Can NA Data Account for Differences among Household Groups? Integration of Micro and Macro Data on Labour Income for Households Accounts

Davide Di Laurea
Istat - Italian National Institute of Statistics

Francesca Tartamella
Istat & Eurostat

Paper prepared for the 35th IARIW General Conference

Copenhagen, Denmark, August 20-25, 2018

Session 4D: Distributional Diversity in the National Accounts

Time: Wednesday, August 22, 2018 [14:00-17:00]

Can NA Data Account for Differences among Household Groups? Integration of Micro and Macro Data on Labour Income for Households Accounts

Davide Di Laurea[1] *and Francesca Tartamella*[1,2]
[1]*Istat - Italian National Institute of Statistics*
[2]*Eurostat*

Preliminary draft

Abstract

National Accounts describe the economic behaviour of a representative household without investigating heterogeneity, while micro data sources provide information on the distribution of income, consumption and wealth among people. However, the latter might fail in covering comprehensively all households' earnings and expenses and do not allow for analysing households vis a vis the other economic sectors. Over the years the macro and micro approaches developed separately, often leading to divergent results even when coping with ex-post fully harmonized population domains and income definitions. Moreover, if micro data are used to derive distributional information for sub-populations, the resulting estimates might be misleading, due to heterogeneous magnitudes of discrepancies across different strata of the relevant population. It is therefore important to investigate and detect all sources of differences and to consequently adjust micro and macro data, in order to derive correct distributional information.

The aim of the paper is to show how administrative archives can be used to integrate micro and macro data on labour input and labour compensation, finding the reasons of discrepancies on these flows. The paper describes the process of record linkage between an household income survey (It-Silc) and the Italian Social Security archives with data on employee and self-employed. The latter, combined with Labour Force Survey data, already constitutes the base of Italian National Accounts estimates on labour input and labour compensation. The massive use of these integrated survey and administrative data let macro data have a sound and coherent micro database; the further integration of It-Silc data provides a valuable framework for distributional purposes.

The comparison among administrative and survey data at the micro level allows for detecting and appropriately correcting inconsistencies. They may come from diverging occupational status – occupied vs non-occupied - or employment status –employee vs self-employed- at the individual level as well as at the job one. It also enables to identify non-registered workers, permitting to impute their own relevant labour income. Once working characteristics are reconciled, it also provides a valid support for minimizing discrepancies between amounts in the data from different sources by imputing non-reported and/or under-reported income components.

1 Introduction

Households' economic condition is a key indicator of economic well-being: enterprises and Governments performances are socially worthy only when they lead to a widespread improvement of economic conditions for the involved population.

Households disposable income and its trend over time is a crucial measure among many well-being indicators. Over the years, they have been developed two different and distinct pathways to estimate disposable income, based respectively on micro and macro approaches. Micro approach commonly uses household survey data; alone or, more recently, integrated with administrative archives. It allows for analysing the association among individuals and households characteristics and the distribution of income and its components. Microdata make it possible to run distributional analyses and to estimate economic inequality or poverty measures by population subgroups. However, survey data on income are typically far from being complete and exhaustive, being affected by sampling and non-sampling errors: selective total non-response, non-reporting and under-reporting responses biases; they could not be able to cover accurately and reliably some kind of incomes, financial income or gains from illegal activities among others. All these factors do not homogenously affect income distribution curves nor their impact is likely to be constant among population subgroups; in turn, they may result in less robust, if not poor, distributional analyses.

On the macro side, National Accounts (NA henceforth) data are exhaustive for all income components by definition, making its totals and trends more reliable. Moreover, households disposable income is consistently estimated with respect to each other NA flows. Thus, NA macrodata allow for combining analyses with other relevant NA measures, such as GDP, enterprises performance, factor productivity, general government spending or deficit. As a negative aspect, it does not offer any insight on individuals or households attributes and their heterogeneity. It should also be considered that some NA concepts and imputed flows make sense only within the general framework of NA as a whole, but they can be misleading when considering the household sector *per se* and hardly to be interpreted from any other perspective; the correction of interests for Fisims is a valid example of this kind of flows.

Having different objectives, the result is that the micro and macro approaches have been – and still are- developed separately, quite often producing different outputs and the aggravating circumstance to prompt reciprocally inconsistent interpretations on the undergoing processes. As pointed out in the 2nd edition of the Canberra Group Handbook on Household Income Statistics (Unece, 2011; pag. 5): "*It is undoubtedly a considerable disservice to users when two sets of statistics both labelled 'household income' appear to produce different results, and possibly have different implications for social and economic policy*".

Many references complaining on the relevance of these statistical "disservices" might be quoted. Nowadays, there is a large acknowledgement about the need to overcome this separation and its limitations: integrating micro and macro approaches could provide consistent and more precise distributional measures, combining virtues of both and containing vices of each one. More accurate indicators on households economic behaviour could improve the quality of official statistics on both sides and would answer the Stiglitz-Sen-Fitoussi report recommendations on NA statistics: i) *giving prominence to the distribution of income, consumption and wealth* - Recommendation 4, p. 409; ii) *making distributional measures compatible in scope with average measures from the national accounts* - SSF report § 43, p. 34; iii) *developing distributional measures of full income (i.e. distributing market income but also imputed income such as imputed*

rents from own-occupied housing, and government services provided in kind) - SSF report § 57, p. 39.

Eurostat, OECD and other statistical agencies have been taking significant steps in order to address these recommendations. Among others, the OECD-Eurostat Expert Group on Disparities in a NA framework is devoted to deepen research on the reconciliation of macro and micro estimates with the ultimate purpose of providing NA estimates broken down by households subsectors[1]. In early 2016, a new initiative by the European Central Bank (ECB) has established an Expert Group on Linking Micro and Macro Household Data (EG-LMM).

Oecd is drafting a manual on the methodology to be applied in order to derive distributional information from macro totals. This methodology is taken as guideline by countries approaching this reconciliation, but it is true that the approach may vary according to country specific situation in terms of variety and reliability of sources (from surveys only to registers for all kinds of income).

The easy way of thinking about the integration of micro and macro data consists in using breakdown indicators derived from an household survey to decompose NA flows. In other words, integration could be performed by re-proportioning NA flows in accordance to distributions by individual or household characteristics, the latter coming from survey data. In this case, differences are not accounted for, as if they were uniformly distributed among all the statistical units, thus running the concrete affecting correct distributional analysis. This concrete risk pushes forward the need for an accurate exploration for differences before using micro data to derive distributional information on macro data. Deeply understanding the reasons and why (and the domains where) the two sources differ can positively influence the reconciliation process.

The aim of this paper is to report the main results of an experimental exercise based on data for Italy for 2011. We first examine the gap between NA and the Italian Survey on Income and Living Conditions (It-Silc henceforth) estimates for labour compensation, once target populations and income concepts were harmonized. We then proceed with a micro-micro data linkage between survey and administrative data; the latter are not used as auxiliary variables, being instead fully integrated at a micro-level with survey data in order to reconcile them.
The result from the reconciliation process is then compared with NA estimates, to measure if and how much the initial gap is reduced. An evaluation of the distributional impact of the exercise is also provided. The paper suggests a way of adjusting micro data, integrating different sources at a micro level, so that survey incomes can be more consistent with macro estimates. It opens the way to a more effective use of micro data to disaggregate macro flows.

The paper is structured as follows: section 2 points out conceptual and empirical differences between micro and macro data on household income items. Section 3 shows how administrative registers can be used to integrate income data on labour input and labour compensation by addressing micro-macro cross-consistency, identifying the reasons of discrepancies and hence allowing for detecting and accordingly correcting inconsistencies. Section 4 shows the impact of reconciliation and integration of household survey and administrative data on household income distribution. Finally, section 5 draws some conclusive comments. The analysis refers to 2011, the reference year for NA benchmark, for which there is a rich set of information already integrated.

[1] M. Fessau and L. Mattonetti (2013b) presented the main conclusions drawn by the activity of this Expert Group in its first phase, whereas J. Zwijnenburg (2016) dealt with gaps between micro and macro aggregates that need to be bridged in the compilation process.

2 Reconciliation of macro and micro data on households' income

When one compares the micro and the macro values for the main income aggregates, as they are published, the differences may be sometimes misleading. In the following, we use It-Silc as the reference source for households survey income in contrast to NA figures. As it is evident from Table 1, some aggregates l - *Wage and salaries* or *Operating surplus from own account production* - have a good, if not excellent, coverage rate; others - *Property income* or *Income from self-employment* - lay far below the corresponding NA figures.

Table 1. Coverage rates between It-Silc and NA for the main income components before harmonization, 2011

Disposable income and its components	It-Silc value/ NA value, %
Wages and salaries	100.2
Self-employment income	61.3
Property income (received less paid)	15.0
Operating surplus from own account production	104.8
Actual rents (including land)	75.9
Current taxes and actual social contribution paid by households	89.4
Social benefits excluding social transfers in kind (received less paid) + other current transfers	89.4
Total disposable income	82.7

However, it has to be ensured to deal with "homogeneous" figures, not affected by specification errors: they should refer to the same statistical domains; furthermore income concepts and definitions should be as close as possible. Thus, the first step consists necessarily in reconciling the two sources, getting rid of these differences in order to sterilize the coverage rates by their effects.

The first difference between micro and macro data is the reference population. NA data refer to all units whose "economic residence"[2] is on national territory: institutional households or illegal immigrants are in the domain of study for NA; they are instead out of scope for household surveys, hence not included in It-Silc[3]. In these cases, any income flows directed to these part of NA population have to be eliminated from NA aggregates. For institutional households, income estimates are based on the distribution by age and work characteristics, drawn from Population Census. As for illegal immigrants, their income is directly estimated within NA operations, in terms of hours worked and sector of economic activities. Both corrections amount to 1.6% of NA

[2] An institutional unit is resident in a country when it has its center of predominant economic interest in the economic territory of that country (ESA10 par. 2.04).

[3] According to ESA framework, Households sector includes also Npishs. In Italy the full Households account is computed separately for Npishs, producer and consumer households (see M. Ascione A. M. M. Carucci, F. C., L. Ciaccia, P. Santoro – 2012). Indeed, in the Italian framework disentangling flows not relevant for households does not imply any additional operations.

disposable income; with respect to income components, the incidence is higher only for *Wage and salaries*, reaching 2.9%.

Micro and macro sources differ also in terms of income concepts and definitions. As a general point, it should be kept in mind that almost all flows can be affected by re-computing or re-classification due to this kind of differences. Far from being complete, in Appendix A we have compiled a list for the items whose harmonization may have a not negligible impact. In the following, the main treatments applied to data for the current exercise:

- Wages and salaries. NA flows include the imputed employers' social insurance contributions for persons for which no real contribution is effectively paid; the opposite holds true in It-Silc. On the other side, arrears for dependent workers have to excluded from It-Silc employee income.
- Self-employment income. From NA perspective the following three distinct flows have to be taken into account: i) *Share of mixed income distributed to consumer households*; ii) *Withdrawals form quasi-corporations*; iii) *Other income distributed from corporations*. In *Table 2* the flows for self-employment income from NA are reported in two versions, depending on the inclusion of income from illegal activities or not. However, it remains incorporated in total NA disposable income.
- Gross operating surplus (GOS). Following EU regulation, in It-Silc imputed rents are considered only for households' main dwellings. Therefore, GOS has been computed both including or excluding *imputed rents on secondary dwellings* from corresponding NA aggregate. Moreover, income flow *from own account production of gross fixed capital formation* is deducted from GOS estimates. However, both flows are not excluded from the total NA household disposable.
- Property income: NA value does not include other non-cash investment income (see Appendix A sub f); interests are not corrected for Fisim.

Table 2. Coverage rates between It-Silc and NA for the main income components after harmonization, 2011

Disposable income and its components	It-Silc harmonized value/NAs value, %
Wages and salaries – gross of arrears	102.2
Wages and salaries – net of arrears	101.6
Self-employment income	61.9
Self-employment income – net of income from illegal activities	64.9
Property income (received less paid)	29.1
Operating surplus from own account production	104.8
Operating surplus from own account production excluding secondary dwellings	128.1
Actual rents (including land)	76.3
Current taxes and actual social contribution paid by households	95.0
Social benefits excluding social transfers in kind (received less paid) + other current transfers	89.4
Disposable income	91.3

* It excludes income from illegal activities and imputed rents on secondary dwellings at households disposal.

As it is evident, beneath the harmonization of population and conceptual contents of the aggregates, the survey values still fall short the macro ones, especially for self-employment and property income. Moreover the over-coverage of GOS becomes manifest.

Differences may be due to non-reporting and under-reporting in surveys but also to selective non-response and scarce sample representativeness, particularly affecting highly concentrated items (as an example: income from financial assets) or related to infrequent events (as for severance payments).

It is therefore necessary to further investigate the reasons of discrepancies and possibly intervene on them at a micro level. In the next section we show how a record linkage between survey data and administrative archives could be a step forward in this direction.

3 Integrating survey and administrative income data

The administrative data have playing a more and more crucial role in the statistical production, being no longer used only as auxiliary sources. As for Italy, they have been used for twenty years in Business surveys processes, while only in more recent years and unevenly in the domain of Social Statistics.

A significant innovation for labour input estimates was experimented and introduced with the occasion of last NA benchmark operations. In this context, an integration among data from different statistical domains was already implemented although at a meso level. The change consisted in pushing the integration at a micro level, jointly using Social Security administrative archives and Labour Force Survey data[4]. This new operational framework increased the informational power of the linked micro-data. It allowed for checking, detecting, identifying and correcting information on the same units whose reciprocal inconsistency would have been ignored following the traditional approach. As it is known, in a stovepipe production process ensuring internal coherence for the data is sufficient to come to an end. By using multiple sources for each population unit increases the complexity of the reconciliation phase, due to the higher number of inconsistent profiles than may occur and are to be correctly treated. On the other hand, having available labour demand and supply side data gives the opportunity detect signals of undeclared jobs, not covered by administrative registers. At the same time, administrative data might be affected by over-coverage, less difficult to check for in presence of multiple information. On the whole, the cost of a more complex data processing phase is compensated by the benefit of a more accurate measure of the phenomenon and the positive spill-over on the survey process.

Moving from this experience, in this section we show that administrative data can be a very useful complement of household survey data also for income analysis. The traditional integration of sources that is the base of NA estimation, has now strengthened the use of micro data: one of the most relevant result of the integration consists in detecting hidden flows so that many corrections and imputations on economic variables (not only labour input) are now feasible at the micro level. This means that most part of labour compensation can be traced, linking employers to each person employed in the productive system.

The focus of the present exercise consists in integrating household survey estimates on labour income with administrative data. The combined process of integration and reconciliation should entail an higher micro data consistency leading. The main potential advantage by working at a

[4] F. Battellini et al. (2015).

micro-level lies in preserving information on multiple jobs, although it is necessary to figure out operational rules able to distinguish them from information on the same job which are misclassified. Each source, in fact, classifies jobs and/or heads according to its own purposes, involving the risk of income being considered as pertaining to different jobs even when they are not. It allows for adding missing information on jobs and income to household survey data[5], while limiting the risk of double-counting for jobs and/or income, when already present (partially or totally) in the survey even if under different classification. As a by-product undeclared jobs and related income may be estimated, ensuing a finer comparison with NA aggregates. As a matter of fact, it is not possible to presume ex-ante that the output gets It-Silc closer to NA; in any case the resulting increased homogeneity will enable to discern discrepancies, identifying weak points of the survey or on the NA side.

3.1 Survey and administrative data used

It-Silc is the main source for household income and living conditions, harmonized at European level. For the current work we have used data from It-Silc 2012, whose income reference year is 2011.

The main variables for labour income at individual level are respectively: the sum of PY010G/N and PY020G/N for cash, near-cash and non-cash employee income; PY050G/N for cash benefits or losses from self-employment. It-Silc, from its very beginning, integrates survey data with tax registers reports[6]. The latter ones, matched at the micro level, serve the scope to reduce the impact of item non-response for quantitative amounts in survey data and to minimise phenomena like voluntary under-reporting, memory effect and telescoping. As it is well known, non-reporting and under-reporting cause important biases particularly for self-employment income[7].

The number of employees/ self-employed is computed as number of persons receiving wages/ self-employed income during the reference year. However, from this computation the number of employed persons is overestimated, since it may well be possible that only a fraction of the year was worked, so the number could be different if computed as annual average. This is particularly true for undeclared employment or even regular jobs with very short duration, which are more and more frequent in Italy. Unfortunately, It-Silc questionnaire is not rich in information on job related variables for the income reference year, so a proper number of persons employed cannot be computed. Therefore the comparison between NA and It-Silc in terms of employed persons should be cautiously interpreted.

The administrative sources used are mainly those from social security and insurance obligations. In principle, they cover every registered job regardless the employment status - employees or self-employed persons - and its characteristics in terms of permanent or temporary contract. They can therefore be useful complementary sources to tax registers, whose coverage may be limited when annual gains from labour income are below the no-tax area threshold.

[5] The problem is known and not specific for the Italian case. With reference to cases of multiple jobs of different nature, it is worth mentioning Eu-Silc guidelines: "The growth in self-employment as a secondary activity for employees poses additional problems. Unless such secondary activities are properly covered in an income survey with questions that are just as detailed as those for the primary employment, this too will be a source of under-reporting" – Eurostat (2013b), pp. 320.

[6] See C. Ceccarelli et al. (2008).

[7] As clearly stated also in Eu-Silc guidelines: "Not only are the self-employed less likely than employees to respond to surveys, those that do respond are more likely to under-report their income" – Eurostat (2013b), pp. 320.

In the appendix B we shortly describe all linked archives, separately for employee (six archives) and self-employed (four archives). The set of information provided by each archive is highly heterogeneous in terms of richness and quality for statistical purposes. Employee archives include information on wages, but for public sector workers and the social insurance archive. As for the first, public sector registered wages are covered through fiscal data. Even without any "quantitative" data, "qualitative" signals from social insurance database prove to be useful to cross-validate information on the covered units available from other sources.

Self-employment archives are, in general, poor with respect to data on earnings: the only exception is for outworkers. For all other self-employed persons, by linking each unit to the enterprise he/she works into, it is possible to proceed with estimating a registered compensation taking into account the following attributes:

- the share of ownership of the enterprise;
- the economic results of the enterprise. This is available in a separate database that contains the economic results of all registered Italian enterprises, the so-called Frame-SBS;
- the amount of re-evaluation: in NA operations for estimating non-observed economy, for each enterprise is estimated the undeclared quota of regularly earned profits, according to industry/size-specific selection models[8].
- As Disposing of fiscal codes for almost all (96%) persons aged 16 and more, interviewed in 2012 (reporting income for 2011), we can link It-Silc observations with administrative archives from the employer side.

3.2 Process of integration

Each interviewed person that is found in at least one administrative archive has been labelled as *Admin_employed*. The administrative archive identifies each registered job, so that it is possible that an individual has at the same time a job (or several jobs) as employee and/or self-employed.

At the same way we label as *Silc_employed* (both employee or self-employed), each individual that reports in the survey to have worked and/or to have received some remuneration for work done during the income reference year.

For each record we can therefore have 3 different cases:

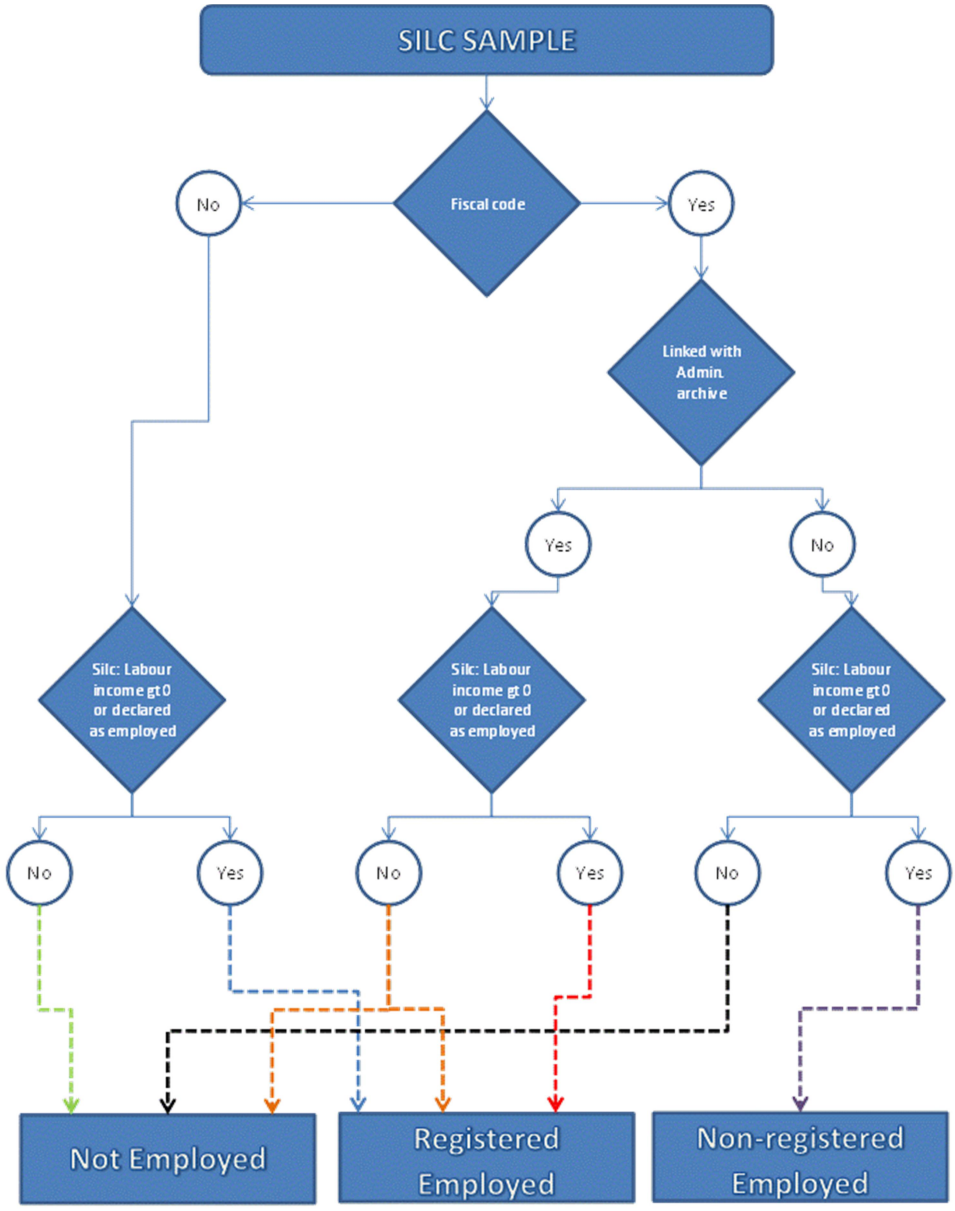
a) units classified as both *Admin_employed* and *Silc_employed*: they constitute the set of potential registered workers;
b) units who are labelled *Silc_employed* and without any information from administrative sources: they form the set of potential undeclared workers as operational hypothesis; at any rate, the latter is valid as a first approximation: for some observations the absence of administrative information might be caused by under-coverage errors;
c) some units are identified only as *Admin_employed*, not having any labour signal from It-Silc. The operational hypothesis for them is twofold: it may be they didn't reply correctly to the survey, both for explicit intent or for accidental problems (comprehension or recalling, for example); in this case they should be assigned to the declared workers category. Alternatively they may come from over-coverage of administrative archives: in such a case, the administrative information should be discarded.

What guides the decision of assigning the activity status and income is, as a general rule, considering the coherence among sources. Figure 1 outlines decisional rules for the occupational

[8] N. Di Veroli et al., 2015.

status assignment, although in an extremely stylized way. It has to be underlined that final assignment entails a cross-classifications for the type of labour income perceived and the occupational status reciprocally consistent.

Figure 1. Decision tree for the integrated activity status assignment



In the following paragraphs, we present the decisional rules of assignment for employees and self-employed. Operationally data analysis to set the rules has to be carried out simultaneously taking into account the type of income and type of archive, in order to detect and solve any inconsistency and, when necessary, re-classify the type of job and/or income. In fact, it may be worth noticing that the operational hypothesis under c) (see supra) is indeed threefold: working by type of labour income and type of archive they may occur combinations *Admin_self-employed* & *Silc_employee* or *Admin_employee* & *Silc_employed* which may be to misclassifications.

3.2.1 Process of integration: assignment rules for employees and results

The potential set of employees is made up by persons who:

e_i. in It-Silc declare to be employee in at least one month or having received wages during 2011; or

e_ii. have been found in at least one employee archive.

As a first step we push out some individuals that cannot be labelled as employees: it happens when they only receive arrears, without any other quantitative or qualitative signals as an employee in the survey. Being NA flows defined on the basis of accrual principle, the decision is straightforward.

The *potential declared employees* are those individuals with signals as employees both in It-Silc and in at least one administrative register (conditions *e_i and e_ii*). The consistent identification coming from both sides is considered as a sufficiently strong signal, without any further investigation. Consequently they are labelled as registered employees and not fatherly treated in terms of income and status in employment. The assigned wage is the maximum between the one declared in the survey and the one resulting from the administrative source. On the hypothesis that on one side the actual income could include some items that is not recorded in administrative data[9] or some part of undeclared compensation. It has to be noted that the resulting income integration is relevant, since it adds up 1.6 percentage points with respect to the total silc starting wage.

The *potential undeclared employees* are those persons who are not found in any employee administrative archive but report wages or employee status in It-Silc (conditions *e_i and not e_ii*, in Table 3). Some of them are reclassified as self-employed: their record matches with self-employed administrative archives, carrying along a self-employment compensation derived from administrative data. The existence of a self-employment register compensation dominates survey's information for them.

The remaining individuals are coded as undeclared employees. As already pointed out, reports for some units of the group may be under-covered by administrative data. However, at this stage, no additional information is available to support a different operational rule. As for their own wages, the survey collects information on income net of taxes and contributions and then the gross value is computed. In this case, labelling them as undeclared jobs, income is accordingly classified as undeclared, but only the net value has been considered; gross income has been set equal to net, because no taxation and contribution are to be paid.

Some more investigation, instead, is performed on the set of sample persons whose information may imply the potential over-coverage of administrative archive or survey non-reporting, i.e. those individuals that are found in an employees' administrative archive but do not have any signal as employee in the survey (conditions *not e_i and e_ii*).

They can be divided into three different subgroups:

- Cases for which it is not possible to validate their employment status, mainly because of the absence of income in the administrative source. They have been considered as not employed: here the hypothesis of archive over-coverage prevails.
- Individuals recoded as self-employed. They are mainly working partners of cooperatives declaring to be self-employed in It-Silc; their classification might be borderline between employee and self-employed, the latter being coherent with social statistics classification. Thus, they were assigned to self-employment, with additional analysis to prevent from potential double-counting for the same job[10].

[9] The administrative data are mainly social security based. Some wages in kind are not subject to contribution.

[10] The alternative would be to classify these individuals as employees and part of the It-Silc income as property income.

- The remaining units are classified as registered employees, their employment income coming from the administrative source: in most cases they gain small entity amounts, with average yearly value below 2,000 €: they may be jobs performed just for small fraction of the year that have been forgotten by interviewed. For this group, the hypothesis of Silc non-reporting prevails. The increase for total wages is negligible.

Table 3. Silc wages and salaries and employees after validation-recoding resulting from administrative archives (in percentage values)

		Wages and salaries	Employees	Silc/NA Wages and salaries, %
	It-Silc starting value	**100.0**	**100.0**	**102.2**
	(-) Arrears	*0.6*	*0.7*	
	It-Silc value net of arrears	**99.4**	**99.3**	
e_i & e_ii	**Potential declared employee**	**91.5**	**89.9**	
	Validated as declared employee & wages integration	93.1	89.9	
	Potential undeclared employee	**4.4**	**6.4**	
e_i & not e_ii	*(-) Recoded as self-employed for misclassification*	*0.1*	*0.2*	
	(-) Subtraction of taxes and contributions for non-registered validated	*1.2*	*-*	
	(=) Validated as undeclared employee	3.1	6.3	
not e_i & e_ii	**Potential Admin over-coverage/misclassification or Silc non-reporting errors**	**0.8**	**3.6**	
	(-) Too weak signals in administrative archives to be validated	*-*	*0.5*	
	(-) Validated as self-employed for the compresence of Admin_Self-employed signals	*0.2*	*0.4*	
	(=) Validated as declared employee; Silc non-reporting error, wages & heads added from administrative sources	0.6	2.7	
	Unmatched observations	**3.4**	**3.7**	
	Total	**100.3**	**102.5**	**102.5**
	Total declared	**93.7**	**92.5**	**101.3**
	Total undeclared	**3.1**	**6.3**	**59.3**

Some more comments on figures from Table 3: 100.3% of the starting value of It-Silc total wages and salaries is validated. It results as the compensation between 1.6 points of integration of compensation from administrative records for registered workers and 0.6 points of integration for Silc under coverage. On the other side, 1.2 percentage points have been discarded for taxes and contribution cancelled for undeclared workers, 0.6 points for arrears and 0.1 for misclassification (workers classified as self-employed).

The total coverage with respect to NA aggregated value (harmonized with It-Silc income concept and population, as explained in paragraph 2) goes from 102.2 to 102.5%. More in deep, this value is the weighted result of a light over-coverage of wages relative to registered workers (101.3%) and under-coverage of wages relative to non-registered workers (59.3%).

The comparison in terms of persons employed is more problematic, as already mentioned. In addition, some cases assigned to non-registered employees could be misclassified in terms of status in employment. Unfortunately, the details of jobs characteristics are not enough to go further.

3.2.2 Process of integration: assignment rules for self-employed and results

The potential set of self-employed is made up by persons who:

se_i. define themselves as self-employed in at least one month or have received some self-employment remuneration during 2011, in It-Silc; or

se_ii. have been found in at least one self-employment archive.

The *potential registered self-employed* are those individuals found with self-employment signals on both sides (conditions *se_i and se_ii*). They are classified as declared self-employed (see Table 4). As for employees, their income has been set to the maximum between Silc income and the income deriving from administrative sources. It has to be specified that in this case the income coming from administrative sources also includes the integration for non-observed economy estimated in Italian NA, so it refers to registered self-employed, not necessarily registered income.

Those discarded as employees belong to the same group (*se_II.1*); they have been classified as declared self-employed persons in the previous stage. Their income is added to self-employment only after checking and eventually sterilizing for double-counting.

The *potential undeclared* are those who have got self-employment status or income in It-Silc, but have not on the administrative side (conditions *se_i and not se_ii*). They remain confirmed as such and their own income is set to Silc net one.

Records satisfying conditions *se_ii and not se_i* (only in administrative archives for self-employment) have their signal validated as declared self-employment when an income in the administrative archive exists. All other cases are discarded, due to the low quality of the sources and the lack of information on income. In this group, also those classified from employees are added.

Table 4 quantifies results for each group.

Table 4. Silc self-employment income and self-employed after validation-recoding resulting from administrative archives

		Self-employment income	Self-employed	Silc/NA Self-employment income*, %
	It-Silc starting value	**100.0**	**100.0**	**64.9**
se_i and se_ii	**Potential declared self-employed**	**91.6**	**87.7**	
	Validated as declared self-employed	110.4	87.7	
	Potential undeclared self-employed	**5.2**	**9.0**	
se_i and not se_ii	(-) Validated as declared self-employed for Admin_employed misclassification	0.0	0.0	
	(-) Subtraction of taxes and contributions for non-declared validated	2.1	0.0	
	(=) Validated as undeclared employee	3.1	8.9	
not se_i and se_ii	**Potential Admin over-coverage/misclassification or Silc non-reporting errors**	**3.1**	**31.6**	
	(-) Too weak signals in administrative archives to be validated	*0.0*	*26.5*	
	(-) Validated as employee for the compresence of Admin_Employee signals	*0.0*	*0.3*	
	(+) Income integration for	*0.2*	*0.0*	

misclassification in Admin_Employee signals			
(=) Validated as declared employee; Silc non-reporting error, added from administrative sources	3.3	4.9	
Unmatched observations	**3.2**	**3.3**	
Total	**120.0**	**104.9**	**77.9**
Total declared	**113.7**	**92.7**	**86.0**
Total undeclared	**3,1**	**8,9**	**14,4**

*NA exclude income from illegal activities

The overall impact of reconciliation is positive and conspicuous, adding +20% for income while only +4.9% in terms of employment, with respect to initial survey figures. The qualitative signals for self-employment are not robust enough too to add more self-employment earners. Income gained from registered workers are, after the integration, over the total initial amount for self-employement: 113.7%.

In terms of coverage of NA aggregates (excluding income from illegal activities), the incidence for total self-employment income after the described process of integration rises from 64.9 to 77.9%: 1/3 of the initial gap between the two sources has been filled. It is worth noticing as the rate for non-declared income flow is remarkably poor: it reaches only 14.4. Survey estimates for registered self-employed fall short macro values, but with a far lower distance: coverage rate reaches 86%.

The under-coverage of Silc total self-employment income can be therefore ascribed to different factors. The number of self-employed from It-Silc seems to be too low, both for non-registered and for registered jobs. This may be due to survey total non-response, item non-reporting or under-reporting, whose impact is known to be consistent for self-employed.

The amount of income for non-registered self-employed is scarce. In addition to what just mentioned, it could be due to a misclassification of the status of employment. Unfortunately, the available information does not allow further investigations. However, it cannot be excluded that NA self-employment compensation for non-registered jobs might be overestimated.

4 Distributional impact of labour income integration

As expected, the integration has an impact on personal income distributions, which is different for the two labour income. Per capita wages and salaries are lower than survey estimates: -1.6% on average and -1.9% for the median. The decrease is more relevant for the lower part of the distribution: the 25th percentile is 6.6% lower. The differences have the same path for males and females and for observations middle-aged. The decrease is higher for residents in Middle Italy.

Table 5. Distribution of wages and salaries

	Initial Distribution				Final Distribution			
	25° pctl	Mean	Median	75° pctl	25° pctl	Mean	Median	75° pctl
Total	11,895	23,323	21,393	30,135	11,104	22,957	20,990	29,776

Sex								
M	14,998	26,110	23,787	32,738	14,160	25,862	23,352	32,418
F	9,884	19,884	18,382	26,699	9,360	19,405	17,876	26,347
Age								
16-29	6,526	14,435	13,142	20,677	5,637	13,760	12,244	20,152
30-49	13,660	23,863	22,290	30,404	13,028	23,610	21,990	30,103
50-64	16,183	28,352	25,648	35,045	15,379	28,190	25,107	34,735
65&over	4,800	15,702	11,709	17,947	3,343	13,558	8,450	15,472
Region								
North	15,104	25,395	23,373	31,632	14,495	25,233	23,012	31,397
Center	12,168	23,873	20,962	29,902	10,929	23,003	20,042	29,085
South & Islands	8,238	19,354	18,260	26,876	7,766	18,994	17,713	26,506

Net income for employees is also lower after integration. However differences are reduced along the entire distribution.

Table 6. Distribution of wages and salaries, net values

	Initial Distribution				Finale Distribution			
	25° pctl	Media	Mediana	75° pctl	25° pctl	Media	Mediana	75° pctl
Total	9,768	16,834	16,332	21,892	9,482	16,528	15,859	21,556
Sex								
M	12,000	18,635	17,901	23,788	11,511	18,350	17,638	23,460
F	8,448	14,612	14,200	19,584	8,216	14,299	13,835	19,255
Age								
16-29	5,432	10,971	10,620	15,400	4,968	10,580	10,182	15,200
30-49	11,148	17,357	17,040	22,203	10,746	17,108	16,701	21,875
50-64	12,725	19,782	18,904	24,655	12,245	19,545	18,530	24,356
65&over	4,263	11,450	8,832	14,400	3,036	10,509	8,276	13,680
Region								
North	12,060	18,175	17,508	22,876	11,612	17,903	17,184	22,564
Center	9,910	17,108	15,832	21,836	9,427	16,617	15,480	21,291
South & Islands	7,105	14,323	14,300	20,089	6,972	14,089	13,835	19,824

Changes are more pronounced for self-employed: average income is 14,5% higher after the record-linkage with administrative data; the increase is even higher for median income: 24.3%. The final distribution is more disperse; in particular for women whose 25th is lower (-1.8%) and the 75th markedly increases: +25%.

Table 7. Distribution of self-employment income

	Initial Distribution				Final Distribution			
	25° pctl	Mean	Median	75° pctl	25° pctl	Mean	Median	75° pctl
Total	7,473	24,467	17,149	30,171	7,396	28,003	21,314	34,000
Sex								

M	9,850	28,037	20,023	33,860	10,333	31,922	24,997	39,316
F	5,011	17,729	12,253	22,422	4,923	20,812	13,960	28,027
Age								
16-29	4,077	13,483	8,971	17,936	3,159	13,200	8,541	18,779
30-49	8,962	24,020	18,155	31,159	9,399	27,514	23,419	35,478
50-64	9,613	30,033	21,099	35,823	10,855	35,288	26,115	41,285
65&over	4,932	26,995	13,544	29,754	5,187	33,126	16,800	35,866
Region								
North	8,621	28,531	19,599	34,518	8,464	31,981	24,677	39,843
Center	8,050	24,674	17,584	30,970	7,953	28,784	21,387	35,378
South & Islands	6,269	17,846	13,872	23,419	6,139	21,054	16,728	28,000

Also net disposable income from self-employment are higher. +16.7% on average and +14.2% for the median value. In this case, change occurs also for the tails: the 25th percentile is 13% lower, while the 75th is 21.1% higher (+25.7% for women). also we can add to It-Silc self-employment income an extra-amount as much as it gets to individual level consistent with NA. As a result we are able to partially fill the gap in Table 5: the coverage rate rises to 90.7% for registered self-employed and 78.1% for the total self-employment income.

Table 8. Distribution of self-employment income, net values

	Initial Distribution				Finale Distribution			
	25° pctl	Media	Mediana	75° pctl	25° pctl	Media	Mediana	75° pctl
Total	5,495	17,014	12,795	22,000	4,780	19,856	14,615	26,643
Sex								
M	7,469	19,330	15,000	24,553	6,505	22,392	18,000	29,000
F	3,654	12,644	9,327	16,663	2,945	15,202	9,721	20,950
Age								
16-29	2,400	9,767	6,987	13,000	1,938	9,652	6,000	13,309
30-49	6,658	16,565	13,650	22,000	6,000	19,260	16,362	27,158
50-64	7,230	20,575	15,689	25,896	6,505	24,781	18,860	30,000
65&over	4,000	19,989	11,534	24,000	3,500	25,205	12,689	28,580
Region								
North	6,724	19,737	14,882	25,000	5,142	22,586	17,000	29,581
Center	5,713	17,169	12,946	22,763	5,300	20,412	14,863	27,139
South & Islands	4,327	12,568	10,214	17,204	3,732	15,072	11,754	21,077

Finally, a look to the household disposable income. As shown in *Table 9* and 10, 87.3% of the population is ranked is the same quintile as compared to survey estimates.

Table 9. Equivalised household disposable income (HY020) quintiles

Initial distribution	Final distribution					
	1	2	3	4	5	Total
1	92.3	4.9	1.7	0.7	0.4	20
2	7.4	86.2	3.2	2.1	1.2	20
3	0.1	8.9	84.0	4.7	2.3	20
4	0.1	0.1	10.9	83.3	5.5	20
5	0.1	0.0	0.2	9.2	90.6	20
Total	20	20	20	20	20	

The percentage of individuals whose quintile is lower than before integration amounts to 7.4%; 5.3% units are, instead, better off. However, the impact may change for different household type: for single, lone parent and couple without children changes in higher quintiles is lower than for other types. Changes occur more frequently for couples with children, both in worse or better situation.

Table 10. Changes in equivalised households disposable income (HY020) quintiles by household type

	Single	Lone parent	Couple w/o children	Couple w 1 child	Couple w 2+ children	Other	Total
Worse off	6.2	8.1	7.0	7.8	8.3	7.3	7.4
Unchange	91.5	88.0	89.7	84.9	84.6	86.6	87.3
Better off	2.2	4.0	3.3	7.3	7.1	6.1	5.3
Total	100	100	100	100	100	100	100

5 Final remarks

In this paper we concentrate on labour income, analyzing how the administrative registers can offer essential insights to explain and partially correct the distance between NA and household survey values: administrative data are the bridge between them.

From the macro viewpoint, they are among the sources regularly used for aggregates estimation. On the other side, they share with household survey the individual perspective. Being characterized by availability of information on each job at a firm level, they allow for simultaneously observing firms and individuals, jobs and persons. However, the counterpart of this potential gain is constituted by the complexity in processing the reconciliation. Complexity arises from the variety of conflicting information and cannot be easily reduced ex-ante.

The current exercise aims to show a preliminary work of the process of reconciliation for the labour market. Our focus mainly consists in validating the It-Silc economic activity status (employed vs not employed) and status in employment (employee vs self-employed), while trying to distinguish registered from non-registered worker and accordingly reclassifying the It-Silc

income. The unique correction for It-Silc values occurs when the assumption of survey under-coverage is sufficiently supported.

The need for complementing NA aggregates with distributional information from household survey data is now well acknowledged. The harmonization of concepts, definitions and statistical domains is a pre-requisite. The core for its implementation consists in disentangling data gaps into their components in order to correctly fill dethe discrepancies avoiding to introduce bias that would hamper the new information potential. The use of administrative registers should be tested and implemented on different domains too, in order to fully develop exhaustive and correct distributive measures.

References

M. Ascione, A. M. M. Carucci, F. C., L. Ciaccia, P. Santoro (2012), *The Households sector in Italy: an analysis for producer and consumer units,* United Nations - Economic and Social Council, Conference of European Statisticians, Geneva, 30 April-4 May 2012

F. Battellini et al. (2015), *Soluzioni metodologiche per l'utilizzo integrato delle fonti statistiche per le stime dell'occupazione*, Istat Working Papers n. 19/2015.

C. Ceccarelli, Coppola L., Cutillo A. and Di Laurea D. (2008), *Combining survey and administrative data in the Italian Eu-Silc experience: positive and critical aspects*, United Nations - Statistical Commission and Economic Commission for Europe, Conference of European Statisticians, Wien, 21-23 April 2008

A. Coli, and F. Tartamella (2014), *Using administrative and survey data to analyse tax evasions from unregistered labour*, IARIW-paper.

C. De Gregorio and A. Giordano (2015), *The heterogeneity of irregular employment in Italy: some evidence from the Labour force survey integrated with administrative data*, Istat Working Papers n. 1/2015.

Eurostat (2013a). European System of Accounts, ESA 2010.

Eurostat (2013b), Description of Target Variables: Cross-sectional and Longitudinal. 2012 operation (Version May 2013), Doc. EU-SILC 065/2013.

M. Fesseau, F. Wolff and M. L.Mattonetti, (2013a), *A cross-country comparison of household income, consumption and wealth between micro sources and national accounts aggregates*, OECD Statistics Working Papers, No. 2013/04, OECD Publishing, 2013.

M. Fesseau and Mattonetti (2013b), *Distributional Measures Across Household Groups in a National Accounts Framework: Results from an Experimental Cross-country Exercise on Household Income, Consumption and Saving.* OECD Statistics Working Papers, No. 2013/04, OECD Publishing.

M. Fesseau and P. Van de Ven (2014), *Measuring inequality in income and consumption in a national accounts framework*, Statistics Brief, Oecd, n. 19.

Stiglitz Commission (2009), *Report on the measurement of economic performance and social progress*. Available from: www.stiglitz-sen-fitoussi.fr

UNECE (2011). *Canberra Group Handbook on Household Income Statistics*; Geneva; Available from: www.unece.org.

J. Zwijnenburg (2016), *Further enhancing the work on household distributional data: Techniques for bridging gaps between micro and macro results and nowcasting methodologies for compiling more timely results* IARIW-paper.

N. Di Veroli, A. Puggioni, F. Sallusti (2015), *L'Economia Non Osservata nei Conti Nazionali*, Italian Economic Association Conference.

Appendix A

We list the main items that have a different content in micro and macro sources.

- Self-employment income. There is not a so-called "self-employment income", but income flows directed to self-employment can be found, in Italian Sector accounts, as mixed income summed to income distributed from quasi-corporations and other income distributed from corporations: these two flows remunerate self-employed working in the corporation sector[11].
- Wages and salaries. In NA concept, wages and salaries do not include the amounts that are paid from employers in case of employees illness, nor include those fringe benefits that can be classified as social benefits. These amount are classified among imputed contributions (in the generation of primary income account) and social benefits (in the secondary distribution of income account). In households micro sources wages and salaries include also these amounts, not social benefits. It is possible (for national accountants) to have separate estimate of these amounts and therefore include them in wages and salaries rather than in imputed contribution and social benefits.
- Imputed rents. In NA any dwelling (including accessory areas) at household disposal generate imputed rents, i.e. residence home and secondary homes (if not rented out). It-Silc collect information only on residence homes, so the value of imputed rents is underestimated in It-Silc by definition. Moreover, the computation method for imputed rents is quite different in the two sources. In NA it is based on a stratification of census data
- Gross operating surplus. In NA it also includes own account production for consumption of agricultural products and gross fixed capital formation, the IT-Silc does not include the information about gross fixed capital formation.
- Interests payed and received. These flows record the most significant difference in definition and content. NA paid and received interests are computed with correction for Fisim, i.e. the intermediation spread is not included in the interests flows but is accounted in consumption of financial services, the same concept apply to insurance claims and premium as computed in NA.
- Investment income attributable to insurance policy holders and pension funds. NA Property income includes also this flow that is not necessarily cashed by households, but are reinvested by insurance company and pension funds. It is an imputed flow that is not recorded among Silc incomes.
- Income from illegal activities. It can be controversial if illegal activity are included in household survey declared income. In principle household may be less reticent when answering to household survey than to fiscal declaration, this is proved when we deal with legal but not observed economy (see section 3) but it may be different when these income arises

[11] Due to the peculiarity of Italian productive system, fragmented in a multitude of small enterprises, Italy derogates Esa2010 that allows for the presence of self-employed only in the household sector (the so called producer households in Italian sector accounts). Many of the small enterprises classified in the corporation sector, have limited responsibility and a full set of accounts, but the economic behavior and the actual management of the enterprise does not significantly differ from the enterprises classified in the household sector, also in the role and the remuneration of its owners. Therefore there are self-employed also in the corporation (and quasi-corporation) sector,. Their remuneration is classified among property income, even if they have a nature of mixed income. The classification of self-employed remuneration in one of the three flows (mixed income distributed to households, withdrawals from quasi-corporation, other income distributed from corporations, relies on the sector (or sub-sector) classification of the enterprise where the self-employed works: productive household, quasi-corporation, corporation.

from illegal activities (in NA drugs and prostitution). So in table 3 we display both coverages, i.e. including and excluding incomes from illegal activities.

- All flows in NA are recorded according to the accrual principle, while in the survey incomes are recorded when they are received. This lead, for example, to a different recording of arrears for labour compensation.

Appendix B

In detail, the administrative archives that have been matched with Silc data are, for employees:

1.1. Social security archive of individual insured position for workers employed in the private sector: it is the archive richest in detail. Each enterprise in the private sector (excluding those in agriculture) with at least one employee has to fill in a form for the social security institution (INPS) each month, with contribution paid by employer and employees and the benefits (family allowances, CIG, maternity allowance, illness allowances etc) that the employer or INPS (through the employer) has to pay to the worker. It therefore contains all information about the 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 (the part subject to contribution) and, for each week of the month, number of worked and not worked days if this affects contribution paid or benefits received;

1.2. Social security archive of individual insured position for workers employed in the public sector: it contains information about the employer and the employee, the date of start and end of the working relationship, but no information on earnings.

1.3. Social security archive of individual insured position for workers employed in the private sector of sport, arts and entertainment: 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.

1.4. Social security archive of individual insured position for workers employed as domestic staff: 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.

1.5. Social security archive of individual insured position for workers employed in agricultural sector: 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, the number of weekly working hours for part-timers and earnings for all workers.

1.6. Social insurance archive 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). For agency workers it also supply the information on the firm where the agency worker is employed. This archive does not always provide reliable date for working period.

For self-employed:

2.1. Social security archive for outworkers: it contains information on each job performed as outworker, the hiring firm and the time length of the professional service. It also have information about the type of work performed, earning and contribution paid. This is a peculiar type of self-employed. In principle, the outworkers are self-employed, since the employer buy the professional service without a specific place and time of work. For this reason they are classified as self-employed. But in practical terms, often enterprises hires outworkers to have a more flexible and less expensive (since social contributions rates are lower) form of fixed term employees and the practical managing of the labour contract is more similar to the one of employees. It is not possible to detect when this happens, so outworkers are always classified as self-employed. But this is the reason why the “perception” of the worker can be different and

they may classify themselves as employees when answering to It-Silc (and classify the income as wage and salary).

2.2. Social security archive 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.

2.3. Social security archive 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 have information about the type of work performed, earning and contribution paid.

2.4. Archive built in Istat for the enterprises census. The enterprise census was entirely performed trough administrative data and produced the number of enterprises and institutions, and their employment (employee and self-employed). So this archive synthetize any 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). This archive has also a procedure that, according to the type of link with the enterprises, validates the number of self-employed. The number of persons employed (employees and self-employed) in each unit is then recorded in the business register.