Using Administrative and Survey Data to Analyse Tax Evasion from Unregistered Labour

Alessandra Coli (University of Pisa, Italy)

Francesca Tartamella (Italian National Institute of Statistics)

Paper Prepared for the IARIW 33rd General Conference

Rotterdam, the Netherlands, August 24-30, 2014

Session 7A

Time: Friday, August 29, Morning

33nd IARIW General Conference, Rotterdam, The Netherlands.
August 24-29, 2014.

Using administrative and survey data to analyse tax evasion from unregistered labour

Alessandra Coli[1], Francesca Tartamella[2]

Draft - This version: July 2014

Abstract

Well-established methods are used to estimate hidden economy, which depend on the characteristics of a country's statistical system. According to the Italian method, hidden economy stems mainly from unregistered labour and from the underreporting of value added. This research aims at providing an alternative estimate of non-registered labour earnings (both from employees and self-employed) using a blending of data sources on households budgets (micro data from different data sources are integrated through deterministic record linkage techniques). One objective is to assess whether and to what extent the novel use of household budgets data sources may affect the current estimate of hidden economy. A further goal is to detect the typical features of individuals supplying irregular labour. Finally, we aim at providing estimates on the distribution of hidden income among households groups with different social, economic or demographical characteristics.

1. Introduction

The increasing use of national accounts statistics as the basis for levying contribution and distributing subsidies requires European Member States to ensure the "exhaustiveness" of their Gross Domestic Product (GDP) estimates. The production boundary, on which Gross Domestic Product (GDP) is defined, includes all production actually destined for the market, be it legal or illegal (as long as these activities are willingly engaged by buyers and sellers). Assuring a complete coverage of

[1] Alessandra Coli, Department of Economics & Management, University of Pisa; alessandra.coli1@ unipi.it
[2] Francesca Tartamella, Italian National Institute of Statistics (ISTAT); tartamel@istat.it
The views expressed in this paper are those of the authors and do not necessarily reflect the views of their affiliation institutes.

economic production is a tremendous challenge for national accountants since part of the economy is deliberately concealed or difficult to be captured by data sources. In official statistics, non-observed economy (NOE) includes the following kinds of production (OECD, 2002): i) underground or hidden production; ii) legal production activities characterized by a low level of organization (informal economy); iii) production from illegal economy; iv) productions omitted due to deficiencies in the basic data collection system (statistic underground).

In 2005-2006, the UNECE secretariat carried out a survey on countries methods in estimating non-observed economy in national accounts (UNECE, 2008). Forty-five countries answered the questionnaire, providing details on the estimation methods used. The UNECE survey shows that the size of the adjustments for NOE varies widely across countries. This depends obviously on the characteristics of a country's economy (higher/lower presence of hidden economy or of informal economy) but most likely also on the methods used for estimating NOE.

Currently, the Italian National Accounts (NAs henceforth) provide an estimate of the GDP stemming from non-observed economy but they do not allow one to analyse it according to the characteristics of individuals/households who benefit from this kind of income, nor to measure the impact of income from hidden economy on households income distribution. The purpose of this paper is to try overcoming these limits for a specific segment of hidden economy, i.e. for the income stemming from non-registered labour, be it from employees of self-employed. In Italy this income accounts for a relevant part of the total according to Istat most recent published estimates (about 6.5% of GDP in 2008, Istat 2010).

The paper is organized as follows. Section 2 provides a short description of the current Istat method used to estimate the underground economy. Section 3 describes an alternative method for detecting non-registered labour, which is mainly based on the record linkage of survey and administrative data on individuals/households. In Section 4 we apply multivariate statistical analysis in order to detect the typical features of non-registered workers. In Section 5 we estimate the impact of income from unregistered labour on households groups according to the households income quintile and according to the household's main source of income. Section 6 draws some final conclusions.

2. The current Italian method

The Italian method to measure NOE assumes that hidden economy stems mainly from: i) the use of non-registered labour ii) the under-reporting of turnover, due to the under-reporting of legal production and/or over-reporting of intermediate costs.

The estimate of non-registered labour relies on the comparison of the evidence stemming from people-focused data sources (population census, labour force survey, administrative data) and firms-focused data sources (census on enterprises, business budgets and other administrative records). The

former are used to estimate the supply of labour, whereas the latter allow one to calculate the demand of labour. Broadly speaking, the method assumes that the surplus of supply over demand is due to the presence of unregistered labour, once removed any other possible cause which may explain the supply-demand difference (i.e. differences in time of recording, in definitions of worker, in the territory where the worker lives and works etc.). The comparison (and estimation of non-registered labor thereof) is carried out within very detailed labor categories identified by the crossing of several characteristics (e.g. the kind of economic activity, the firm's size and the region). Thus, at a further stage, it possible to analyse GDP from THE hidden economy according to such characteristics.

The revaluation of turnover, uniquely relies on business surveys (and business administrative record – profits and losses accounts). The method is based on the analysis of the business' costs and receipts, following A. Franz (1985). The level of net enterprise's income is compared with the compensation of employees: if the former is lower, the firm is classified as "under-declaring" and its receipts are revalued. The underlying hypothesis is that net enterprise income should guarantee to self employed a remuneration not lower than the compensation of an employees working in the same economic activity field and with analogous working time.[3] If the economic flows are not coherent with this hypothesis, it is assumed that the self-employed did not declare all entries or pushed up the intermediate costs. The firms found in this condition are identified as under declaring and therefore are subject to revaluation (see Coli, Tartamella 2012 for details).

In this paper, we present an alternative method to estimate the income generated by non-registered labour. The first step requires an alternative estimate of the underlying volume of labour, which is the topic of the following section.

3. Integrating survey and administrative data to detect non-registered employment

In this section, we describe the method used to detect non-registered employees among the individuals sampled by the Italian Survey on Income and living conditions (It-Silc henceforth). The idea is that of checking whether individuals earning income from labor according to the survey, are recorded in (at least one of) the administrative archive(s) stemming from the fiscal and/or social security's obligations of registered employees. To this end, we apply an exact record linkage procedure, which compares records contained in the It-Silc and administrative archive datasets, in order to determine pairs of records pertaining the same worker. Whenever this happens, the individual is classified as registered employee, non-registered otherwise. Before describing the method in detail, it is necessary to give some details on the data sources involved in the record linkage procedure.

[3] In fact, if the self employed should be in a position to earn less than an employees with the same characteristics, then the hypothesis is that he/she would prefer to modify his/her occupational status from self-employed to employees, to increase his/her income.

3.1 The data

It-Silc provides cross-sectional and longitudinal data. Cross-sectional data focus on income, poverty, social exclusion and living conditions, whereas longitudinal data cover only part of the topics. The questionnaire consists of the REG, FAM and IND forms. The REG form collects personal information on each member of the family such as gender, date of birth, occupational status, main source of income, etc. The FAM form is the household questionnaire. It collects data on lodging, housing costs and on the general economic status of the family. Finally, the IND form (Individual questionnaire) must be compiled by members aged 15 or more at the moment of the interview. The questionnaire collects information mainly on education, health, status in employment, income and saving. The sample is selected on the basis of a two-stage sample design (municipalities first and then households). Municipalities were selected once and for all in 2004 (first wave) from homogeneous strata in terms of demographic size and geographical region. Households are randomly selected each year from the municipalities' register offices. The sample renewal concerns the elementary units only: municipalities update their samples partially or totally depending on the stratum they belong to. The It-Silc is carried out yearly, on the basis of a face-to-face interview. Furthermore, Istat uses administrative data to reduce the measurement errors, integrating the It-Silc and revenue taxes information (Consolini, 2008). Sampling weights are calibrated to external data sources. The known totals concern the distribution of people by sex, age class, geographical area and demographic size of the municipality.

So It-Silc data are already integrated with administrative data, but fiscal data are not necessarily exhaustive since not all earners have to fill in fiscal declarations and pay only withholding taxes. Administrative data from employer side, instead, have a more likelihood to cover all working transaction, since employer has to act as withholding agent to pay taxes and contribution to fiscal authorities and social security institutions, even when the worker is not asked to pay additional taxation.

We could dispose of several administrative archive from the firm side, mainly related to social security obligation, signaling, with a different degree of strength, a working relationship between a firm and a worker. Each of this archive has individual data and the fiscal code of both the individual and the enterprise and therefore allows to perform a record linkage with individuals interviewed in a household survey and the firms/institutions in the enterprise/institution archive. Moreover, these archives better allow to isolate different jobs relating to the same person.

Each source has a different detail in the information provided, in terms of:

- period length of the working relationship,
- characteristics of the working relationship and type of job,
- payment.

In detail, the administrative archives used for the record linkage are, for *employees*:

1. Social security archive of individual insured position for workers employed in the private sector (INPS_e): each enterprise in the private sector (excluding those in agriculture) with at least one employee each month has to fill in a form for the social security institution (INPS) with all information to compute the contribution due to INSPS (both from enterprise and employees) and the benefits (family allowances, CIG, maternity allowance, illness allowances etc) that the employer or INPS (directly or through the employer) has to pay to the worker. It therefore contains all information about worker (fiscal code, residence), the firm, the characteristics of the jobs that affect contribution (position, type of job), date of start and end (if existing) of the working period, type of contract, number of paid days, earnings and, for each week of the month, number of worked days and not worked if this affects contribution paid or benefits received (from employer or Inps)
2. Social security archive of individual insured position for workers employed in the public sector (INPDAP_e): it contains information about the employer and the employee, the date of start and end of the working relationship, but no information on earnings.
3. Social security archive of individual insured position for workers employed in the private sector of sport, arts and entertainment (ENPALS_e): it contains information about the employer and the employees, the type and category of activity performed, the date of start and end of the working relationship, the earnings and contribution paid.
4. Social security archive of individual insured position for workers employed as domestic staff (INPS_d): it contains the fiscal code of worker and the employer and, for each quarter of the year, the number of weeks paid in the quarter, the number of hours paid in the quarter, hourly earnings, the total earning and contribution of each quarter.
5. Social security archive of individual insured position for workers employed in agricultural sector (INPS_agr): it contains the fiscal codes of workers and enterprise, date of hiring and firing, year and quarter of reference and days worked in that period, if part time the number of weekly working hours and the earnings
6. Social insurance archive (INAIL_e) for all employees: it contains all information about employer and employees and relationship (data start, data end) and information that affect the amount paid as insurance contribution (position, type of job, type of contract). This archive does not always provide reliable date for working period.
7. Social insurance archive (INAIL_aw) for agency workers: same information as the previous archive plus information on the firm where the agency worker is employed (with data start and end)

For *self-employed*:

1. Social security archive (INPS_out) for outworkers: it contains information on each job performed as outworker, the hiring firm and the time length of the professional service. It also contains information about the type of work performed, earning and contribution paid.
2. Social insurance archive (INAIL_out) for outworkers
3. Social security archive (INPS_se_agr) for self-employed working in agriculture: it contains all information about the enterprise (fiscal code, name address etc) and the worker (fiscal code, name, address, place and date of birth, residence), year and number of days worked in the year.
4. Social security archive (INPS_p) for professionals and freelancers: it contains information on each job performed as outworker, the hiring firm and the time length of the professional service. It also has information about the type of work performed, earning and contribution paid.
5. Archive built in Istat for the administrative enterprises census that synthetize information from fiscal agencies (VAT numbers) and chamber of commerce (partners of corporations and persons that have some implication in the administration of a partnership or a corporation (SE).

3.2 Record linkage

Record linkage is a technique, which compares records contained in two files A and B, in order to determine pairs of records pertaining the same population unit. Through record linkage, record pairs are singled out and recorded in a unique archive (matched file), which contains information from A and B for the each linked unit.

In order to apply record linkage the A and B files must have a non-empty set of units in common; furthermore A and B must have an identifier variable in common or a set of variables (k variables) which jointly permit to identify a population unit univocally.

Record linkage between two files is very simple, provided that each record in both files contains the same identifier and this identifier is recorded without errors. In this case, the problem is solved by simply picking out the records (if any) with the same identifier value. Obviously, errors may occur because the identifier variable is incorrectly recorded or some values of the k variables may be missing so that the K-variable may not be known exactly for some of the records in *A* or *B* (Copas and Hilton 1990). Due to such errors, two records for the same unit may not agree, and two records which agree may refer to different units.

In this work, we aim at linking units (persons) from the IT-Silc survey (FI) and from a set of administrative archives separately. We use the Italian fiscal code (Codice Fiscale) as identifier, an alphanumeric code of 16 characters, with characters reflecting personal information like name,

surname, date and place of birth. The fiscal code is a good identifier in that it identifies univocally each person; furthermore it is possible to detect recording errors by checking the values of variables (when available) from which the fiscal code derives. Assuming that errors are negligible, we apply a deterministic linkage instead of probabilistic record linkage.

The result of the record linkage procedure is shown in table 1. Column A shows the incidence of interviewed found in each archive, out of the about 45 thousands interviewed, while column B shows the incidence out of the about 22 thousands interviewed found in at least one archive (these persons can be labelled as registered worker). If a person is found in more than one archive it is counted in each archive. The table in the last row, as Asia, displays the number of self_employed that figure as self employed in the register of active enterprises. So while SE archive is a list of "potential" self employed, Asia is a list of "actual" self-employed.

The result is that 44.5% of individuals interviewed are found in at least one archive. Some archive should overlap by definition (notably Inps and Inail: each individual has to pay both social security and social insurance contribution).

Table 1: results of the micro linkage

	A incidence of micro linkage on all Interviewed	B incidence of micro linkage on registered workers
INPS_e	24.0	54.0
INPDAP_e	6.5	14.5
ENPALS_e	0.4	0.9
INPS_d	0.3	0.7
INPS_agr	1.3	2.8
INAIL_e	27.5	61.9
INAIL_aw	0.7	1.6
INPS_out	2.4	5.3
INAIL_out	1.9	4.4
INPS_se_agr	0.8	1.7
INPS_p	0.4	0.9
SE	14.8	33.2
ASIA	10.0	22.5
ALL (at least 1 linkage)	44.5	100.0

Each administrative source has a different degree of reliability, not in terms of the relationship individual-firm (or institution), but mainly in terms of the time spam of the actual working activity. For example, the fact that a person is recorded as artisan, does not necessarily imply that a service has been actually rendered in the analyzed period. At the same time these archives can more easily capture the existence of an actual job even when it is not reported (or under-reported) to fiscal authority.

Consequently, the individual-enterprise relationship cannot be simply labeled as present/non present, but also in weak/strong. For employees the weak presence happens when the only source where the employee is present is the archive of social insurance, which is updated less frequently, while we assume a strong presence if we find the employee in the social security archive. For self-employed, the strong presence happens when the worker is defined as self-employed in the archive of active enterprises (Asia) or, again, when it is present in the social security archive, that reports precisely the start and the end of each position (Inps_p).

From the survey side, we defined as worker for IT-Silc (employee or self-employed) those who perceived some income (of the corresponding type) in the survey, so that if a person perceived both the income type, it is present in both status, as employees and as self-employed, (as it is present in both administrative record). The administrative data are by position, while in the survey a person is classified as employee or self-employed according to the main labour activity, but all incomes are reported.

Therefore, according to the presence in the administrative archive and the definition as worker/non worker, the IT-Silc individual can be labelled as:

1. Registered worker: presence in the administrative archive (strong or weak) and identified as worker in IT-Silc;
2. Non-registered worker or under coverage of the administrative source: absence in the administrative archive and worker in IT-Silc
3. Mis-reporter in the survey or over-coverage of the administrative source (presence in the administrative record, not worker for IT-Silc).

Cases 2 and 3 could be as well interpreted as mistakes in the fiscal code, from both sides.

Table 2 shows the results of the micro-linkage. We did not include all interviewed, but those who can be defined as employed in at least one source (the rest of the sample are not employees or self-employed in both sources). The tables show therefore the percentage distribution of those who can be defined as employed for at least one source, according to their presence in an administrative archive and in the survey and as employed according to the presence of a labour income in IT-Silc.

Note that we simply classified individual in employed- not employed (from both sources) and we did not divided the sample in employees and self-employed because if the categorization of income according to the administrative archive is quite straightforward, the same does not apply for IT-Silc. The reason is twofold. For those record derived from fiscal source the classification of the labour income can be not always correct for our purposes, since some income that in registers (and in NA) are labelled as self-employment income are fiscally treated as employees income (they are "assimilated" to employees income from a fiscal point of view). Moreover also when the classification is made directly by the interviewed the "perception" can be different since sometimes some actual employee jobs are masked as self-employment (and registered as self-employment) income when the employer have a

fiscal advantage, so that in the survey the actual subordination relationship is reported and self-employement income are reported as employees. Also the opposite is true: some self-employed record themselves as employees and derive from their enterprise an employee income plus a share of the profit. So the actual relationship with the enterprise is as self-employed and therefore they report their income accordingly, but are registered in administrative archive as employee. This lead to a different classification in the administrative archive and IT-Silc and a possible mis-match, so that separating the sample into employees and self-employed would lead to over-counting the groups 2 and 3 listed above.

Table 2: Persons employed in Esuilc and presence in administrative data.

		Employed in Administrative sources			
		No	Weak	Strong	Total presence in Adm. sources
Employed in IT-Silc	No	0.0	5.9	2.5	8.4
	Yes	9.5	8.8	73.3	91.6
	Total presence in It-Silc	9.5	14.7	75.8	100.0

The table shows that the presence in administrative record and at the same time absence of income in the survey is less likely than the opposite situation i.e. employed in the survey and not present in administrative records. This could be interpreted as under-coverage of survey or over-coverage of the administrative source, even if the fact that those found only in the administrative sources come from a "strong" administrative archive, makes plausible the under-coverage of the survey: it is true that the integration of survey data with fiscal data performed on IT-Silc should remedy the under-reporting, but fiscal record data may be less timely than the administrative archive used in the micro linkage, moreover not everybody is compelled to fill in a fiscal declaration. In most cases, for those we found in a "strong" administrative source, those employed worked for a short period of time (one month or less) and the amount involved is small. This confirms that it is less likely that small amounts and short period of work are reported in the survey.

Non-registered workers, i.e. workers not present in any administrative sources, represent about 9.6% of persons with labour income in IT-Silc, they are more frequent among employees (about 60%, this makes an incidence of about 12% on IT-Silc self-employed and 7% on IT-Silc employees). These values are highly comparable with the published incidence of non-registered worker, according to NA data (Istat, 2011), is 10.3 in terms of employed persons and 17% in terms of jobs. The incidence is higher for employees (11.2% for employees, 18.8 in terms of jobs) than for self-employed (10.2 in terms of employed persons, 17.2% in terms of jobs).

4. What are the typical features of registered/non-registered labour?

Current official statistics analyse the distribution of registered/unregistered labour by economic activity and separating employed from self-employed labor. The analysis is made for different unit of analysis, i.e. for heads, jobs and equivalent labour units. Istat also provides an insight on the territorial distribution of registered and unregistered labour units. On the contrary, information lacks on other social and demographic characteristics of the non-registered workers such as gender or age. Information is also lacking on the characteristics of the families whom non-registered workers belong.

In this section, we present the result of a statistical analysis aimed at detecting the specific features of the registered/unregistered employee. Particularly, we apply a logistic regression on the Ii-Silc records relating to individuals aged 16 years or more who declared to had earned income from labour in 2010. The aim is to detect the characteristics of individuals, which better help explaining their belonging to the registered/non registered groups. Table 3 describes the variables used in the analyses.

Table 3 Description of the variables used in the logistic regression and/or cluster analyses

Name	Description	Category	Meaning of the category
reg	registered/non-registered worker	0	non-registered worker
		1	registered worker
quanty	Personal net disposbale income	1	1st quintile
		2	2nd quintile
		3	3rd quintile
		4	4th quintile
		5	5th quintile
quantfy	Household net disposbale income	1	1st quintile
		2	2nd quintile
		3	3rd quintile
		4	4th quintile
		5	5th quintile
Region	Region (nust2) where the interviewed persons (families) lives	10	Piemonte
		20	Val d'Aosta
		30	Lombardia
		41	Bolzano
		42	Trento
		50	Veneto
		60	Friuli Venezia Giulia
		70	Liguria
		80	Emilia Romagna
		90	Toscana
		100	Umbria
		110	Marche
		120	Lazio
		130	Abruzzo
		140	Molise
		150	Campania
		160	Puglia
		170	Basilicata
		180	Calabria
		190	Sicilia
		200	Sardegna
Gender		1	male
		2	female
YWORK	Number of years spent in paid work		
AGE	Age in 2010		

Table 3 –continues

CITIZ	Italian/ foreigner	1	Italian
		2	Foreigner
FTPT	full time/part time	1	Full time
		2	Part time
NJOBS	one job/more jobs	1	one job
		2	more than one job
ACTIVITY_1	Economic activity of the local unit of the main job. Nace_Rev2 Sections.	A	Agriculture, forestry and fishing
		B	Mining and quarrying
		C	Manufacturing
		D	Electricity, gas, steam and air conditioning supply
		E	Water supply, sewerage, waste management and remediation activities
		F	Financial and insurance activities
		G	Construction
		H	Wholesale and retail trade; repair of motor vehicles and motorcycles
		I	Accommodation and food service activities
		J	Transportation and storage
		K	Information and communication
		L	Real estate activities
		M	Professional, scientific and technical activities
		N	Administrative and support service activities
		O	Public administration and defence; compulsory social security
		P	Education
		Q	Human health and social work activities
		R	Arts, entertainment and recreation
		S	Other service activities
		T	Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use
		U	Activities of extraterritorial organisations and bodies

Table 3- continues

ACTIVITY_2	Economic activity of the local unit of the main job. Aggregations of Nace_rev2 Sections.		
		AGR	A
		MAN	B+C+D+E
		COS	F
		RETAIL	G+H+I+J
		FIN	K
		PROF	L+M+N
		PA+OS	O+P+Q+S+U
		HH	T
CLYWORKED	Class of years worked	1	<9
		2	10-15
		3	16-22
		4	23-31
		5	32 and more
TIPOLAV	Status in employement	1	employees
		2	self-employed
		3	earned income from both employee and self-employed labour
CLAGE	Class of age	1	16-32
		2	33-40
		3	41-47
		4	48-55
		5	56 and more
CWORKERS	Number of persons working at the local unit where the person works	1	1-10
		2	11-19
		3	20-49
		4	50 and more

The response variable (*reg*) is whether the subject is a registered worker (1 = yes, 0 = no). The model contains several explanatory variables. Two variables measure the position of the worker in the ranking of personal and households disposable incomes (*quinty* and *quintfy* respectively). Five variables concern personal characteristics of the worker (*gender*, *age*, *region*, *citiz, edu*). Six variables relates to labour characteristics (*yworked*, *FTPT*, *NJOBS*, *Activity_1*, *Tiplav*, *Cworkers*). Tables 4a and 4b present the results of the analysis.[4] The first table shows the change in deviance obtained by adding each of the terms in the order listed in the model formula. A chi-test is performed to assess whether the contribute of each term is significant. We can see that all terms were highly significant (P-value< 0.001) when they were

[4] We used the R software -R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

introduced into the model, with the only exception of education attained (*edu*), whether the worker has a full-time or a part-time job (*FTPT*) and whether he has got one or more than one job (*NJOBS*).

Table 4a Change in deviance obtained by adding each of the variables

	Df	Deviance	Resid.Df	Resid.Dev	Pr(>Chi)	
NULL			15343	9797.4		
quanty	4	833.95	15339	8963.5	<2.20E-16	***
quantfy	4	158.98	15335	8804.5	<2.20E-16	***
Gender	1	8.53	15334	8796	0.0034976	**
Region	20	236.17	15314	8559.8	<2.20E-16	***
Yworked	1	11.81	15313	8548	0.0005894	***
age	1	24.6	15312	8523.4	7.06E-07	***
Citiz	1	91.88	15311	8431.5	<2.20E-16	***
Cworkers	3	166.77	15308	8264.7	<2.20E-16	***
tipolav	2	299.29	15306	7965.4	<2.20E-16	***
Activity_2	7	302.26	15299	7663.2	<2.20E-16	***
edu	2	2.25	15297	7660.9	0.3243601	
FTPT	1	0.26	15296	7660.7	0.6133828	
NJOBS	1	0.12	15295	7660.6	0.733253	
Pseudo-R^2 = 0.22						

The model does not seem to fit data very well (Pseudo-R2 = 0.22). However, our purpose here is to discover which characteristics are significantly associated with the "propensity to be a registered worker instead of non-registered workers". To this end, it is necessary to look at the tests on coefficients shown in Table 4b.

Looking at the sign of coefficients, we notice that the probability of being a registered workers increases with the number of years worked (positive sign for *yworked*) and decreases with the age of the workers (negative sign for *age*). We also notice that female workers have a higher probability of being registered workers with respect to male workers (positive coefficient for *gender* category 2). Analogously, workers employed in bigger local units have a higher probability to be registered than non-registered with respect to workers employed in firms with 9 employees at most (positive coefficient for Clworkers, all categories). The opposite happens for foreign workers with respect to Italian workers (negative coefficient for *citiz* category 2) and for self-employed with respect to employees (negative coefficient for *citiz* category 2). The probability of success (*reg*=1) decreases with the increase of personal net disposable income (all coefficients are positive for *quanty* categories) whereas only the third and fourth quintiles of households disposable income seem to imply a higher propensity to registered against non-registered labour with respect to the first quintile. All regions with p-value lower than 0.05 (one star), present negative coefficients, i.e. a lower probability of success with respect to Piemonte (*region*=1). Such regions are all in the South of Italy with the exception of Toscana and Lazio (Centre

Italy). Only the province of Trento, Veneto and Emilia Romagna perform better than Piemonte. Finally, it is worth noting how "Agricolture" is one of the economic activity with lower probability of success. Only workers employed by households (*HH*) seem to present an even lower probability to be registered.

Table 4b Logistic regression – Tests on the coefficients.

	Estimate	**Std. Error**	**Z**	**Pr(>|z|)**	
(Intercpet)	0.523172	0.302947	1.727	0.084178	
quanty(2)	1.000764	0.090615	11.044	< 2.00E-16	***
quanty (3)	1.474842	0.107837	13.677	< 2.00E-16	***
quanty (4)	1.405335	0.110211	12.751	< 2.00E-16	***
quanty(5)	2.013311	0.128089	15.718	< 2.00E-16	***
quantfy2	0.154283	0.086886	1.776	0.075781	.
quantfy3	0.427271	0.098626	4.332	1.48E-05	***
quantfy4	0.256724	0.103324	2.485	0.012968	*
quantfy5	0.069683	0.108132	0.644	0.5193	
Gender (2)	0.289929	0.072144	4.019	5.85E-05	***
Region (20)	0.202015	0.292623	0.69	0.48997	
Region (30)	-0.031395	0.163166	-0.192	0.847421	
Region (41)	-0.340832	0.243451	-1.4	0.161513	
Region (42)	0.795698	0.358211	2.221	0.02633	*
Region (50)	0.519315	0.188279	2.758	0.005812	**
Region (60)	0.021773	0.194229	0.112	0.910745	
Region (70)	-0.133664	0.195295	-0.684	0.49371	
Region (80)	0.49743	0.193233	2.574	0.010046	*
Region (90)	-0.394054	0.169938	-2.319	0.020405	*
Region (100)	0.057343	0.199284	0.288	0.773541	
Region (110)	0.527956	0.216225	2.442	0.014618	*
Region (120)	-0.887055	0.154951	-5.725	1.04E-08	***
Region (130)	0.060899	0.264232	0.23	0.817722	
Region (140)	-0.49855	0.23127	-2.156	0.031107	*
Region (150)	-1.104617	0.161804	-6.827	8.68E-12	***
Region (160)	-0.444123	0.184099	-2.412	0.015847	*
Region (170)	-0.659917	0.229767	-2.872	0.004077	**
Region (180)	-0.370454	0.200838	-1.845	0.065104	.
Region (190)	-0.318443	0.185361	-1.718	0.085805	.
Region (200)	-0.459606	0.208149	-2.208	0.02724	*
yworked	0.025509	0.005195	4.91	9.10E-07	***
Age	-0.010894	0.004855	-2.244	0.024856	*
Citiz(2)	-1.002991	0.109162	-9.188	< 2.00E-16	***
Clworkers(2)	0.222783	0.096449	2.31	0.020897	*
Clworkers(3)	0.766851	0.13195	5.812	6.18E-09	***
Clworkers(4)	0.38182	0.095351	4.004	6.22E-05	***
Tipolav (2)	-1.343907	0.081612	-16.467	< 2.00E-16	***
Tipolav (3)	0.677629	0.187083	3.622	0.000292	***

Table 4b - continues

	Estimate	Std. Error	Z	Pr(>|z|)	
Activity_2 (cos)	1.012274	0.157406	6.431	1.27E-10	***
Activity_2 (fin)	1.682004	0.290627	5.787	7.14E-09	***
Activity_2 (pa+os)	0.481921	0.133949	3.598	0.000321	***
Activity_2 (HH)	-1.618169	0.269195	-6.011	1.84E-09	***
Activity_2 (MAN)	1.18021	0.142481	8.283	< 2.00E-16	***
Activity_2 (PROF)	1.405405	0.156926	8.956	< 2.00E-16	***
Activity_2 (RETAIL)	1.234073	0.130894	9.428	< 2.00E-16	***
Edu(2)	-0.047545	0.07141	-0.666	0.505536	
Edu (3)	0.087903	0.104854	0.838	0.40184	
FTPT (2)	-0.044333	0.083378	-0.532	0.594924	
NJOBS (2)	-0.069492	0.20493	-0.339	0.734534	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The results of the analysis confirm (in line with Istat statistics on labour) that non registered labour is more present in the South of Italy, in small firms and in some specific economic activities such as Agricolture, Housholds activities as employers, Other services and Construction.

5. An alternative estimate of income from non-registered labor and its impact on the distribution of disposable income

In this final section, we present some analysis on the incidence of income of non-registered worker on total household income and sub component by quintile, main source of income, geographical area and householder age group. Total gross household income is computed as the sum of the sub-component: wages and salaries, self-employment income, property income (excluding imputed rents), social benefits and other transfers.

More in detail, we present four sets of tables. The first set of tables focuses primarily on households with at least one non-registered worker. Such tables display:

- the total amount of self employment income earned by each group in percentage of the total self employment income earned by all households in the group (quintile, age group etc)
- the quota of this self employment income derived from non registered workers
- the total amount of wage and salaries earned by each group in percentage of the total wage and salary earned by all households in the group (quintile, age group etc)
- the quota of this wage and salary derived from non registered workers
- the total amount of property income earned by each group in percentage of the total property income earned by all households in the group (quintile, age group etc)
- the total amount of social benefits and other transfers earned by each group in percentage of the social benefits and other transfers earned by all households in the group (quintile, age group etc)

- the total amount of taxes and contribution paid by each group in percentage of the taxes and contribution paid by all households in the group (quintile, age group etc)
- the total amount of net household income earned by each group in percentage of the net household income earned by all households in the group (quintile, age group etc)
- the number of households with at least one non registered worker in each group in percentage of the total number of households in the group (quintile, age group etc)
- the non registered self employment income in percentage of the total self employment income earned by all households in the group (quintile, age group etc)
- the non registered wage and salaries in percentage of the total wage and salary earned by all households in the group (quintile, age group etc)

The second set of tables, shows, always for households with at least one non-registered worker, the incidence of each group in the total income flows, the third set of table display the same by group income composition, for households without non-registered worker. The last set of table, finally, shows the incidence of each group in the total income flows for all households.

The main evidences are:

- From tables 2s we see that most of non registered incomes are in high income families, living in the center and in the south of Italy, with householder below 50 years.
- From tables 1s we see that the incidence of non registered income on the corresponding self employment or employee income of households with at least one non registered worker is very high for the first quintile (in this quintile almost all labour income is non registered), for households whose main source of income is not labour. It is above average in the Centre and South of Italy and for older households.
- With respect of the distribution of taxation by group of households without non registered workers (tables 3s), households group with higher incidence of non registered income shows lower taxation quota (tables 2s).

Tab 5 Households by quintile

Tab 5.1 and 5.2: households with at least one non-registered employed

	self employment income (on total self employment income of the quintile)	incidence of non registered self employment income on self employment income	wages and salaries (on total self employment income of the quintile)	incidence of non registered wages and salaries on wages and salaries	property income (on total property income of the quintile)	Social benefits and other transfers on total social benefit of the quintile	total tax and contribution paid by the quintile	total net income (on total income of the quintile)	n. of households (on total number of hh of the quintile)	incidence of non registered self employment income on total self employment income of the quintile	incidence of non registeredwages and salaries on total wages and salaries of the quintile
1	23.1%	86.4%	24.0%	92.0%	-10.5%	1.6%	2.0%	10%	11.2%	19.9%	22.1%
2	14.5%	63.7%	8.2%	72.1%	-1.7%	2.6%	2.6%	7%	6.8%	9.2%	5.9%
3	14.4%	64.2%	10.1%	57.5%	-0.5%	4.0%	4.0%	9%	9.0%	9.2%	5.8%
4	14.6%	46.5%	8.6%	39.4%	1.4%	7.3%	6.0%	10%	9.9%	6.8%	3.4%
5	10.4%	37.4%	10.2%	29.1%	7.6%	9.7%	7.5%	11%	10.7%	3.9%	3.0%
total	12.1%	46.8%	9.9%	42.6%	5.2%	6.0%	6.2%	10%	9.5%	5.7%	4.2%

	self employment income	non registered self employment income	wages and salaries	non registered wages and salaries	property income	social benefits and other transfers	tax and contribution	total net income	n. of households
1	5%	9%	6%	13%	-1%	3%	1%	6%	24%
2	8%	10%	7%	12%	-1%	7%	3%	8%	14%
3	11%	15%	16%	21%	-1%	14%	8%	15%	19%
4	21%	21%	24%	22%	5%	26%	22%	23%	21%
5	55%	44%	48%	33%	98%	50%	66%	48%	22%
tot	100%	100%	100%	100%	100%	100%	100%	100%	100%

Tab 5.3 Households without non-registered employed

	self employment income	non registered self employment income	wages and salaries	non registered wages and salaries	property income	social benefits and other transfers	tax and contribution	total net income	n. of households
1	2%		2%		1%	11%	2%	6%	20%
2	6%		8%		4%	18%	8%	12%	21%
3	9%		15%		10%	21%	13%	16%	20%
4	17%		28%		20%	21%	23%	24%	20%
5	65%		47%		66%	30%	54%	43%	20%
tot	100%	0%	100%	0%	100%	100%	100%	100%	100%

Tab 5.4 all households by quintile

	self employment income	non registered self employment income	wages and salaries	non registered wages and salaries	property income	social benefits and other transfers	tax and contribution	total net income	n. of households
1	3%	9%	2%	13%	0%	10%	2%	6%	20%
2	6%	10%	8%	12%	4%	17%	7%	11%	20%
3	9%	15%	15%	21%	9%	21%	13%	16%	20%
4	18%	21%	27%	22%	19%	21%	23%	24%	20%
5	64%	44%	47%	33%	68%	31%	54%	43%	20%
tot	100%	100%	100%	100%	100%	100%	100%	100%	100%

Tab 6 Households by main source of income

Tab 6.1 and 6.2: households with at least one non registered employed

	self employment income (on total self employment income of the group)	incidence of non registered self employment income on self employment income	wages and salaries (on total self employment income of the group)	incidence of non registered wages and salaries on wages and salaries	property income (on total property income of the group)	Social benefits and other transfers on total social benefit of the group	total tax and contribution paid by the group	total net income (on total income of the group)	n. of households (on total number of hh of the group)	incidence of non registered self employment income on total self employment income of the group	incidence of non registeredwages and salaries on total wages and salaries of the group
Wages and salaries	13.7%	49.7%	9.8%	42.9%	15.6%	11.0%	6.2%	11%	12.7%	6.8%	4.2%
Self employment income	11.9%	44.0%	11.7%	24.5%	7.8%	14.8%	7.6%	14%	17.9%	5.2%	2.9%
Property income	4.3%	69.1%	8.2%	15.1%	2.8%	1.2%	2.6%	3%	2.7%	3.0%	1.2%
Social benefits and other incomes	12.6%	71.9%	9.9%	56.6%	3.9%	4.4%	4.8%	5%	3.6%	9.0%	5.6%
total	12.1%	46.8%	9.9%	42.6%	5.2%	6.0%	6.2%	10%	9.5%	5.7%	4.2%

	self employment income	non registered self employment income	wages and salaries	non registered wages and salaries	property income	social benefits and other transfers	tax and contribution	total net income	n. of households
Wages and salaries	14%	15%	88%	88%	20%	28%	53%	56%	58%
Self employment income	79%	74%	6%	4%	33%	14%	30%	26%	26%
Property income	0%	0%	0%	0%	16%	0%	1%	0%	0%
Social benefits and other incomes	7%	10%	6%	8%	31%	58%	16%	17%	16%
total	100%	100%	100%	100%	100%	100%	100%	100%	100%

Tab 6.3 Households without non-registered employed

	self employment income	non registered self employment income	wages and salaries	non registered wages and salaries	property income	social benefits and other transfers	tax and contribution	total net income	n. of households
Wages and salaries	12%		89%		6%	14%	53%	48%	42%
Self employment income	80%		5%		21%	5%	24%	18%	12%
Property income	1%		0%		31%	1%	1%	1%	1%
Social benefits and other incomes	7%		6%		42%	80%	21%	33%	44%
total	100%	0%	100%	0%	100%	100%	100%	100%	100%

Tab 6.4 all households by main source of income

	self employment income	non registered self employment income	wages and salaries	non registered wages and salaries	property income	social benefits and other transfers	tax and contribution	total net income	n. of households
Wages and salaries	13%	15%	89%	88%	7%	15%	53%	49%	44%
Self employment income	80%	74%	5%	4%	22%	5%	25%	19%	14%
Property income	1%	0%	0%	0%	30%	1%	1%	1%	1%
Social benefits and other incomes	7%	10%	6%	8%	41%	79%	21%	31%	41%
total	100%	100%	100%	100%	100%	100%	100%	100%	100%

Tab 7 Households by geographical area

Tab 7.1 and 7.2: households with at least one non registered employed

	self employment income (on total self employment income of the area)	incidence of non registered self employment income on self employment income	wages and salaries (on total self employment income of the area)	incidence of non registered wages and salaries on wages and salaries	property income (on total property income of the area)	Social benefits and other transfers on total social benefit of the area	total tax and contribution paid by the area	total net income (on total income of the area)	n. of households (on total number of hh of the area)	incidence of non registered self employment income on total self employment income of the area	incidence of non registeredwages and salaries on total wages and salaries of the area
North-West	10.8%	39.5%	8.0%	37.4%	4.1%	4.6%	5.6%	8%	7.2%	4.3%	3.0%
North-East	7.3%	38.9%	6.5%	32.8%	5.6%	4.3%	4.6%	7%	5.7%	2.9%	2.1%
Centre	14.9%	54.1%	12.2%	48.2%	6.5%	7.1%	7.2%	12%	11.8%	8.0%	5.9%
South	18.3%	54.3%	14.6%	43.7%	6.4%	8.8%	8.6%	14%	13.6%	9.9%	6.4%
Islands	12.6%	40.8%	11.8%	54.1%	2.6%	5.3%	5.8%	10%	10.5%	5.1%	6.4%
total	12.1%	46.8%	9.9%	42.6%	5.2%	6.0%	6.2%	10%	9.5%	5.7%	4.2%

	self employment income	non registered self employment income	wages and salaries	non registered wages and salaries	property income	social benefits and other transfers	tax and contribution	total net income	n. of households
North-West	30%	25%	26%	23%	27%	23%	31%	26%	22%
North-East	14%	12%	15%	11%	20%	14%	17%	14%	12%
Centre	26%	31%	25%	29%	29%	25%	25%	26%	25%
South	23%	27%	24%	25%	20%	29%	21%	26%	30%
Islands	6%	5%	9%	12%	3%	8%	6%	9%	12%
total	100%	100%	100%	100%	100%	100%	100%	100%	100%

Tab 7.3 Households without non-registered employed

	self employment income	non registered self employment income	wages and salaries	non registered wages and salaries	property income	social benefits and other transfers	tax and contribution	total net income	n. of households
North-West	34%		33%		36%	31%	34%	32%	29%
North-East	25%		24%		19%	20%	23%	23%	21%
Centre	21%		20%		23%	21%	21%	20%	19%
South	14%		16%		16%	19%	15%	17%	20%
Islands	6%		8%		6%	9%	7%	8%	11%
total	100%	0%	100%	0%	100%	100%	100%	100%	100%

Tab 7.4 all households by geographical area

	self employment income	non registered self employment income	wages and salaries	non registered wages and salaries	property income	social benefits and other transfers	tax and contribution	total net income	n. of households
North-West	34%	25%	32%	23%	35%	30%	34%	31%	29%
North-East	24%	12%	23%	11%	19%	20%	23%	22%	20%
Centre	22%	31%	21%	29%	23%	21%	21%	21%	20%
South	15%	27%	17%	25%	17%	20%	15%	18%	21%
Islands	6%	5%	8%	12%	6%	9%	7%	8%	11%
total	100%	100%	100%	100%	100%	100%	100%	100%	100%

Tab 8 Households by householder age group

Tab 8.1 and 8.2: households with at least one non registered employed

	self employment income (on total self employment income of the group)	incidence of non registered self employment income on self employment income	wages and salaries (on total self employment income of the group)	incidence of non registered wages and salaries on wages and salaries	property income (on total property income of the group)	Social benefits and other transfers on total social benefit of the group	total tax and contribution paid by the group	total net income (on total income of the group)	n. of households (on total number of hh of the group)	incidence of non registered self employment income on total self employment income of the group	incidence of non registeredwages and salaries on total wages and salaries of the group
Wages and salaries	8.6%	59.8%	8.6%	59.1%	5.5%	12.1%	3.9%	11%	12.7%	5.2%	5.1%
Self employment income	12.3%	54.6%	8.2%	46.3%	12.2%	9.9%	4.8%	11%	12.3%	6.7%	3.8%
Property income	15.1%	35.9%	12.4%	32.0%	7.1%	9.3%	9.3%	13%	13.1%	5.4%	4.0%
Social benefits and other incor	9.1%	56.0%	9.4%	47.6%	3.5%	3.9%	4.3%	5%	3.3%	5.1%	4.5%
total	12.1%	46.8%	9.9%	42.6%	5.2%	6.0%	6.2%	10%	9.5%	5.7%	4.2%

	self employment income	non registered self employment income	wages and salaries	non registered wages and salaries	property income	social benefits and other transfers	tax and contribution	total net income	n. of households
Wages and salaries	14%	18%	21%	30%	-5%	7%	11%	17%	26%
Self employment income	28%	33%	26%	28%	14%	8%	19%	23%	26%
Property income	47%	36%	46%	35%	49%	43%	55%	44%	36%
Social benefits and other incor	11%	14%	7%	8%	42%	41%	15%	16%	12%
total	100%	100%	100%	100%	100%	100%	100%	100%	100%

Tab 8.3 Households without non-registered employed

	self employment income	non registered self employment income	wages and salaries	non registered wages and salaries	property income	social benefits and other transfers	tax and contribution	total net income	n. of households
Wages and salaries	21%		25%		-5%	3%	18%	16%	19%
Self employment income	28%		32%		6%	5%	24%	21%	20%
Property income	36%		36%		35%	27%	36%	32%	25%
Social benefits and other incor	16%		7%		64%	65%	22%	31%	37%
total	100%	0%	100%	0%	100%	100%	100%	100%	100%

Tab 8.4 all households by householder age group

	self employment income	non registered self employment income	wages and salaries	non registered wages and salaries	property income	social benefits and other transfers	tax and contribution	total net income	n. of households
Wages and salaries	20%	18%	25%	30%	-5%	4%	18%	16%	19%
Self employment income	28%	33%	31%	28%	6%	5%	24%	21%	20%
Property income	37%	36%	37%	35%	36%	28%	37%	33%	26%
Social benefits and other incor	15%	14%	7%	8%	63%	64%	22%	29%	34%
total	100%	100%	100%	100%	100%	100%	100%	100%	100%

6. Final considerations

This paper presents an alternative method for estimating income from non-registered labour, both from employees and self-employed. The idea is that of detecting non-registered workers by crossing information from survey and administrative data sources on household's budgets. Particularly, we use the Italian survey on income and living conditions (It-Silc), which provides lots of information on workers and their families and several administrative archives, mainly related to social security obligation. We apply an exact record linkage procedure, which compares records contained in the It-Silc and administrative archive datasets, in order to determine pairs of records pertaining the same worker. IT-Silc workers who do not find a match are labeled as non-registered. Applying a logistic regression analysis, we find that non-registered workers are, other things being fixed, more likely man than female, old than young, foreign than Italian. They work more frequently in small farms and in the South. Furthermore, their disposable income is lower as well as the disposable income of their family. The results of the analysis are in line with Istat statistics on non-registered labour and this supports the validity of our method.

The analysis shows that administrative data provide key information, moreover it confirms that household surveys capture at least a part of underground components (it is likely that some non-registered workers do not declare to be employed and therefore to earn some labour income). Therefore, their combined use not only allows to identify non-registered workers, but also to quantify their income and to measure its distribution among different households groups.

Our analysis confirms that non-observed income flows are not evenly distributed, but are more likely in some sub-population.

The National Accounts department is currently carrying out a revision for the adoption of ESA 2010 and in this occasion, it is reviewing all data sources and estimation methodologies. The work presented in this paper strongly benefited from the on-going revision process. In the next future, administrative sources will be even more extensively combined with micro data from households surveys. This will further improve the coherence and accuracy of the overall accounting system and will enhance the households' perspective in national accounts .

References

Brandolini, A.: The Distribution of Personal Income, in Post-War Italy: Source Description, Date Quality, and the Time Pattern of Income Inequality, Temi e discussioni, Banca Italia,1999.

Calzaroni, M., "The exhaustiveness of production estimates: new concepts and methodologies", International Conference on Establishment Surveys, Buffalo, USA, 17-21 June, 2001

Coli A., Tartamella F.: Micro-Macro Integration: Survey Data on Household Income for the Estimate of the Italian GDP, Iariw Conference, St. Gallen, Switzerland, August 2010.

Consolini P. (2009), "Integrazione di dati campionari Eu-Silc con dati di fonte amministrativa", Collana Metodi e Norme Istat, 38, Roma.

Copas J. B., F. J. Hilton (1990) Record Linkage: Statistical Models for Matching Computer Records, *Journal of the Royal Statistical Society. Series A* (Statistics in Society), Vol. 153, No. 3 (1990), pp. 287-320.

Eurostat (1995) European system of Accounts

Eurostat (2010). Cross Sectional UDB (User Data Base)

Faramondi, F. Foschi, A. Puggioni, 2006, "Le innovazioni introdotte nel trattamento dei dati di impresa per le stime di contabilità nazionale" in La revisione generale dei conti nazionali del 2005 Roma, 21-22 giugno 2006

Istat (2000) Le nuove stime dei consumi finali delle famiglie secondo il SEC 95, Metodi e norme n. 7, 2000

Franz A. (1985), Basic Model in Estimates of the hidden economy in Austria on the basis official statistics - The Review of Income and Wealth, 4, 1985

Istat (2010) *La misura dell'economia sommersa secondo le statistiche ufficiali, anni 2000 - 2008*. Statistiche in breve.

Istat (2011) La misura dell'occupazione non regolare nelle stime di contabilità nazionale, http://www.istat.it/it/archivio/39522

OECD (2002) Handbook for Measuring the non-observed economy

United Nations (2008) Non observed economy in National Accounts – Survey of countries practices, New York and Geneva.