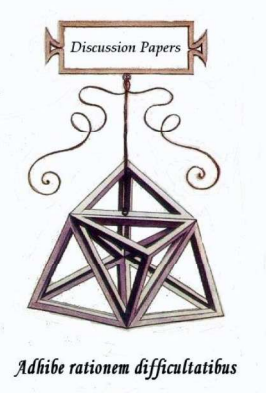
Discussion Papers
Collana di
E-papers del Dipartimento di Economia e Management – Università di Pisa



Alessandra Coli, Francesca Tartamella

Features of the hidden labour in Italy. An empirical analysis based on the matching of survey and administrative data

Discussion Paper n. 259

2020

Discussion Paper n. 259, presentato: **Ottobre 2020**

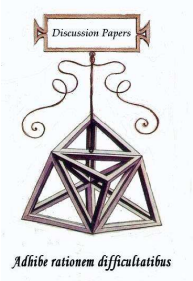
Indirizzo dell'Autore:

Alessandra Coli
Dipartimento di Economia e Management, via Ridolfi 10, 56100 PISA, Italia
Email: alessandra.coli1@unipi.it

Francesca Tartamella
Istituto Nazionale di Statistica, Via Cesare Balbo, 16, 00184 Roma, Italia
Email: tartamel@istat.it

Si prega di citare così:
Coli A., F. Tartamella (2020), "Features of the hidden labour in Italy. An empirical analysis based on the matching of survey and administrative data", Discussion Papers del Dipartimento di Economia e Management – Università di Pisa, n. 259 (http://www.ec.unipi.it/ricerca/discussion-papers).

Discussion Paper
n. 259



Alessandra Coli, Francesca Tartamella

Features of the hidden labour in Italy. An empirical analysis based on the matching of survey and administrative data

Abstract

In this paper, record linkage is applied to detect hidden (or irregular) workers in the Italian labour market. The idea is to trace each single worker in a survey as well as in a set of administrative registers, where the worker should appear if regular. When a worker is sampled by the survey but he does not appear in any of the administrative registers, that worker is identified as irregular. Subsequently, we estimate a logistic regression model to identify the individual and household features, which characterize the typical hidden Italian worker.

Keywords: record linkage; non registered labour; hidden economy

JEL: J21, C81

Features of the hidden labour in Italy. An empirical analysis based on the matching of survey and administrative data

Alessandra Coli*, Francesca Tartamella†

1 Introduction

Official statistics estimate income from undecleared work in the context of National Accounts (NAs henceforth). In fact, to meet the requirement of "exaustiveness", Gross Domestic Product (GDP) must include the value added generated by all production activities including those that use hidden work (OECD 2002). Italian NAs publish annual estimates of the number of irregular workers and shows their distribution by employment status (employees or self employed) and by kind of economic activity. In addition, on a more occasional basis, Istat provides estimates of the terrirorial distribution of irregular workers (see e.g. Istat 2020). These estimates are obtained at a macro level and are mainly produced to capture the characteristics of those production units that use undeclared work. Conversly, NAs do not allow one to analyse the characteristics of individuals/households who

*Corresponding author: Department of Economics and Management – University of Pisa, Italy (alessandra.coli1@unipi.it)

†Italian National Institute of Statistics, Italy

benefit from hidden income. To trace the profile of the typical irregular worker is thus necessary to abandon the macroeconomic scheme and resort to the use of micro data having the "worker" as the unit of analysis. This paper presents the results of an empirical analysis based on the use of the Italian Statistics on Income and Living Conditions (IT-SILC henceforth) and a set of administrative archives stemming from the fiscal and/or social security's obligations of registered employees. First, irregular workers are detected among the ones sampled by IT-SILC, then a logistic regression model is estimated to identify the peculiar characteristics of irregular workers. The paper is structured as follows. In Section 2, we describe the micro data sources employed in the empirical analysis. Section 3 describes the method used to detect hidden workers whereas Section 4 presents the logistic regression results. Section 5 includes concluding remarks.

2 Micro data sources on workers

The empirical application described in this study requires the use of survey as well as administrative datasets, both having the worker as unit of analysis. Among available surveys we decided to use IT-Silc because it collects very detailed information not only on the employment conditions of the interviewed but also on the economic situation of the family they belong to. For example, we could include the household income quintile among the independent variables of the logistic regression model (see Section 4). IT-Silc is carried out yearly, on the basis of a face-to-face interview. Collected data are currently integrated with information taken from revenue taxes in order to reduce measurement errors (Consolini, 2008). In this study we used the 2011 IT-Silc round (2010 reference year) since for the year 2010 many administrative registers were available. In reality, the year is not so important for the objective of our analysis, since it is reasonable to assume that

the characteristics of the hidden work are stable over time or that they change very slowly in time.

For employees, the following administrative archives were considered.

- **INPS** The social security archive of individual insured position for workers employed in the private sector (with the exception of agricultural activities).
- **INPDAP** The social security archive of individual insured position for workers employed in the public sector.
- **ENPALS** Social security archive of individual insured position for workers employed in the private sector of sport, arts and entertainment.
- **INPS (domestic)** Social security archive of individual insured position for workers employed as domestic staff.
- **INPS (agriculture)** Social security archive of individual insured position for workers employed in agricultural sector.
- **INAIL** Social insurance archive for all employees.
- **INAIL(agency workers)** Social insurance archive for agency workers.

For self-employed the following registers were considered.

- **INPS (outworkers)** The social security archive for outworkers.
- **INAIL (outworkers)** The social insurance archive for outworkers.
- **INPS (self-employed in agriculture)** The social security archive for self-employed working in agriculture.
- **INPS (professionals)** The Social security archive for professionals and freelancers.

Administrative Archive	Percentage of links
INPS	24.0
INPDAP	6.5
ENPALS	0.4
INPS (domestic)	0.3
INPS (agriculture)	1.3
INAIL	27.5
INAIL (agency workers)	0.7
INPS (outworkers)	2.4
INAIL (outworkers)	1.9
INPS (self-employed in agriculture)	0.8
INPS (professionals)	0.4
SE	14.8
ASIA	10.0
ALL (at least 1 link)	44.5

Table 1: Percentage of links with respect to the total number of interviewed, by administrative archive

- **SE** Istat archive on self-employed. This archive stems from the combination of fiscal agencies data (in particular those concerning VAT numbers) and chambers of commerce data.

- **ASIA** Istat statistical business register on active enterprises. The register integrates administrative and statistical data sources on firms belonging to the industrial sector.

3 The detection of the hidden workers

The method applied in this study derives from the observation that surveys are able to capture hidden work since at least some of the irregular workers identify themselves as employed or self employed (see e.g. Kazeimer 2014). Consequently, only part of interviewed workers (namely the regular ones) are expected to appear in the administrative archives. Through the record linkage, we aim to find out which of the workers interviewed are not present in the administrative archives, thus providing an estimate of irregular workers.

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 each linked unit. In order to apply this technique, 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. The procedure is straightforawrd 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. This procedure is known as deterministic record linkage. Obviously, errors may occur because the identifier

variable is incorrectly recorded (Copas and Hilton 1990). Due to such errors, two records for the same unit may not agree (false unmatched pair), and two records, which agree may refer to different units (false matched pair) . Formalizing the linking procedure into a statistical model, it is possible to evaluate the quality of the matching by measuring the probability of generating false-matched-pairs and false-unmatched pairs (Fellegi and Sunter 1969). This proedure is known as probabilistic record linkage.

In this work, we consider IT-Silc dataset as the file A whereas each administative archive represents the B file. The objective is that of tracing each interviewd worker througth all the administrative archives described in the previous section. The identifier is an anonymized code, which corresponds to the Italian fiscal code (Codice Fiscale). Fiscal code (and consequently our anonymized code) is a good identifier since it identifies univocally each worker. Furthermore, possible recording errors can be identified by checking the values of the variables from which the fiscal code itself is derived[1]. Since registration errors are assumed to be negligible, deterministic linking appears to be the best method.

The first step of the procedure consists in the identification and selection of the " IT-Silc workers ". For this application we have decided to define them as those individuals who declared that they received an income from work in 2010.

For each IT-Silc worker, the record linkage procedure looks for a link in each administrative register. Table 1 shows the number of links found out in each register (see Section2) as a percentage of IT-Silc interviewed. The result is that 44.5% of all surveyed individuals are found in at least one archive. We observe that INAIL contains the highest number of links, followed by INPS. It is worth noting that some archives are expected to overlap. For example a worker recorded

[1]The Italian "codice fiscale" is an alphanumeric code of 16 characters, with characters reflecting personal information like name, surname, date and place of birth.

in INPS is expected to be found also in INAIL, since regular workers have to pay both social security and social insurance contributions.

Eventually, we lable as "regular" the IT-Silc workers for whom record linkage found out at least one link in the administrative archives and as "irregular" those who could not be matched to any administrative record. According to our results, irregular workers represent about 9.5% of total IT-Silc workers, with higher presence among employees (about 60% of irregular workers are employees). These values are in line with National accounts estimtes (Istat,2011), according to which hidden labour represents 10.3% in terms of employed persons and 17% in terms of jobs.

During the matching procedure links were found also for some individuals that were identified as unemployed according to the survey. Such links are likely due to incorrect responses to the survey, since an in-depth analysis of the administrative archives has shown that these cases correspond mostly to short-term (one month or less) jobs with low earnings.

The method described above was firstly proposed in (Coli, Tartamella 2014). A similar application in shown in De Gregorio (2016) where the Labour Force Survey is used instead of IT-Silc.

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

In this section, we present the result of a statistical analysis aimed at detecting the specific features of irregular workers. The response variable (*Irreg*) indicates if the individual is a hidden worker or not.

The model contains several explanatory variables. Two variables measure the position of the worker in the ranking of personal and households disposable in-

Name	Description	Reference category
irreg	Irregular(1)/regular(0)	
quanty	Quintile of personal disposable income	1st quintile
quantfy	Quintile of household disposbale income	1st quintile
region	Region (Nust2) where the interviewed persons (families) lives	Piemonte
gender	Gender of the worker	Female
ywork	Number of years spent in paid work	
age	Age (in years) in 2010	
citizenship	If Italian or foreigner	Foreigner
education	Level of educattion attained (3 levels)	Lowest level
timework	Full time or part time	Part time
jobs	One job or more than one job	More than one job
industry	Economic activity of the local unit of the main job	Agriculture
employment	Employed, Self-employed, Both	Employed
firmsize	Size of local unit in terms of employees: 1-10, 11-19, 20-49, 50 and more	1-10 employees

Table 2: Description of the variables used in the logistic regression

comes. Five variables concern personal characteristics of the worker like gender or age. Six variables relates to labour characteristics like the number of years worked or the industry of the firm where the worker is employed. Table 2 describes the variables in detail whereas Table 3 and 4 present the results of the analysis.

Table 3 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 explanatory variables is significant. We can see that all terms were highly significant (P-value$<$ 0.001) when they were introduced into the model, with only three exceptions: level of education attained, whether the worker has a full-time or a part-time job and whether he has got one or more than one job.

Table 4 shows the estimates of coefficient and the results of the test.

For easy of interpretation, we limit to discuss the significance and sign of estimated coefficients. A significant positive coefficient indicates that the predictor increases the probability of being classified as irregular, whereas a significant negative sign means that the predictor reduces the risk of being a hidden worker. In case of categorical variables, the coefficient sign indicates whether the observed category increases (positive sign) or decreases (negative sign) the risk, compared to the reference category (see Table 2).

Our results show that the probability of being an irregular worker decreases with the number of years worked but increases with the age of the worker. We also notice that male workers have a higher probability of being irregular with respect to female workers. The risk of being irregular decreases with the size of the firms; in fact workers employed in local units with 9 employees at most, record the highest probability. Foreign workers have a higher probability to be irregular compared to Italian workers and the same is true for self-employed workers compared to employees. The probability of being irregular decreases with the amount

	Df	Deviance	Resid.Df	Resid.Dev	Pr(>Chi)	sig
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	***
ywork	1	11.81	15313	8548	0.0005894	***
age	1	24.6	15312	8523.4	7.06E-07	***
citizenship	1	91.88	15311	8431.5	<2.20E-16	***
firmsize	3	166.77	15308	8264.7	<2.20E-16	***
employment	2	299.29	15306	7965.4	<2.20E-16	***
industry	7	302.26	15299	7663.2	<2.20E-16	***
education	2	2.25	15297	7660.9	0.3243601	
timework	1	0.26	15296	7660.7	0.6133828	
jobs	1	0.12	15295	7660.6	0.733253	

Table 3: Change in deviance obtained by adding each explanatory variables

of personal and household disposable income. The probability of being irregular is higher for workers of the South compared to those working in the North. In fact, regions with p-value lower than 0.05 (one star) and positive coefficients are in the South, with the only exceptions of Toscana and Lazio. Only the North and namely province of Trento, Veneto and Emilia Romagna perform better than Piemonte (reference category). Finally, workers employed by Households (mainly domestic staff) record the highest risk to be irregular followed by those employed in Agriculture and in Services.

	Coeff.	Std. Error	z value	Pr(>|z|)	Sig.
(Intercept)	0.286	0.315	0.909	0.364	**
2nd quanty	-0.992	0.090	-10.991	< 2e-16	***
3rd quanty	-1.456	0.107	-13.542	< 2e-16	***
4th quanty	-1.380	0.109	-12.624	< 2e-16	***
5th quanty	-2.040	0.127	-15.985	< 2e-16	***
2nd quantfy	-0.168	0.086	-1.948	0.051	.
3rd quantfy	-0.433	0.098	-4.414	1.01e-05	***
4th quantfy	-0.260	0.103	-2.528	0.011	*
5th quantfy	-0.081	0.108	-0.753	0.452	
Male	0.259	0.072	3.614	0.000	***
Val d'Aosta	-0.185	0.292	-0.633	0.527	
Lombardia	0.015	0.162	0.096	0.923	
Bolzano	0.351	0.243	1.446	0.148	
Trento	-0.778	0.356	-2.182	0.0290	*
Veneto	-0.516	0.187	-2.753	0.005	**
Friuli Venezia Giulia	-0.024	0.193	-0.127	0.898	
Liguria	0.126	0.194	0.651	0.515	
Emilia Romagna	-0.482	0.192	-2.511	0.0120	*

	Coeff.	Std. Error	z value	Pr(>|z|)	Sig.
Toscana	0.372	0.169	2.203	0.0276	*
Umbria	-0.038	0.198	-0.195	0.845	
Marche	-0.528	0.215	-2.452	0.014	*
Lazio	0.873	0.154	5.667	1.46e-08	***
Abruzzo	-0.0462	0.263	-0.175	0.860	
Molise	0.497	0.230	2.160	0.031	*
Campania	1.097	0.161	6.819	9.19e-12	***
Puglia	0.445	0.183	2.426	0.0152	*
Basilicata	0.648	0.228	2.834	0.0046	**
Calabria	0.337	0.200	1.684	0.092	.
Sicilia	0.338	0.185	1.830	0.067	.
Sardegna	0.457	0.208	2.199	0.027	*
Ywork	-0.025	0.005	-4.839	1.31e-06	***
Age	0.011	0.005	2.189	0.028	*
Italian	-0.998	0.108	-9.203	$< 2e-16$	***
firme size $(11-19$ employees)	-0.219	0.096	-2.274	0.023	*
firm size $(20-49$ employees)	-0.739	0.132	-5.587	2.31e-08	***
firm size (50 and more employees)	-0.358	0.097	-3.691	0.000	***
Self-employed	1.334	0.082	16.353	$< 2e-16$	***
Self-employed and employed work	-0.644	0.187	-3.447	0.000	***
Manufacturing	-1.176	0.142	-8.255	$< 2e-16$	***
Construction	-0.995	0.157	-6.328	2.49e-10	***
Financial and insurance activities	-1.658	0.290	-5.706	1.15e-08	***
General gov., education, health	-0.555	0.143	-3.869	0.000	***
Households activities	1.604	0.269	5.966	2.43e-09	***
Services other than retail	-0.913	0.137	-6.642	3.10e-11	***

	Coeff.	Std. Error	z value	Pr(>|z|)	Sig.
Retail	-1.224	0.131	-9.359	< 2e-16	***
Education (medium level)	0.020	0.071	0.280	0.779	
Education (highest level)	-0.153	0.105	-1.457	0.145	
Full time	0.044	0.083	0.532	0.594	
One job	- 0.048	0.202	-0.238	0.812	

Table 4: Logistic regression – Tests on the coefficients.

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

5 Concluding remarks

This paper shows a method to obtain an estimate of hidden labour. Furthermore it compares regular and irregular workers and tries to detect the peculiar features of Italian hidden labour. The results of the analysis are in line with those stemming from National accounts. The typical irregular worker earns a low income, lives in the South and works in local unit of small size. He is more frequently male and foreigner. However, further research is required to validate the method. In particular, the application should be repeated for several years, using both IT-Silc and LFS from the survey side. Furthermore administrative archives should be continuously improved in order to favour record linkage with surveys.

Acknowledgments

This work was supported by: University of Pisa PRA2018-19 project.

The views expressed in this paper are those of the authors and do not necessarily reflect the views of their affiliation Institutes.

References

[1] Coli A., F. Tartamella (2014), Using administrative and survey data to analyse tax evasion from unregistered labour, 33nd IARIW General Conference, Rotterdam, The Netherlands. August 24-29, 2014. http://www.iariw.org/papers/2014/ColiPaper.pdf.

[1] 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.

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

[4] De Gregorio C., A. Giordano (2016) The heterogeneity of undecleared work in Italy: some results from the statistical integration of survey and administrative sources. Rivista di statistica ufficiale. N. 2/2016.

[5] Fellegi I. P., A. B. Sunter. (1969). A theory for record linkage, Journal of the American statistical association. vol. 64, pp. 1183-1210.

[6] Istat (2020) Conti economici territoriali, anni 2016-2018, Statistiche Report, pubblicato il 20/01/2020, https://www.istat.it/it/files//2020/01/Conti-economici-territoriali.pdf.

[7] Kazemier B. (2014) Hidden workers and the hidden worker potential in the Netherlands. Economic analysis and Policy 44 (2014), 39-50.

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

Discussion Papers
Collana del Dipartimento di Economia e Management, Università di Pisa

Email della redazione: lfanti@ec.unipi.it