An Application of Statistical Matching Techniques to Produce a New Microeconomic Dataset on Farming Households' Institutional Sector in Italy

Edoardo Pizzoli, Benedetto Rocchi, and Giuseppe Sacco

Abstract

A new microeconomic database on farm households in Italy was created using statistical matching techniques. Information on total households' income and well-being gathered by the EU-SILC survey on living conditions for Italy was attached to the observations included in the FBS database for Italy. The new dataset, still representative of agriculture as an industry, also allows a proper statistical representation and socio-economic characterization of farm households as an institutional sector.

The quality of the new microeconomic information was assessed analysing the statistical properties of key analysis variables and the distributive features of the current UE Common Agricultural Policy.

1 Introduction

In carrying out insightful analyses of distributive implications of alternative agricultural policy options, suitable microeconomic information on potential beneficiaries is needed. Two main features seem to be relevant to the analysis. First, the institutional sector of farm households needs to be properly placed within the economy-wide income distribution, observing the total household income (THI) (Unece et al. 2007); second, information should be available to classify households both using information on the farm (such as size, product typology, management

E. Pizzoli (✉) • G. Sacco
ISTAT, Rome, Italy
e-mail: pizzoli@istat.it; sacco@istat.it

B. Rocchi
University of Florence, Firenze, Italy
e-mail: benedetto.rocchi@unifi.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_15,

form) and information on well-being of the household itself (such as composition, age, education, health).

The main sources of microeconomic information on the institutional sector of farm households, such as the Farm Business Survey (FBS) carried out by ISTAT or the European Farm Accountancy Data Network (FADN), fail to comply with both these characteristics: their focus on technical aspects and the centrality given to income from farming makes these surveys suitable for analysis only within an industry (agricultural) perspective.

This paper aims to propose a possible solution to this information problem. In the next paragraph a description of data and methods used in the analysis will be proposed. The assessment of new dataset produced through matching techniques with some exemplificative results will follow. In the subsequent paragraph a statistical analysis will be performed on key objective variables included in the datasets result of several tests. A final paragraph will show some figures on distributive features of the farming households' sector in Italy resulting from the new selected dataset.

2 Data and Methods

A new microeconomic database on farm households in Italy was created using statistical matching techniques (Rassler 2002; D'Orazio et al. 2006). Information on total households' income and well-being gathered by the EU-SILC survey on living condition for Italy (ISTAT 2010) was attached to the observations included in the FBS database for Italy (ISTAT 2011). The new dataset, still representative of agriculture as an industry, also allows a proper statistical representation and socio-economic characterization of farm households as an institutional sector (Rocchi 2010).

The FBS, designed to supply information for national accounts, yearly surveys a sample of agricultural holdings representative of the Italian agriculture. The database includes a detailed set of variables on farm structures (such as cultivated area, livestock number, labour employment) and on costs and revenues from farming. According to these information a good estimate of income from farming can be obtained. Furthermore, for the farm households (the largest part of the sample) a small set of variables on household's composition as well as on extra-farm source of income (classes of income by four types of sources) is available (Pizzoli 2005).

The EU-SILC is a sample of Italian households designed to gather detailed information on incomes as well as on living condition and well being. The sample is representative of total Italian population but, given the optimization criteria adopted in the design of the survey, *farm* households are under-represented (520 observations from a total of 20,982, that is 2.48 %). The dataset includes variables on occupation, professional position and income sources by type of single household's members; a number of nominal variables expressing well-being of household's members and family living condition are available as well.

Table 1 Matching variables

Variable name	Description	Continuous	Type
ncomp	Number of household members	No	Scale
ex_i	Extra-farm net income from self-employed labour	Yes	Scale
ex_d	Extra-farm net income from hired labour	Yes	Scale
ex_p	Extra-farm net income from pensions	Yes	Scale
ex_c	Extra-farm net income from capital assets	Yes	Scale
yagrn	Net income from farming	Yes	Scale
redtotale2	Total household income (THI)	Yes	Scale
Quex_i	Share of extra-farm income from self-employed labour	Yes	Scale
Quex_d	Share of extra-farm income from hired labour	Yes	Scale
Quex_p	Share of extra-farm income from pensions	Yes	Scale
Quex_c	Share of extra-farm income from capital assets	Yes	Scale
agrn	Net income from farming more than 50 % of THI	No	Binary
redtotale2_pc	Per capita total household income	Yes	Scale
decile	Income decile	No	Ordinal

Farm households in FBS and EU-SILC samples can be assumed to be homogeneous, as they represent the same typology of statistical units, and coming from the same target population, according to the following units definition: "households . . . that derived any income, however minor, from agriculture or contributed some labour input to agricultural production". ("broad" definition, Chapter IX, The Agricultural Household—Concepts and Definitions; Unece et al. 2007). Farm households, by construction, belong to the larger households population and are a specific typology (socio-professional) group.[1]

For the aim of the analysis a sub-sample of 9,858 observations representative of 1,586,193 farm households from the FBS (year 2007) was considered as the "recipient" database. A set of 14 "matching variables" on households' characteristics was defined according to available information. The criterion followed in the variables selection was the possibility to exactly replicate them for each observation included in the "donor" EU-SILC database (year 2007). Table 1 lists the matching variables with some information on them.

[1]In the EU-SILC survey a "'private household' is a person living alone or a group of people who live together in the same private dwelling and share expenditures, including the joint provision of the essentials of living" (Art. 2, Definitions; EU 2003). This definition is equivalent to the UN definition (UN 1998) adopted by Eurostat and EU members countries.

A farm household in the FBS sample is defined as a household with at least a spouse that manages an unincorporated or quasi-corporate agricultural holding (individual farms; communal tenures), and works in the agricultural holding. A farm household in the SILC sample is defined a household with at least a family member earning incomes from self-employed labour in agriculture, according with individual records where incomes are classified by sector of economic activity. For a discussion on the definition of agricultural household see Chapter IX of the UN Handbook (UN 2011).

Table 2 Regional stratification of observations in the original datasets

Region	Frequency in recipient	Frequency in donor	Donor to recipient ratio
1	1,552	4,973	3.20
2	2,607	4,990	1.91
3	1,951	4,950	2.54
4	2,876	4,400	1.53
5	872	1,669	1.91
Total	9,858	20,982	2.13

Both donor and recipient samples were stratified according to a space variable (the region each to which observation belongs). Two different regional stratifications were assessed (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 2.

The donor to recipient ratio shows a good distribution of the 20,982 donors with respect to the 9,858 recipients.

The integrated archive was built by means of statistical matching techniques based on nonparametric imputation methods (hot-deck). More precisely in the realization of the matching between the two files was used the method of nearest-neighbour imputation where the proximity between two records is expressed by an appropriate distance function.

The distance function chosen for the matching procedure is the mixed distance (Gower distance), in order to take into account the presence of discrete variables between the matching variables. Given the value assumed for the observations a and b by k variables x_j available in both databases,

$$\text{Gower} : \frac{1}{k}\sum_{j=1}^{k} c_j d_j\,(a,b)$$

where: for categorical variables: $c_j = 1$, $d_j(a,b) = 0$ if $x_{a_j} = x_{b_j}$ and 1 otherwise; for continuous variables: $c_j = 1/\text{Range}(x_j)$, $d_j\,(a,b) = \left|x_{a_j} - x_{b_j}\right|$

For each matching variable, the range is calculated considering the observations of both samples. Indicated with A and B respectively the set of possible values that can assume the variable x_j in the set of donors and in the set of recipients will have:

$\text{Range}\left(x_j\right) = x_{j_1} - x_{j_2}$, where $x_{j_1} = \max\left(x_{j_i}/x_{j_i} \in A \cup B\right)$ and $x_{j_2} = \min\left(x_{j_i}/x_{j_i} \in A \cup B\right)$

The matching was achieved by placing the constraint that a record could not be donated more than two or 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:

Table 3 Parameters adopted in the replications of matching

Name	Maximum number of donations	w_{thi}
Test1	2	0.50
Test2	3	0.50
Test3	2	0.75
Test4	3	0.75

$$d_{\min} - 0.01 \leq d\ (a, b) \leq d_{\min} + 0.01$$

where $d(a, b)$ is the observed distance between the units a and b and $d_{\min}$ is the minimum distance observed.

Different weights can be assigned to the matching variables. Given the aim of the analysis (to create an improved dataset to ground the estimate of the total income of farm households) in the matching procedure the largest weight was assigned to *redtotale2*, the variable representing the THI, respectively 0.50 or 0.75.[2] Given w_{thi} the weight assigned to THI, a weight equal to $(1 - w_{\text{thi}})$ was equally subdivided among the other matching variables.[3]

Combining the maximum number of donation of the same record from donor dataset with the two set of weights results in 4 replications of the matching procedure according to Table 3.

3 Statistical Checking

To reliable final analysis of results from statistical matching procedure, it is important to check if small changes of the key matching variable (THI) weight, w_{thi}. significantly affect the parameters of the resulting distribution for the variables of analysis generating probable unstable results. This statistical checking has been carried out on *ncomp*, the control variable available for all the datasets (including FBS), and *redtotale2_pc*, the objective variables of the Tests.

Descriptive statistics for *ncomp* and *redtotale2_pc* are reported in Table 4.

The distributions of the two variables have the same shape, asymmetric with long tails to the right-end-side and more peaked than normal distribution. Parameters slightly change with the different Tests.

[2]So far the choice of the distribution weight was oriented by *a priori* considerations of the subjective nature: it was thought that the variable THI (redtotale2) should be more decisive in the matching process. In a future development the correlation between matching variables and the variables of interest in the donor database will be assessed as a possible criterion in choosing distribution weights.

[3]The software package used in this paper was originally built for the production of an integrated archive for the social accounting matrix of Italian economy. A short documentation on the software is available in the Manual (Sacco 2008) at the site http://cenex-isad.istat.it.

Table 4 Descriptive statistics (valid $N = 9858$)

Var.	Mean	Confid. −95.0 %	Confid. +95.0 %	Median	Min.	Max.	SD	SK	KU
ncomp									
FBS	1.7736	1.7552	1.7921	2	1	16	0.93548	1.859	9.304
Test1	1.5021	1.4875	1.5167	1	1	6	0.73799	1.537	2.395
Test2	1.5346	1.5199	1.5494	1	1	6	0.74648	1.419	1.895
Test3	1.5371	1.5223	1.5518	1	1	6	0.74704	1.421	1.969
Test4	1.5469	1.5321	1.5618	1	1	6	0.75257	1.414	1.937
redtotale2_pc									
Test1	16,146.4	15,921.3	16,371.4	13,594.8	0	86,165.4	11,398.9	0.876	0.302
Test2	16,595.4	16,373.6	16,817.2	13,594.8	0	92,923.2	11,236.5	0.926	0.852
Test3	16,609.7	16,386.9	16,832.5	13,594.8	0	86,866.1	11,284.1	0.957	1.045
Test4	16,639.1	16,415.4	16,862.9	13,672.5	0	94,167.2	11,334.0	1.018	1.445

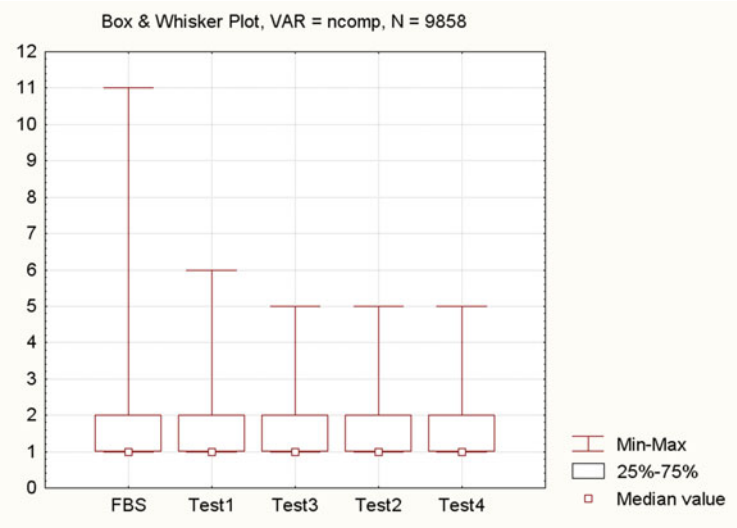

Fig. 1 Box and Whisker plot for the number of household members

Considering the number of household members, FBS variable has a much higher range and variability, while Test2, Test3 and Test4 variables are very close (see Fig. 1)

Considering per capita THI, all the Tests variables are very close and only Test4 variable shows a higher range of variability (see Fig. 2).

A *t*-test has been used to evaluate the differences in means between pairs of variables (Table 5).

Considering the first variable, *ncomp*, the *p*-levels reported suggest that the research hypothesis about the existence of a difference in means can be accepted,

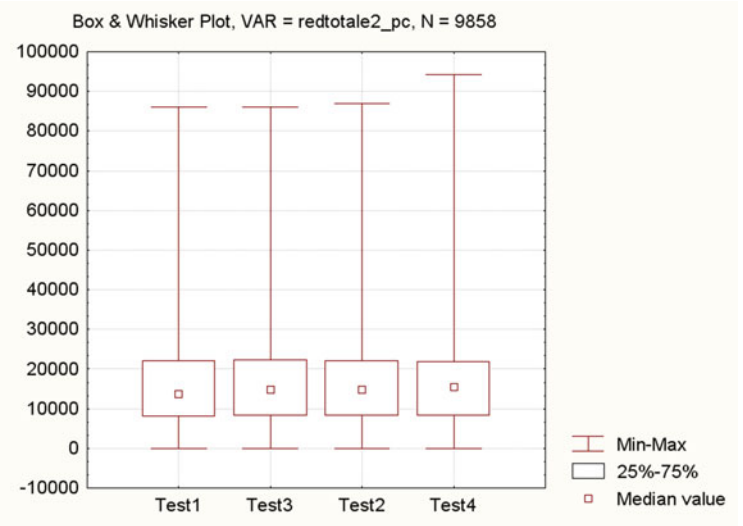

Fig. 2 Box and Whisker plot for per capita total household income

Table 5 *t*-Test for difference in means, independent samples[a] (valid $N = 9858$)

Var.	*t*-value	*df*	*p*	*F*-ratio variances	*P* variances
ncomp					
FBS vs.Test1	22.62803705	19,714	0.000000000	1.606811236	0.00000000
FBS vs.Test2	19.82692898	19,714	0.000000000	1.570483757	0.00000000
FBS vs.Test3	19.61920737	19,714	0.000000000	1.568120907	0.00000000
FBS vs.Test4	18.74891138	19,714	0.000000000	1.545132341	0.00000000
Test1 vs. Test2	−3.079968607	19,714	0.002073072	1.023131394	0.25631243
Test1 vs. Test3	−3.308986123	19,714	0.000938023	1.024673052	0.22632157
Test1 vs. Test4	−4.223467879	19,714	2.41651E-05	1.039918195	0.05202248
Test2 vs. Test3	−0.228886566	19,714	0.818959462	1.001506804	0.94042060
Test2 vs. Test4	−1.149698167	19,714	0.25028213	1.016407278	0.41918106
Test3 vs. Test4	−0.921314689	19,714	0.356897437	1.014878056	0.46349586
redtotale2_pc					
Test1 vs. Test2	−2.78528	19,714	0.005353	1.029113	0.154293
Test1 vs. Test3	−2.86820	19,714	0.004133	1.020455	0.314840
Test1 vs. Test4	−3.04351	19,714	0.002341	1.011479	0.570993
Test2 vs. Test3	−0.08936	19,714	0.928797	1.008485	0.674905
Test2 vs. Test4	−0.27205	19,714	0.785583	1.017434	0.390916
Test3 vs. Test4	−0.18251	19,714	0.855185	1.008873	0.660998

[a]The selected samples, from FBS survey and the four tests, are assumed to be independently generated with respect to the two objectives variables: number of household members (*ncomp*) and per capita total household income (*redtotale2_pc*). Relaxing this assumption, the power of *t*-test should be considered with respect to other tests

Table 6 Extra-farm income estimates (Mio€, 2007)

	FBS	Test1	Test2	Test3	Test4
ex_i	3,005	6,038	6,180	6,401	6,197
ex_d	6,956	12,898	13,477	13,547	13,688
ex_p	6,801	4,077	4,345	4,375	4,359
ex_c	151	4,831	5,358	5,378	5,356
Total	16,914	27,844	29,360	29,702	29,600

as expected, comparing FBS variable with the Tests variable. Test1 variable also do not pass the test with respect to the other Tests variable.

Considering the second variable, *redtotale2_pc*, the same results are confirmed between the Tests variable.

If a one-way ANOVA is computed on all four *ncomp* and *redtotale2_pc* Tests variables, the previous results are confirmed only for the first variable: *ncomp* mean in Test1 significantly differs from the other means in Test2–4 (F-value = 6.71, $p = 0.0002$), while for *redtotale2_pc* the means can be considered not significantly different (F-value = 2.48, $p = 0.0589$). If the same analysis is replicated only on Test1–3 variables, the null hypothesis of equal mean can be accepted at 5 % significance level for both *ncomp* and *redtotale2_pc*.

This is an indicative result for the matching procedure: Test1 considers a weight for the key matching variable equal to 0.5, and changing the donation from 2 (Test2) to 3 (Test1) can change the mean estimation. A higher weight for THI assures a greater stability of results.

Finally, a comparison among the four final databases is proposed in Table 6. Each row shows alternative estimates of the total extra-farm income by different sources. The first column displays the totals that could be estimated using only information included in the original FBS database[4] while the others show the estimates obtained using information originally included in the EU-SILC database and "matched" with FBS records according with the procedure described above.

Two relevant results can be stressed. First, the use of the new database, whatever the replication considered, leads to a quite different estimate of totals (namely, larger for self-employed labour, hired labour and capital asset incomes, and smaller for income from pensions); second, the outcome of the matching procedure does not seem to be sensibly affected by changes in the parameters (max number of donations and weights assigned to matching variables).

[4]More precisely the estimate was based on the matching variables. In the FBS only classes of extra-farm incomes (by source) are collected. To estimate the absolute value of each income component an average value was associated to each class. The average value of each class for each income component was estimated through regression using the EU-SILC database, where single income sources are collected in absolute value, and used to prepare the matching variables both in the recipient and in the donor database.

Table 7 Comparison among alternative matching results

	Test1	Test2	Test3	Test4
Total income: combined vs. matched				
Percentage difference	45.2	42.7	42.3	42.4
Correlation	0.585	0.596	0.626	0.610
Average Gower distance	0.043	0.044	0.036	0.046

A further comparison between the four replications is proposed in Table 7. In the first row the total income of households estimated combining the income from farming from FBS data with extra-farm income matched from EU-SILC, is compared with the THI of "donor" observations (as quantified in the original EU-SILC dataset). Not surprisingly the large, positive percentage difference shows that the farming component of total income would be underestimated using EU-SILC data alone. Test1 shows a larger difference (>45 %) in the estimate of totals, while the other three replications lead to quite similar results. The best correlation between "combined" and "matched" total income variables is shown by Test3. Finally, the average value of the Gower distance in the space of the matching variables between recipient (FBS) and donor (EU-SILC) records is proposed in the last row. Again, the best performance is shown by Test3, the only one with an average distance lower than 0.04.

4 Some Preliminary Results

Overall, these results seem to show that a relevant information may be added to the original FBS using statistical matching techniques. Furthermore, the matching procedure yields results quite robust in front of variation in the values of parameters. To highlight the potential interest of the matching experiment in this paragraph the Test3 database is used to estimate some figures on the distributive features of the farming households' sector in Italy.

In Table 8 some figures on the distributive features of the farming households' sector in Italy are displayed. Families are classified according to the prevalence of income from farming[5] (agricultural vs. non-agricultural) and by income quintile. The reader should bear in mind that income quintiles were defined taking into account the whole Italian population, not only the sector of farm households. As a consequence in Table 8 the households are not equally distributed among quintiles: figures in the first column show the position of households managing agricultural activities in Italy within the *overall* income distribution.

For the largest part of households involved in agriculture farming is only a secondary source of income. "Agricultural" households in a narrow sense (income from farming is more than 50 % of THI) are less than 20 %. Noticeably, in the

[5] A household is classified as "agricultural" if farming supplies more than 50 % of the THI.

Table 8 Distributive features of the farming households' sector in Italy, 2007

Income quintile	Percentage of households	Average per capita income (€)	Percentage of net income from farming	Percentage of SFP	SFP/total household income (%)	SFP/net income from farming (%)	Well-being index
Non-agricultural 1	43.1	7,714	9.3	11.4	2.9	26.5	5.9
Non-agricultural 2	15.2	14,777	6.2	6.4	3.0	22.3	6.4
Non-agricultural 3	12.1	19,218	6.6	6.1	3.2	19.9	6.7
Non-agricultural 4	6.8	24,713	6.9	6.5	4.6	20.3	7.5
Non-agricultural 5	5.7	46,925	10.4	11.0	5.1	23.0	8.5
Agricultural 1	7.7	4,557	5.4	6.3	18.9	25.1	6.0
Agricultural 2	1.5	14,619	3.2	2.7	11.7	18.7	6.3
Agricultural 3	1.3	19,252	3.6	3.5	13.2	20.8	6.7
Agricultural 4	1.3	25,523	4.4	4.2	12.6	20.6	7.4
Agricultural 5	5.3	66,266	44.0	41.8	14.9	20.5	8.3
Total	100.0	16,904	100.0	100.0	6.4	21.6	6.5
Q5/Q1 agr	0.1	6.1	1.1	1.0	1.8	0.9	1.4
Q5/Q1 non-agr	0.7	14.5	8.1	6.6	0.8	0.8	1.4

lower quintiles, agricultural households show a lower per capita income than non-agricultural ones. Conversely, in the higher one, agricultural households show an average per capita income higher than non-agricultural.

As expected agricultural households earn the largest part of net income from farming (about 60 % of total). The share of the richest among agricultural households is over 44 % of total: a figure that should be read together with their small number (5.3 %). Overall, the 10 % of families included in the higher quintile (agricultural and non-agricultural) earn more than 50 % of total income from farming.

Another good example of the potential utility of the new dataset is the analysis of the distribution of support from sector policy among different household groups. The Single Farm Payment (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. The percentage of SFP accruing to each household group is shown in the fourth column. The figures reveal the existence of a distributive bias: the 11 % of agricultural households included in the highest quintile gather more than 50 % of SFP; furthermore for the richest agricultural households about 15 % of total income is represented by SFP. The support contributes to create about 20 % of farm incomes but with some interesting differences among household groups revealing an imperfect targeting of the measure.

The last column shows the average value of a composite well-being indicator including beside income level also information on housing conditions, education, health status and social exclusion.[6] The index is based on new information from the EU-SILC survey assigned to observations included in the FBS sample through the matching procedure. The availability of well-being indicators may represent a powerful tool in enhancing the targeting of agricultural policy. The index shows, as expected, a value increasing with income level; interestingly, the largest part of support from policy accrues to a small group of households with a well-being index well above the average in the total population.

References

D'Orazio, M., Di Zio, M., Scanu, M.: Statistical Matching. Theory and Practice. Wiley, Chichester (2006)

EU: Regulation Concerning Community Statistics on Income and Living Conditions (EU-SILC), N. 1177, Brussels (2003)

ISTAT: La distribuzione del reddito in Italia. Argomenti n. 38, ISTAT, Roma (2010)

ISTAT: I risultati economici delle aziende agricole. Anno 2008. Statistiche in breve, ISTAT, Roma (2011)

OECD: Handbook on constructing composite indicators. Methodology and user guide. Paris (2008)

[6]The index is the geometric average of a set of class variables; the aggregation through geometric averaging expresses a partial substitutability among different dimensions of well being (OECD 2008).

Pizzoli, E.: Redditi nelle aziende agricole a conduzione familiare. In: Approfondimento, Rapporto annuale sulla situazione del Paese nel 2004, ISTAT, Roma (2005)

Rassler, S.: Statistical Matching : A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. Springer, New York (2002)

Rocchi, B.: 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, ISTAT, Roma, pp. 521–528 (2010)

Sacco, G.: SAMWIN: a software for statistical matching. Manual, European Centres and Networks of Excellence (CENEX) – Integration of Surveys and Administrative Data (ISAD), Rome (2008)

UNECE, Eurostat, FAO, OECD, World Bank: Rural households' livelihood and well-being. Statistics on rural development and agricultural household income. United Nations, New York (2007)