microdata.

Before using the integrated data set as input file of SM2-EU-SILC, a further procedure of coherence analysis and correction of the net and gross amounts and the related taxes and social insurance contributions was needed.

The administrative data in terms of net incomes, tax credits and income deductions are utilized with survey data as input file and as benchmark for microsimulation results. Hence fiscal data and microsimulation estimates are both applied for reciprocal comparison and validation and for the construction of the final data set of gross incomes at individual and household level.

In order to employ the tax data and the survey data in the input file of the model, three relevant data sets are used: the “730 tax returns” generally filled in by employees and pensioners; the “UNICO tax returns” used by all the taxpayers and in particular by self employed, and the employers’ CUD statements. The CUD statements are included only for those taxpayers not obliged to fill in the tax

returns (employees or pensioners without tax credits due to consumption expenditures), according to the national income tax schedule.

The “730 tax returns” and the “UNICO tax returns” have provided data on net income, gross income, taxes at national and regional level and also data on tax credits and income deductions.

The available registers on compulsory social insurance contributions cover data on employees of private sector (not employers) and on employees and employers of public sector. Furthermore there is only a partial coverage of the social insurance contributions of self employed drawn on the UNICO tax returns.

Some adjustments were needed in order to use the integrated data set as input file of the model. Tax exempt incomes (i.e. disability benefits, family allowances, education related benefits, social assistance) and incomes subject to a separate taxation (severance pay or arrears) have to be calculated in order to exclude them from the total taxable income and avoid a current taxation rate. In fact, these income components are handled in the model as a component-specific deductions.

4.2 The final data set of gross income variables

The final data set of individual and household gross income variables is definitely the result of the mixed strategy of using in conjunction microsimulation and administrative data.

Survey data could be typically affected by under-reporting and tax data, from individual tax returns, generally have an incomplete coverage in respect of non taxable incomes collected through income survey. Furthermore register data may not display all the authorized deductions allowed in the calculation of taxable income (tax avoidance) and of course, do not take account of illegal tax evasion.

For the construction of EU-SILC gross income variables the starting point was the availability of data on withholding taxes and taxes paid for the surveyed individuals with non zero income in the administrative data.

However, the incomes of surveyed individuals not present in register data and the social insurance contributions only partially covered by administrative data could be estimated by SM2-EU-SILC. Instead of applying the microsimulation model only for those individuals not present in tax data, all the available information (survey and registers) have been used as input file of the model. In particular the administrative data in terms of net incomes, tax credits and income deductions have been finally utilized. In the microsimulation models, as in the previous SM2, the income deductions and tax credits based on consumption expenditure generally needed to be estimated by regression technique based on external sources.

After using all the existing information, the SM2-EU-SILC outputs have been compared with the available administrative gross figures in order to assess the quality of microsimulation estimates.

Moreover the comparison of the two data sets (SM2-EU-SILC outputs and tax data) has been very useful for detecting some irregularities in administrative data (i.e. self employed contributions) and after a validation with the National Accounts figures, the SM2-EU-SILC estimates have been preferred.

For what concerns the gross income data production process, SM2-EU-SILC integrated with survey and register data has estimated taxes and social insurance contributions for those individuals not present in register data due to errors in the record linkage procedure (errors in the identification numbers of individuals). Furthermore the model has simulated the employers’ social insurance contributions of the private sector and also the self employed’ social insurance contributions not fully covered by available administrative data.

The final EU-SILC individual and household gross income variables are computed as net amounts plus taxes and social insurance contributions provided by register data, if available, or estimated by SM2-EU-SILC. In order to anonymize the administrative data used, a stochastic component has been added to the withholding taxes and to the taxes paid from registers.

In more details, the final data set of gross income target variables has been built up as follow:

- a) **when the net administrative incomes are higher than the survey incomes**, the net and gross amounts of incomes, the taxes and the employees' social insurance contributions derived from register data, while the employers' social insurance contributions and the self employed contributions are estimated by SM2-EU-SILC. The final EU-SILC gross variables do not differ from the tax gross variables;
- b) on the opposite, **when the survey incomes are higher than the register data**, the net incomes are those taken from the survey (collected or imputed), while the taxes and the employees' social insurance contributions derived from register data. As in the "a" case, the employers' social insurance contributions and the self employed contributions are estimated by SM2-EU-SILC. The final EU-SILC gross variables can be different from the tax gross variables.

It is worth to mention 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 measure of illegal tax evasion. In fact it is not possible to distinguish between the legal tax avoidance, allowed by the national fiscal system, and the tax evasion. Moreover it is highly probable that individuals, who do not pay taxes at all, do not answer to an income survey.

5. MAIN RESULTS

Table 6 shows the distribution of estimated gross income by components. 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 much lower for income from work (60.8%) than for the other components, due to the social insurance contributions to which such income is subject. The ratio of net to gross taxable income of other incomes varies approximately from the low of 73.9% for property income, to 95.9% for various taxable benefits, to of course 100% for housing, social assistance and other tax-exempt benefits.

The distribution of gross income shows clearly that the main income component is represented by income from work (72.9%), followed by old age benefits (19.9%). The differences in gross and net distribution demonstrates that the tax burden is higher in income from work than the other components, like taxable benefits and property income.

Table 7 shows the comparison of EU-SILC data with figures by National Accounts. The table also reports the breakdown of total gross income into social insurance, tax and net components. On the average, net income, after tax and social insurance contributions including employers' contributions, accounts for 67.5% of total gross. The agreement of EU-SILC results and National Accounts figures is relatively good and let the EU-SILC results reasonably satisfactory. In effect, some aspects have to be considered in explaining the discrepancies with National Accounts. The 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, lower than the other income components. For this reason, the EU-SILC personal income and financial taxes turns out over-estimated of 1.4 percentage point. On the other hand EU-SILC survey presents an higher share of income from work compared with that of National Accounts, which consequently lower the EU-SILC gross taxable income rate (-1.4 percentage points). It should be noted that the National Account definition of income incorporates the imputed rent which is not yet included in the EU-SILC target income variables.

Finally, it is expected that the combined effect of the above mentioned aspects explains the difference on net income (-2.8 percentage points) between the two data sources.

Table 6 EU-SILC target variables: distribution of per capita income by component

		per capita amount		Ratio	% distribution	
		Gross	net	Net/gross	gross	Net
		(1)	(2)	(3)	(4)	(5)
Income from work		12,733	7,747	60.8	72.9	65.7
PY010	employee cash or near cash income	6,824	5,504	74.1	42.5	46.7
	employer's SI contribution	2,172			12.4	
	employee's SI contribution	606			3.5	
PY021	non cash employee income-company car	13	9	69.2	0.1	0.1
PY050	cash benefits or losses from self-employment	2,748	2,234	71.6	15.7	18.9
	Self-employed SI contribution	370			2.1	
Property income		457	350	76.6	2.6	3.0
HY090	interest, dividends, profit from capital investments in unincorporated business	250	197	78.8	1.4	1.7
HY040	income from rental of a property or land	207	153	73.9	1.2	1.3
Taxable benefits		4,018	3,434	85.5	23.0	29.1
PY090	unemployment benefits	274	236	86.1	1.6	2.0
PY100	old-age benefits	3,483	2,959	85.0	19.9	25.1
PY110	survivor' benefits	115	99	86.1	0.7	0.8
PY130	disability benefits	146	140	95.9	0.8	1.2
Tax-exempt benefits		259	259	100.0	1.5	2.2
PY140	education-related allowances	14	14	100.0	0.1	0.1
HY050	family related allowances	109	109	100.0	0.6	0.9
HY060	social assistance	16	16	100.0	0.1	0.1
HY070	housing allowances	8	8	100.0	0.0	0.1
HY080	regular inter-household cash transfer received	112	112	100.0	0.6	0.9
Total		17,466	11,789	67.5	100.0	100.0

Table 7 Comparison with National Accounts (N.A.): distribution of total gross income

	EU-SILC 2007 (income reference year 2006)		N.A.	Error (% point)
Gross including SI	17,466	100.0	100.0	
SI contributions	3,148	18.0	16.6	1.4
- Employers' contribution	2,172	12.4	11.9	0.5
- Employees' contribution	606	3.5	2.8	0.7
- Self-employment contribution	370	2.1	1.9	0.3
Gross taxable	14,319	82.0	83.4	-1.4
Personal income tax and financial tax	2,530	14.5	13.1	1.4
Net income	11,789	67.5	70.3	-2.8
Euro 2006, per capita.				

Sources. ISTAT: EU-SILC and National Accounts (2009)

REFERENCES

- Betti G, Verma V, Ballini F, Natilli M and Galgani S (2003) 'Statistical Imputation in Conjunction with Micro-Simulation of Income Data', *Rivista Italiana di Economia, Demografia e Statistica*, 58(3), 35-43.
- Canberra Group (2001), *Final Report and Recommendations*, Ottawa, 2001.
- Consolini P., Di Marco M., R. Ricci 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, 2006.
- Consolini P. (2008), *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, version downloadable at the web site: <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, Rome, March 2009.
- 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, 2006.
- Eurostat (2004) 'Income in EU-SILC: Net-Gross-Net Conversion; Common Structure of the Model; Model Description; and Application to ECHP Data for France, Italy and Spain', EU-SILC 133/04, Working Group on Statistics on Income and Living Conditions (EU-SILC) 29-30 March 2004, Luxembourg: Eurostat.
- Herzog T.N., Scheuren F.J., Winkler W.E. (2007), *Data quality and Record Linkage Techniques*, New York: Springer.
- Istat (2009) *National Accounts Years 1990-2007*, Rome: Istat.
- Newcombe H.B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford, UK: Oxford University Press.
- van der Laan P. (2000), *Integrating administrative registers and household surveys*, Netherlands Official Statistics vol.15, Summer 2000.