4th meeting ▪ Rio de Janeiro, Brazil ▪ 9-11th November 2011

Micro-Data Reuse and Micro-Micro Data Linkages to Integrate Macro-Estimations on Agricultural Households' Disposable Income: Istat Experience for Italian National Accounts[1]

Fabio Boncinelli
University of Florence - Deistaf
Piazzale delle Cascine, 35
50144 Florence

Domenico Ciaccia
Istat
Via Cesare Balbo, 16
00184 Roma

Edoardo Pizzoli
Istat
Via Cesare Balbo, 16
00184 Roma

Giuseppe Sacco
Istat
Via Cesare Balbo, 16
00184 Roma

[1] This paper is the result of statistical and economic research in Istat in collaboration with University of Florence and within Blue-Ets European research project www.blue-ets.istat.it .

ABSTRACT

Several statistical surveys on households and businesses are already carried on in most of the countries and their databases have been consolidated. In an era of economic crisis, facing increasingly stringent budget constraints, it is necessary to improve the information capability of available micro-data and to strengthen national accounts estimations with their support .

Although the budget survey is a typical source of micro-data on households, usually there is not a specific focus on the agricultural socio-professional group. However, additional information on agricultural households' income are available from other surveys and national censuses.

Based on micro-data from European survey on households' income (EU-SILC) for Italy, the authors present a study on the possibilities to estimate agricultural household disposable income. As a further step, micro-data from the national survey on agricultural holdings' economic performance (REA) are combined with previous micro-data from EU-SILC, throughout statistical matching methodologies to improve estimations.

Keywords: Income, agricultural holdings, agricultural households, households' budget survey, statistical matching

1. Introduction

Efficiency in surveys design and in available data exploitation is a critical criteria that has to be satisfied in many countries with a stringent public budget constraint. The Italian statistical system is a typical case: it is reach of surveys and data sources but new demand of information are continuously coming from final users, private and public actors, in the economy and society.

This paper considers the estimation of a key economic variable for policy purposes, agricultural household income, that is not anymore produced among Italian statistics. Two separate data-sources are available at the moment in Italy: European survey on households' income (EU-SILC) and agricultural holdings' economic performance (REA) survey. A research effort is underway to test the re-estimation of EU-SILC and statistical matching of micro-data from the two databases. In the next two paragraph the data and methods are briefly summarized. Preliminary results will be presented in paragraph three and possible future developments will be suggested in the final conclusions.

2. Data

2.1 EU-Silc survey

The Statistics on Income and Living Conditions survey (EU-Silc) is a European annual sample survey carry out in Italy by Istat. This survey started in 2004 and is the main source of micro data on income, living conditions and social exclusion in Europe. The reference population consist of private households residing in the country at the time of data collection. The data are gathered for all family members who are more than 15 years old (16 in the rest of Europe).

The sample design has been set up by Istat (ISTAT, 2008) and is regulated by EU (Regulation N° 1177/2003) which defines the minimum effective sample sizes to be achieved. The reference is

to the effective sample size in relation to the 'risk of poverty rate' variable equal to one[2]. The plans are to achieve the minimum effective sample size of around 121.000 over in the EU as a whole. The allocation of the EU sample among countries is a compromise between two aims: the production of results at the countries and EU level as a whole (Eurostat, 2006). Eurostat has delegated national institutes of statistics to implement sample design under constraint to meet the minimum effective sample size.

The list of Italian municipalities is divided into two sets with respect to their demographic weight. The first set represents the largest Italian municipalities by population and is called Self-Representative (SR). The second set is called Non-Representative (NR) and contains the remaining municipalities. Each municipality included in the SR is considered as a stratum and from the list in its registry office the statistical units (households) are systematically extracted. For NR, on the other hand, a two-steps sample design has been implemented. In the first step a list of municipalities are extracted from NR with probability proportional to population size of the municipality and without replacement. In the second step from the list of the NR municipalities households are extracted with the same probability. The sample extraction is carried out systematically under the constraints to select at least 24 households in each municipality and 4 municipalities in every cluster of NR. The use of this sampling design and sampling frame ensure both unbiased estimates and a good regional coverage (ISTAT, 2008). In Italy, the sample size is about 21,000 households which includes about 50,000 individuals.

2.2 REA survey

The agricultural holdings' economic performance (Risultati Economici delle Aziende agricole - REA) survey, in short Agricultural Business Survey, is an Italian annual sample survey carry out by Istat in collaboration with Inea and Regions. This survey started in 1997 and from 2002, after an institutional agreement, Istat is responsible of survey design, methodology and estimations, while Inea and Regions collect the data. REA survey is integrated with European RICA/FADN network carry out in Italy by Inea: a sample for REA that incorporate RICA sub-sample is selected by Istat and data for large agricultural holdings are extracted from RICA database.

This survey is the main source of micro data for holdings of any dimension on the following characteristics: intermediate costs, sales, own final consumption, incomes from all other farm and off-farm sources.

The reference population is the Italian holdings recorded by agricultural census but information are also gathered on holdings' households.

The sample design is stratified and considers several dimensions: regional level, kind of agricultural activity and holdings' dimension. Units are randomly selected without replacement with constant probability under the constraint of a fixed maximum sample error on key economic objective variables. The final estimates are unbiased with a regional coverage. The sample size is about 20,000 holdings and is periodically renewed on a panel based structure.

3. Methods

3.1 EU-Silc ri-estimation

In spite that EU-Silc is clearly not an agricultural budget survey it considers also agricultural households. The questionnaire includes some questions that allow to identify socio-professional

[2] The design effect is the loss of effectiveness by the use of cluster sampling, instead of simple random sampling. The design effect is basically the ratio of the actual variance, under the sampling method actually used, to the variance computed under the assumption of simple random sampling.

categories. The survey gathers information on employee income, self-employment income, current transfers received, capital and rental incomes. Moreover, it is possible to know the economic sector in which each family member is involved. Thanks to this information, it is possible to estimate total households disposable income and the agricultural income component. Therefore households can be classified according to the main source of income and agricultural households can be selected from the set of Italian households. Two typologies of agricultural households ("narrow" and "broad") can be identified according to different definitions of agricultural households (Unece, 2007). An agricultural household is a household where agriculture is the main source of income, that is more than 50% of total income (narrow definition). This is an household of Type A. A further definition of agricultural household includes all households where at least one member earns some income from independent agricultural activity (broad definition). This typology of agricultural household will be called of Type B.

One of the main advantages using EU-Silc instead of an ad-hoc agricultural budget survey, is that it lets to compare income and welfare of agricultural households to the other socio-economic categories and to the whole Italian households. Moreover, it is possible to identify other agricultural household typologies, more than Type A and B, simply using different agricultural household definitions.

The main advantages of this approach is that re-estimation is very simple and fast to be implemented based on a ready-to-use data from an on-going survey. However the sample design and the sample frame are not focused on agricultural household and the estimates incorporate a greater element of error.

3.2 Record-linkage approach

EU-Silc is a survey on total Italian households and doesn't collect specific information on the agricultural households (ISTAT, 2010). However, statistical matching techniques (Rassler, 2002; D'Orazio et al., 2006) allow to produce a new microeconomic database linking the data and introducing more information on agricultural households. Information on total households' income and well-being, gathered by EU-SILC survey can be integrated with observations included in REA database. The resulting dataset, still representative of agriculture as an industry, allows a better statistical representation and a socio-economic characterization of farming households as an institutional sector (Rocchi, 2010).

This exercise has been done by means of statistical matching techniques based on nonparametric imputation methods (hot-deck). To achieve integration between the two data sources the nearest neighbor imputation method was adopted, where the similarity between two records is expressed by an appropriate distance function, calculated with respect to the set of chosen matching variables. For our purposes, the distance function chosen for the matching procedure is the mixed distance (Gower distance) function, in order to take into account of categorical variables among the matching variables. Given the value assumed for the observations a and *b* by *k* variables x_j, available in both databases (matching variables), Gower distance is the following:

$$\tfrac{1}{k}\sum_{j=1}^{k} c_j d_j(a,b)$$

Where, for categorical variables:

$c_j = 1 \quad d_j(a,b) = 0$ if $x_{a_j} = x_{b_j}$ and 1 otherwise;

for continuous variables:

$$c_j = 1/Range(x_j), \quad d_j(a,b) = \left| x_{a_j} - x_{b_j} \right|$$

Different weights can be assigned to the matching variables. Both donor and recipient samples were stratified according to a space variable (Region, to which each observation belong to). Two different regional stratification were tested (5 and 20 regions corresponding to Nuts1 and Nuts2 classifications). To ensure a well balanced stratification both in the recipient and in the donor database the 5 regions stratification was finally adopted. The result of layering is shown in Table 1.

The matching was achieved by placing the constraint that a record could not be donated more than three times; have also been considered as donors not only those with minimum distance but all those who had a distance *d(a, b)* within the range:

Erro! Não é possível criar objetos a partir de códigos de campo de edição.

where: *d(a, b)* is the observed distance between the units a and *b* and d_{min} is the minimum distance observed[3].

Table 1: *Stratification of Observations in the Original Datasets*

Stratum	Freq. in recipient	Freq. in donor	Donor recipient ratio
1	1552	4973	3.20
2	2607	4990	1.91
3	1951	4950	2.54
4	2876	4400	1.53
5	872	1669	1.91
Total	9858	20982	2.13

Source: REA and EU-Silc datasets

4. The size of Italian agricultural households

The weight of Italian agricultural households (Table 2) based on EU-Silc re-estimation is steadily decreasing over the years. In 2004 the agricultural households were only 2% of Italian households, in 2008 the weight reduced to 1.6%.

Table 2: *The Weight of Italian Agricultural Households*

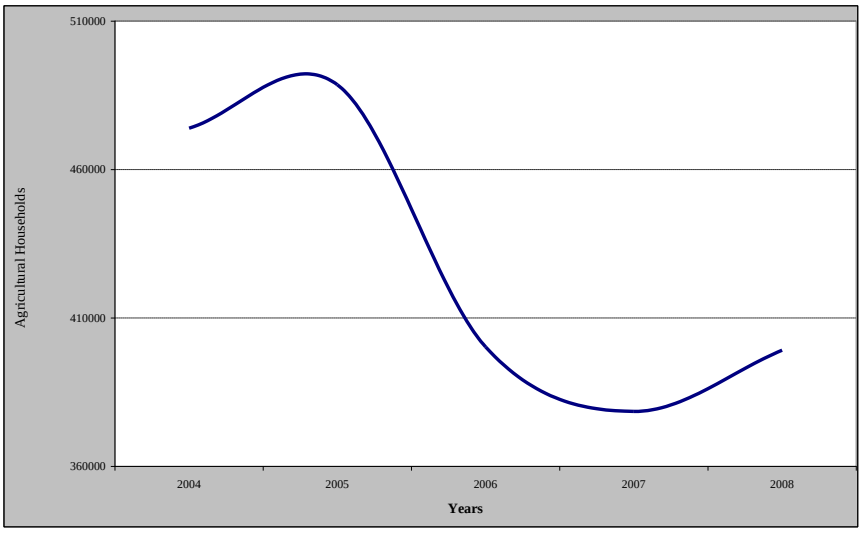
	2004	2005	2006	2007	2008
Non-agricultural households	98.0%	97.9%	98.3%	98.4%	98.4%
Agricultural households	2.0%	2.1%	1.7%	1.6%	1.6%
Total	100%	100%	100%	100%	100%

Source: EU-Silc datasets

[3] The software package used was originally built for the production of an integrated archive for the social accounting matrix of Italian economy. A short documentation for the software is available at the site: http://cenexisad.istat.it/CENEX1/document/Technical_reports_and_documentation/software_on_statistical_matching/SAMWIN_manual.pdf

The Italian agricultural households are nowadays a small socio-professional category, despite they still represent about 1.5 million people. During the same period about 2% per year of agricultural households have left the sector.

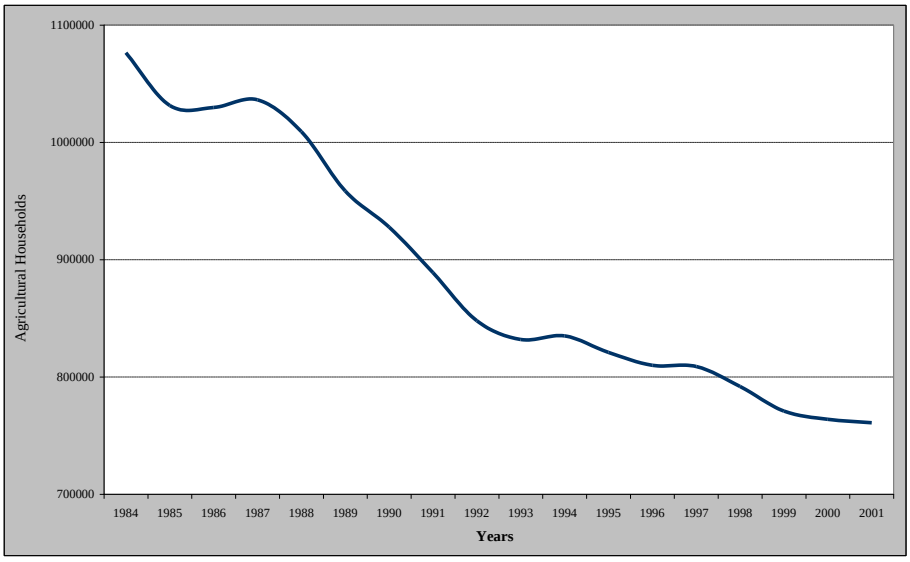
Figure 1: *Italian Agricultural Households. Years 2004 - 2008*



Source: EU-Silc datasets

This reduction confirms the previous trend registered since 1984 (first year in which the data are available). The number of agricultural households has steadily declined since 1984 (Figure 2). In 1984 the number of agricultural households was about one million, in this number reduced to about 700,000 units. From 1984 to 2001 the Italian agricultural households declined steadily by about 1.7% annually (Ciaccia, 2004).

Figure 2: *Total Number of Agricultural Households. Year 1984-2001*



Source: Istat (Ciaccia, 2004)

Unfortunately, these estimates are not completely homogeneous and there are no estimates on the number of agricultural households from 2001 to 2004. In 2004 the Labor Force Survey, at the base of the first period, was completely updated: a new methodology and sampling strategy were adopted. The decline of the number of agricultural households agree with the negative trend of agricultural sector and the difficulty to offer good income opportunities. With respect to the

definitions of agricultural household, in 2008 Type B households are approximately 30% more than Type B households. This is a confirmation of previous results.

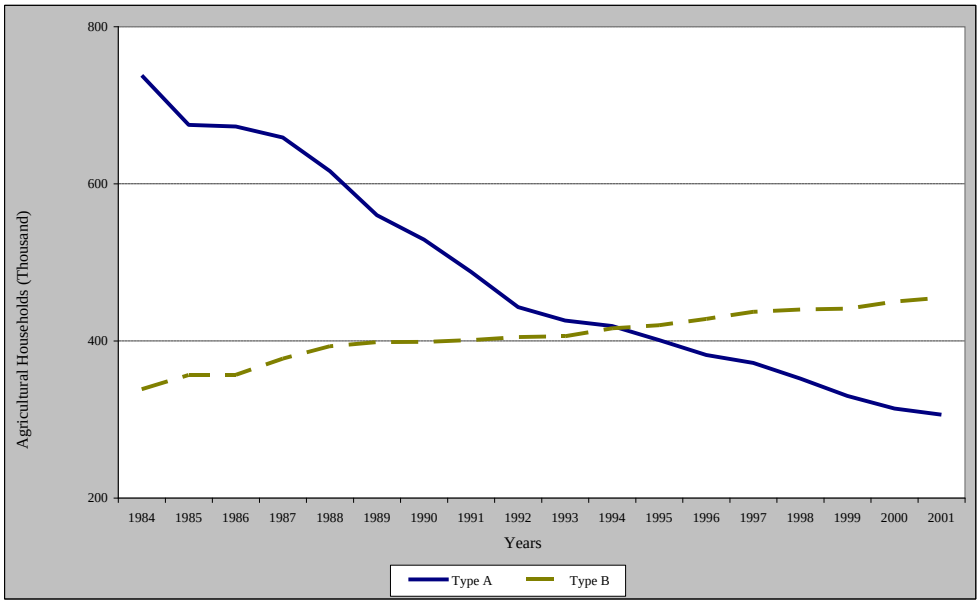
Table 3: *Structure of Agricultural Households*

Years	Type A	Type B	Agricultural A+B
2004	42.6%	57.4%	100%
2005	42.0%	58.0%	100%
2006	40.0%	60.0%	100%
2007	46.4%	53.6%	100%
2008	43.3%	56.7%	100%

Source: EU-Silc dataset

Type A households were the largest group of agricultural households until the mid-nineties, when Type B households became the first group (Figure 3). Type A households reduce in number, while Type B increase because agricultural households extended their income sources to non-agricultural activities. Moreover the decrease of the whole agricultural households in the period 1984-2001 is merely a reduction of Type A households.

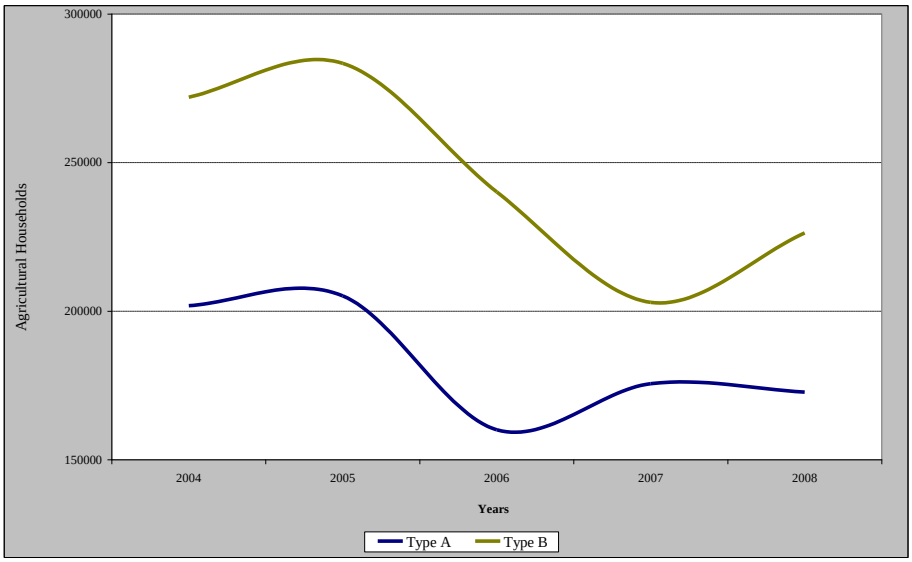
Figure 3: *Number of Type A and Type B Households. Years 1984-2001*



Source: Istat (Ciaccia, 2004)

During the second period of analysis, 2004-2008, the Type A and B households had the same decreasing trend, that is to say that the previous trend has changed and type A are not anymore turning into Type B households.

Figure 4: *The Type A and the Type B Households. Years 2004-2008*



Source: EU-Silc dataset

5. Italian agricultural household's income

Total agricultural households' income estimated with EU-Silc data is defined in the following way: the sum of cash profits (or losses) from agricultural self-employment, cash profits (or losses) from non-agricultural self-employment, the employees' compensation (wages and benefits), pensions (old age, survivors, disability); capital and rental income (interest, dividends, rents of land and buildings), current transfers (subsidy unemployment, sickness, school allowances and bonuses, allowances for accommodation, allowances for family allowances and deductions not otherwise classified) and imputed rent, which are the only figurative items.

In 2004-2008, the total income of agricultural households represented about 2% of Italian households' disposable income, in line with the weight of this sector on the total value added. Considering the total income per household (Figure 4), agricultural households show an average income higher than non-agricultural ones. Especially Type B households have an average income of about 20% higher than the non-agricultural households, while Type A households show an income levels close or smaller with respect to non-agricultural households. However, income per household does not take into account the size of the family. As shown in Table 4, agricultural households are on average larger than non-agricultural ones.

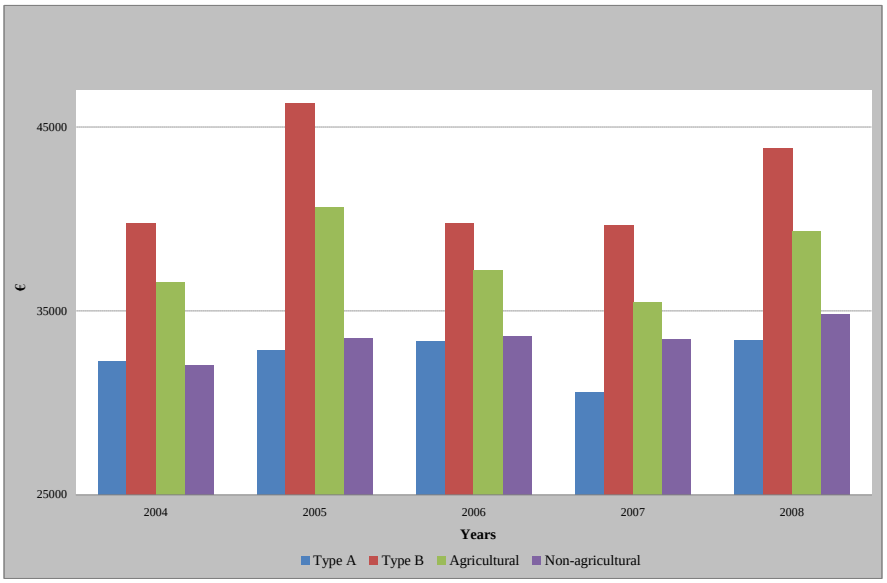
Table 4: *Average Size of Italian Households*

	Year				
Household Categories	**2004**	**2005**	**2006**	**2007**	**2008**
Type A	2.96	2.88	3.01	2.82	3.05
Type B	3.58	3.65	3.74	3.73	3.58
Non-agricultural	2.52	2.53	2.52	2.50	2.49

Source: EU-Silc dataset

Furthermore, Type B agricultural households are larger than Type A and nearly 40% larger than non-agricultural one. This "size effect" explains why Type B agricultural households receive a significant higher portion of income.

Figure 5: *Income per Household*



Source: EU-Silc dataset

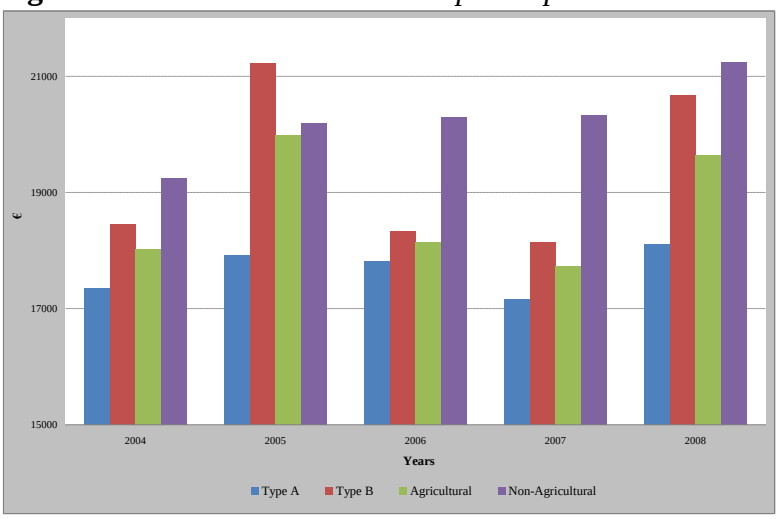
The number of household members is transformed in adult-equivalent with equivalence scales (Deaton, 1997). OECD equivalence scale is the following[4]:

$$\textit{Adult-equivalent household size} = 1 + 0.5 * (HM_{14+} - 1) + 0.3 * HM_{13-}$$

Where HM_{14+} is the number of members older than 14 years and HM_{13-} is the number of members under age 13.

Agricultural households, after transformation, have a systematically lower income than the non-agricultural households. The income gap between agricultural and non-agricultural households is about 10%.. Furthermore, while total income of non-agricultural household shows an average annual growth rate of 2.6%, income of agricultural households grows by 2.3%, despite that income of Type B household has increased of about 3%. However, also Type B households have an income systematically lower (except for 2005 estimation) than non-agricultural households.

Figure 6: *Total Household Income per Capita*



Source: EU-Silc dataset

[4] http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Glossary:Equivalised_disposable_income .

The higher income of Type B households with respect to the Type A households is primarily due to a larger share of income deriving from more profitable non-agricultural activities.

However, the better income performance of Type B households could be largely attributable to the demographic structure of this household category. As shown in Table 5, Type B households have more active members than the Type A and non-agricultural ones.

Table 5: *Adult-Equivalent Household Size by Socio-Professional Category*

	Years				
Household categories	**2004**	**2005**	**2006**	**2007**	**2008**
Type A	1.9	1.8	1.9	1.8	1.8
Type B	2.2	2.2	2.2	2.2	2.1
Agricultural household	2.0	2.0	2.1	2.0	2.0
Non-Agricultural household	1.7	1.7	1.7	1.6	1.6

Source: EU-Silc dataset

Median incomes (Table 6) are systematically lower for agricultural households than non-agricultural households, while non-agricultural households have the highest per capita incomes despite they are composed by only one and half economically active members and agricultural households are composed by 2 members.

Agricultural households earn systematically lower median incomes even in non-agricultural economic sectors (Table 6). This is an evidence of a disadvantaged economic environment for people living in rural areas. Although they are willing to diversify into other economic sectors of activity, they meet worse income opportunities than non-agricultural households.

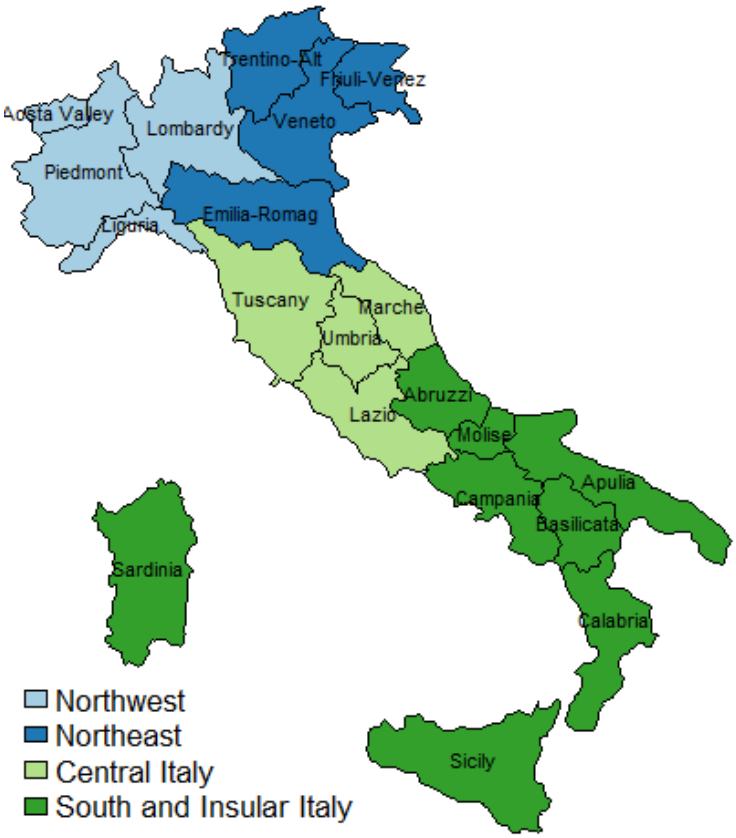
Table 6: *Median Income by Socio-Professional Categories. Year 2008*

	Agricultural Income	**Non-Agricultural Income**	**Imputed rent**	**Employment income**	**Pensions**	**Capital and rental income**	**Transfers**
Type A	17 013	5 735	6 239	7 333	8 476	639	1 080
Type B	8 948	13 149	6 236	15 748	13 130	744	1 200
Non-agricultural	-	16.000	6 370	20 056	14 391	650	1 044

Source: EU-Silc dataset

EU-Silc re-estimations allows to compare regional differences, subdividing Italy into four macro-areas: Northwest, Northeast; Central Italy, South and Insular Italy.

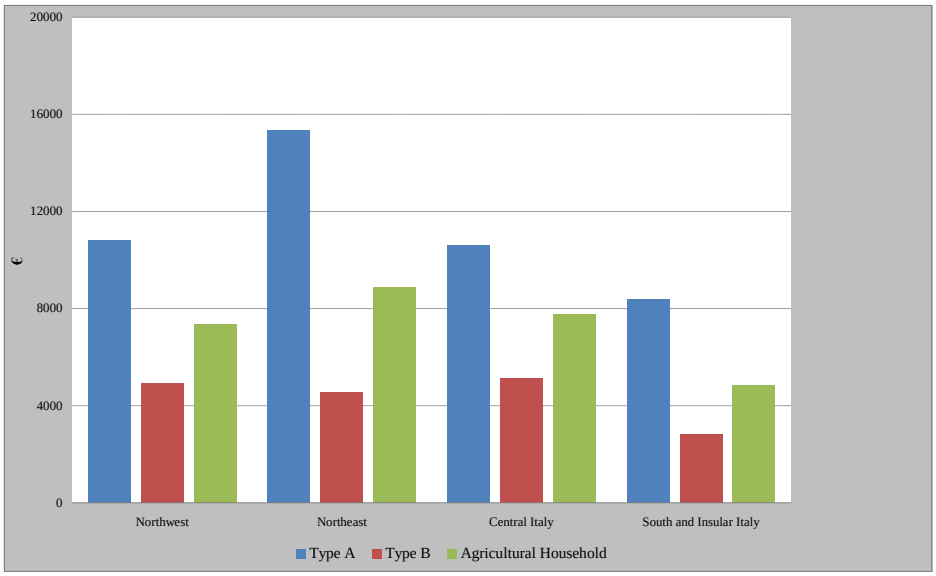
Figure 7: *The Italian Macro-regions*



Source: EU-Silc dataset

North-East is the area with the highest per-capita agricultural income of Type A household and with the largest gap between Type A and B households. In other words, North-East has the best income opportunities in agricultural activity. Households of Type A earn an agricultural income double than the same group in South of Italy. This area seems to be the most disadvantaged one according to income opportunities in agriculture that seem to reflect the low productivity in this sector.

Figure 8: *Agricultural Income per Capita by Macro-areas. Year 2008*

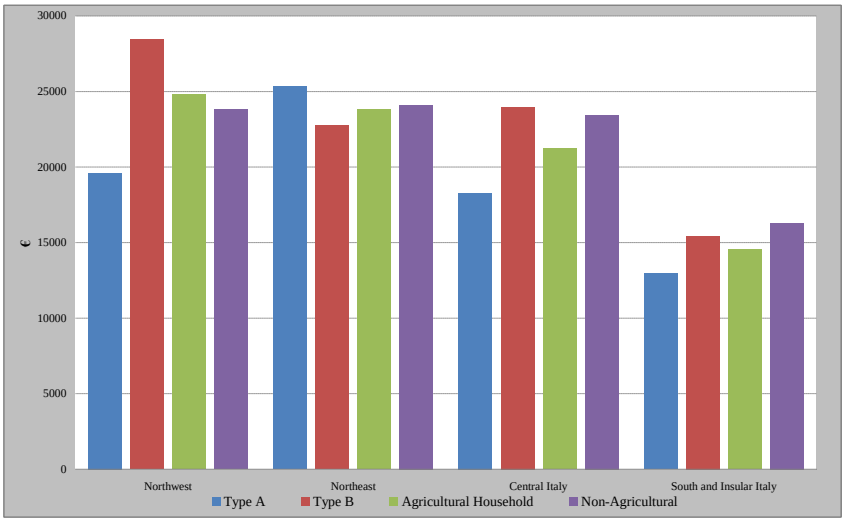


Source: EU-Silc dataset

Despite in this area there are almost 40% of agricultural households, only 26% of the total income of agricultural households belongs to them.

Moreover, total income shows how the regional income gap between Northern and Southern Italy are reflected on income of agricultural households. On average, Southern agricultural households' income is about 60% of their counterparts in other areas. The Central and North-west instead are characterized by Type A households with a very low total income than non-agricultural households, while households of type B show similar levels in these two areas or, in the case of the Northwest, better than non-agricultural households. The Northeast is characterized by Type A households with higher total income as a confirm of the development of agriculture in this area, that generate fair incomes for agricultural households.

Figure 9: *Total Income per Capita by Macro-areas. Year 2008*



Source: EU-Silc dataset

Thanks to the matching procedure, analysis on agricultural households' income can be improved using information from REA survey. The matching procedure was based on a set of variables including: total household's income, number of household's members with at least some source of personal income, total income composition by source, prevalence of income from agriculture activity. The definitions of these variables in the two datasets are harmonized. The 5 regions matching procedure, implemented in this work, was tested with two different set of weights assigned to matching variables: the second test, carried out assigning a larger weight to total household income was retained.

The potential interest of the matching experiment can be highlighted with results in Table X, where summary figures on the distributive features of the agricultural households sector are displayed for Italy. Households are classified according to the prevalence of income from agriculture (agricultural vs. non agricultural) and by income quintile.

A good example of the potential utility of the new dataset is the percentage of Single Farm Payment (SFP) from the Common Agricultural Policy, accruing to each household group (fourth column). This kind of analysis is not possible with only Eu-Silc information.

The SFP, a direct transfer decoupled from the level of farm production, is the most important measure within the EU Common Agricultural Policy, in supporting farmers' income. There is a distributive bias from REA&EU-Silc date analysis: the 7.7% of agricultural households included in the two highest quintiles gather more than 40% of SFP; furthermore for the richest agricultural households more than 15% of total income is represented by SFP.

Table 7: *Distributive Features of the Family Farming Sector in Italy. Year 2008*

Income Quintile	% Households	Average per capita equivalent income (€)	%SFP	SFP/total household income (%)	Well-being index
agricultural 1	26.1	3 214	25.1	31.9	0.4
agricultural 2	4.0	10 626	11.0	27.7	0.5
agricultural 3	3.2	15 077	12.3	24.3	0.6
agricultural 4	3.9	19 224	19.6	27.2	0.6
agricultural 5	3.8	39 427	20.7	15.3	0.8
non agricultural 1	33.7	4 789	5.1	3.1	0.4
non agricultural 2	10.3	10 647	2.3	2.0	0.5
non agricultural 3	6.3	14 780	1.0	0.9	0.6
non agricultural 4	4.7	19 559	1.6	1.5	0.7
non agricultural 5	3.9	40 823	1.3	0.8	0.8
Total	100.0	10 153	100.0	9.7	0.5
Q5/Q1 agr	0.1	12.3	0.8	0.5	1.8
Q5/Q1 non agr	0.1	8.5	0.3	0.3	1.9

Source: REA & EU-Silc datasets

The last column shows the average value of a composite well-being indicator including income level as well as information on housing conditions, education level, health status and social exclusion[5]. The index is based on new information from EU-Silc survey assigned to farmers included in the REA sample through the matching procedure. The availability of well being indicators may represent a further possible development in statistical supply enhancing the targeting of agricultural policy.

Conclusions

Results presented in this paper on micro-data integration and re-estimation suggest that information available in statistical databases are already very rich and useful to satisfy new user needs if appropriate statistical techniques are applied to data. Furthermore, this exercise indicates that limited interventions on survey design, introducing or changing some key questions in the questionnaires or a different stratification in the sample design, can improve future data usefulness.

The statistical techniques tested in the paper seem to be appropriate and developments are underway. A further disaggregation of agricultural households and a smaller territorial level, such as rural areas, is feasible and should be investigated.

REFERENCES

Ciaccia D. (2004) *Azione TAPAS 2002*. Roma: Istat-Eurostat.

Deaton A. (1997) *The Analysis of Household Survey – A Microeconometric Approach to Development Policy*. Baltimore, USA: The Johns Hopkins University Press.

Eurostat (2006) *Comparative EU statistics on Income and Living Conditions: Issues and Challenges*, Proceedings of the EU-SILC conference (Helsinki, 6-8 November 2006).

D'Orazio, M., Di Zio, M. and Scanu, M. (2006) *Statistical matching. Theory and practice.* Chichester (UK), John Wiley and Sons.

Istat (2008) *L'indagine Europea sui Redditi e le Condizioni di Vita delle Famiglie* (EU-Silc), Metodi e Norme N. 37. Roma: Istat.

[5] The index is the geometric average of a set of class variables normalized with reference to the variation field; the index as assumes a value included between 0 and 1; the aggregation through geometric averaging expresses a partial substitutability among different dimensions of well being (OECD, 2008).

Istat (2010) *La distribuzione del reddito in Italia*, Collana Argomenti N. 38. Roma: Istat.
OECD (2008) *Handbook on constructing composite indicators. Methodology and user guide*, Paris: OECD.
Pizzoli, E., Rocchi, B. and Sacco, G. (2011) *An application of statistical matching techniques to produce a new microeconomic dataset on farming households' institutional sector in Italy*, SIS Conference, Bologna.
Rassler S. (2002) *Statistical matching : A frequentist theory, practical applications and alternative Bayesian approaches*. New York, Springer.
Rocchi B. (2010) *Gathering information on total household income within an "industry oriented" survey on agriculture: methodological issues and future perspectives*. In Pizzoli, E. (ed) Statistics on Rural Development and Agriculture Household Income. Proceedings of 2nd Meeting of the Wye City Group. Rome, 11-12 June 2010. Roma, ISTAT: 521-528.
UNECE (2007) *The Wye Group Handbook, Rural households' Livelihood and Well-being*, New York and Geneva.