

# Combining household data on income and expenditure from sample surveys and National accounts

Alessandra Coli

University of Pisa, Department of Applied Statistics and Mathematics,  
e-mail: [a.coli@ec.unipi.it](mailto:a.coli@ec.unipi.it)

Preliminary version - Please do not quote

## Abstract

Analyzing the economic behavior of the households requires data on households budgets at a very detailed level of disaggregation. It is unusual to have a single data source containing detailed information both on household expenditures and incomes and this problem is generally overcome with the “fusion” of independent data sources. In Italy, the Household Budget Survey (Hbs) collects detailed information on consumption. On the other hand the European survey on income and living conditions (Eusilc) and the Bank of Italy survey on income and wealth (Shiw) provide detailed information on income and saving but not on consumption. On the macro side, National accounts provide the totals of income and consumption expenditure for the Households sector as a whole.

The aim of this research is to combine data from the household budget surveys and National accounts in order to estimate consumption propensities by specific groups of households (region of residence, number of employees, number of children etc).

As a first step Shiw/Eusilc and Hbs datasets are fused through a statistical matching in order to obtain a complete and consistent dataset on household economic behavior.

As it is well known, household surveys underestimate income and expenditure, with actually a stronger underestimation for income. As a consequence, the simple matching of the surveys datasets might lead to inconsistent data, with consumption propensities exceeding one for most households. For this reason part of the paper is devoted to the realignment of surveys income and consumption to the National accounts corresponding aggregates.

Keywords: micro-macro reconciliation, households consumption propensities, statistical matching.

## 1. Introduction

In statistically developed countries information on households economic behavior is provided by several data sources. Two main categories may be distinguished. On the one hand National accounts (NA) describe the economic performance of Household from a macro perspective allowing economists to understand relationships between income, consumption and saving within a consistent and integrated framework. On the other, sample surveys on households provide insight on the economic behavior of single families. As it is well known economic micro and macro estimates do not

always combine properly. For this reason information in this field is at the moment fragmentary and incomplete.

Households consumption (or saving) propensity is one of the crucial economic issue which cannot be properly investigated on the basis of available data.

NA allow to calculate propensity to consumption of a representative household without providing further details about differences in consumption choices due to social, demographic or economic characteristics of the household. This limit is counterbalanced by the international comparability of NA consumption propensities.

Households sample surveys provide insights on the specific issues they investigate. In the European scenario usually countries may count on surveys on consumption expenditures and surveys on income. Frequently surveys on consumption collect information also on income and surveys on income contain few general questions on consumption expenditures. Households propensities to consumption are therefore calculated separately on the basis of each survey dataset. Two problems typically may occur:

- there are as many different estimates of consumption propensities as the number of surveys we dispose of;
- the average consumption propensities calculated on the basis of surveys (micro-approach) differ from consumption propensity derived from NA (macro-approach) to a noteworthy extent

The aim of this research is to provide a method to reconcile micro and macro estimates on income and consumption within the boundaries of NA, in order to obtain consumption propensities by groups of households and kinds of consumption. NA disposable income and consumption expenditure are considered the top macro aggregates to be further subdivided according to patterns coming from the surveys microdata sets once these have been fused in an inner consistent dataset. An application is described for the Italian households, for the year 2004.

Section 2 gives a quick view on household income and consumption statistics in Italy. Data are taken from the following data sources: National Accounts, the Bank of Italy Survey on Household Income and Wealth (SHIW), the ISTAT Survey on Households Budgets (HBS), the European survey on income and living conditions (EU-SILC). Comparisons point out how compelling is the need of a more complete and consistent information.

Section 3 is devoted to the merging of sample surveys datasets on household budgets. Particularly a statistical matching is used to merge the SHIW and the HBS for the years 2002 and 2004. For 2004 also a preliminary statistical matching between the EU-SILC and HBS is presented.

The “matched dataset” provides estimates of income and consumption expenditure by groups of households. Such information is used to disaggregate NA disposable income and consumption by groups of households. Results are shown in section 4.

The estimate of income and consumption of groups of households within the National accounts system is a first step towards the compilation of a complete set of accounts by household subsectors. This should allow a better understanding of households’ economic behavior and better comprehension of the interrelations between economic and social aspects. As a matter of fact the European System of Accounts (ESA95) already suggests the compilation of a complete set of national accounts by household

categories, but, to our knowledge, most of the European national statistical offices do not compile them on a regular basis<sup>1</sup>.

## **2. Statistics on households income and consumption in Italy**

National accounts provide a complete set of accounts for the Households' sector as a whole<sup>2</sup>. This allows to analyze the economic role of households in each step of the economic process, from production, to saving. Consistency among the several interrelated economic aggregates is granted by the NA building process.

The Italian statistical office (Istat) provide further details into the Households' sector analysis. On the basis of the main economic function criterion, two sub-sectors are distinguished: the Consumer households (whose main function is consumption) and the Producer households (entrepreneurs whose economic behavior cannot be separated from the economic behavior of the household they belong to). For each category a complete set of accounts is compiled yearly. Occasionally Istat provides also an interesting set of accounts for the households classified according to the region of residence (NUTS 2). Unfortunately this set is not complete since the sequence of accounts breaks off with the Secondary distribution of income account, i.e. with the estimate of households' disposable income by region. Finally, in the context of National accounts estimates of domestic consumption expenditure by region (NUTS 2) and function (consumption by purpose) is provided.

The evidence stemming from these data sources is not always coherent even once the definitions of variables have been harmonized<sup>3</sup>. In 2004 consumption expenditure per household was around 35000 euros according to NA and around 29000 according to the HBS survey. Differences are even more striking for income. In 2004, income per household was on average around 29000 euros according to SHIW, 33000 according to ESUSILC and around 43000 in National accounts. Inconsistencies do not affect only the variable levels. Territorial distribution both of consumption and income is captured differently<sup>4</sup> as shown in table 1 and table 2.

Inconsistencies are not only between macro and micro data. The Households' sample surveys do not always provide unambiguous information on common monetary variables once these have been harmonized in definitions and classifications. For example the level and distribution of income may be significantly different, income from one source may not be economically coherent with consumption expenditure coming from the other source and vice-versa. Table 3 shows an example of the possible incoherent outcomes that can be obtained by using SHIW and HBS without any previous matching process.

---

<sup>1</sup> National accounts have traditionally given relevance to the analysis of productive processes and final uses. On the contrary the information on institutional sectors has not been satisfying for a long time. With the publication of ESA95, the relevance of institutional sectors has sensibly increased. As far as the Households' sector is concerned, ESA95 suggests to consider six sub-sectors identified according to the largest income category of the household as a whole. The objective is to point out different economic behaviors inside the Households sector exactly as the subdivision by industries is aimed at stressing different production processes inside the general group of productive units.

<sup>2</sup> Istat website <http://www.istat.it/conti/nazionali/>

<sup>3</sup> For details see the companion paper Coli, Tartamella (2008)

<sup>4</sup> NA data refer to domestic consumption instead of national consumption. Nevertheless this conceptual difference cannot account for the territorial discrepancies in the distribution, which are macroscopic for not weighted frequencies.

Table 1 Distribution of Households' consumption by region of residence -Italy, 2004

	North-West	North-East	Centre	South	Italy
National accounts (domestic expenditure)	29.64%	21.73%	21.10%	27.53%	100.00%
SHIW	26.01%	23.11%	24.23%	26.64%	100.00%
SHIW-Weighted values	30.20%	21.80%	22.95%	25.05%	100.00%
HBS	26.56%	22.58%	19.07%	31.79%	100.00%
HBS- Weighted values	32.07%	21.92%	19.68%	26.32%	100.00%

Table 2 Distribution of Households' income by region of residence - Italy, 2004

	North-West	North-East	Centre	South	Italy
National accounts	31.43%	21.57%	20.68%	26.28%	100.00%
SHIW	26.86%	24.49%	23.72%	24.92%	100.00%
SHIW-Weighted values	31.19%	23.25%	22.68%	22.88%	100.00%
EUSILC	24.60%	26.57%	25.15%	23.68%	100.00%
EUSILC- Weighted values	30.03%	21.00%	20.95%	28.01%	100.00%

Table 3 Consumption propensities by household regional area. Year 2004

	Consumption propensities (SHIW data)	Consumption propensities (HBS expenditure distribution on SHIW income distribution )
North-west	0.84	0.87
North-east	0.82	0.80
Centre	0.89	0,68
South	0.93	1.13

### 3 The matching of the sample surveys data sets

In this section we apply statistical matching in order to merge the micro data on households' income (SHIW and EUSILC) with the micro data concerning households' consumption (HBS). Our purpose is mainly to obtain a matched data set where the economic variables are coherent by groups of households (subsectors) rather than at the single household level. Statistical matching seems particular attractive with respect to other integration techniques (linear regression modelling for example). Once income and consumption data sources have been matched, we can rely on a coherent and inner consistent set of information for splitting up most of NA variables (not only disposable income and total consumption expenditure) by households groups. Furthermore the matched file allows to group households according to a very large spectrum of characteristics.

Statistical matching is used to link independent samples of data, A and B, by means of some variables common to both data files. Some variables Y appear only in A whereas some variables X appear only in B. A set of variables Z can be observed in both samples. Since usually a sample recording Y, Z and X at the same time does not

exist, it is necessary to generate an artificial data set where each unit records Z, Y and X values. Various methods can be used to match A and B. We refer here to the matching technique that uses the nearest neighbour matches.

According to this method statistical matching can be regarded as an imputation problem. Let us consider sample A as an incomplete data set where X variables are missing. A is then defined as the *recipient* sample. For every unit  $a_i$ , with  $i = (1, 2, \dots, n_A)$ , one  $x$  value from the observations of the donor sample B is selected. The donor unit  $b_j$  with  $j = (1, 2, \dots, n_B)$  is searched among the units belonging to B which present Z values ideally identical to those of the recipient unit  $a_i$ . The perfect match in terms of the common variables is not always possible. When this occurs the donor unit is selected on the basis of a distance measure  $d(Z)$ . The donor unit is the unit with the smallest distance. When more donors are identified a random selection is performed.

The application of traditional statistical matching implies the so called Conditional independence between Y and X given Z (see especially Rodgers 1984)). Conditional independence is produced for the variables not jointly observed even when such variables are conditionally dependent in reality. The problem is that often the relationship between the never jointly observed variables is unknown.

The Conditional independence Assumption (CIA) is a strong limit to the application of traditional statistical matching. Sceptics assert that statistical matching does not bring any additional information on the relationship between the not jointly observed variables: the outcome is already well known. The advocates argue that statistical matching is the only practical solution when the merging of data sets with hundreds of variables is necessary (see for example Ruggles 1974). According to this viewpoint, CIA can be roughly satisfied by carefully selecting the common variables (see Rässler 2002 for the debate on the pros and cons of statistical matching). For this reason it is essential to identify common variables significantly connected both with Y and Z variables.

### 3.1 The matching of SHIW and HBS data sets

In the following we apply statistical matching to SHIW and HBS for the years 2002 and 20045.

As a first step, surveys have been harmonized in order to make the data comparable. This is a very time-consuming and difficult step. Inconsistencies must be solved through recoding of variables, imposing assumptions etc. It is necessary to carry out harmonization with care since changes on original data sets have relevant effects on the entire matching procedure (see Ruggles 1974).

The head of the households characteristics have not been considered among the potential matching variables due to the different definitions of the head of the household (or reference person) in the three surveys.

Some common variables cannot be used in the matching process even though their categories have been perfectly harmonized. This happens when the distributions of such variables are significantly different in the data sets, as if samples were extracted from different populations. Unfortunately this happens also for variables which it

---

<sup>5</sup> Details on this application are in the companion paper Coli, Tartamella (2008). The method was developed within the research studies for the building of a Social accounting matrix. In this context Istat organised a working group aimed at integrating SHIW and HBS data sets. Full results are given in Coli et al. (2006).

would be very desirable to include among the matching ones, given their strict link with income and consumption.

Particularly, we notice relevant differences in the distribution of households by number of retired members (pens) and by number of members neither employed nor retired (naltrac). Therefore we have decided to exclude these two variables from the set of the potential matching variables.

The second step consists in selecting the matching variables, i.e the common variables most strictly connected to household income and consumption. In order to satisfy the Conditional Independence Assumption as much as possible, it is essential to select variables that are strictly connected both with income and consumption expenditure. If we exclude monetary variables we have to resort to the demographic and social variables. It has already been pointed out that the HBS survey collects data on household income and saving. Particularly, it includes a categorical variable on the total monthly entries of the household. This variable underreports income but in our opinion it supplies a good piece of information on the rank of the households in the income cumulative distribution. For 2002 we have defined both for HBS and SHIW the new variable TM which classifies households according to their relative position in terms of income. Eight categories have been established, from the poorest household (TM=1) to the richest (TM=8). Each category contains approximately the same number of households. This variable gives a relevant contribution to the matching process. Unfortunately it is not possible to define TM for 2004, since from 2003 onwards ISTAT does not supply any information on income and saving through the HBS. As an alternative to the TM variable we have used the quintile of food consumption expenditure. Again we assume that both SHIW and HBS correctly collect the rank of the household in a range from lowest consumption to the highest. . On the basis of several analysis<sup>6</sup> ( like (Cramer 's V coefficient, the Analysis of variance, multiple regression models) the following variables have been selected as the most strictly connected to households income and consumption.

TM: income class (only for 2002)

Qalim: quintile of food consumption expenditure

Ncomp: numbers of members

Nocc: numbers of members with a job

Ndip: number of employees

Ndiploma: number of members with 11-13 years' schooling

Nlaurea: Number of members with a university degree

Nadul: number of members aged 40-64

Tipoanz: presence of at least one member aged  $\geq 75$

Area: geographical area of resident

Tbtr: house renting

#### Performing statistical matching

The SHIW microdata set is the recipient sample whereas the HBS microdata set is defined as the donor sample. The size of the SHIW sample is about 8000. Since the HBS sample is on average three times as large, more than one donor unit can be selected for each SHIW unit.

In the application of the nearest neighbour distance matching, different combinations of the matching variables can be chosen. As a first step SHIW and HBS sample units are classified on the basis of the matching variables with the strongest relationship

---

<sup>6</sup> Various techniques can be used (see D'Orazio et al 2007)

with income and consumption expenditure, the so called strata variables. The SHIW and HBS units are grouped into subsets (strata) whose units share the same exact value of the strata variables.

We have considered the following households stratifications of SHIW and HBS samples:

Year 2002

- Income categories <sup>TM</sup> by Number of members (Ncomp): 40 strata
- Quantile of consumption expenditure (Qalim) by Number of members (Ncomp): 25 strata

Year 2004

- Quantile of consumption expenditure (Qalim) by Number of members (Ncomp): 25 strata
- Quantile of consumption expenditure (Qalim) by Number of members with a job (Nocc): 15 strata

The nearest neighbor of each SHIW unit is searched among the HBS units belonging to the same stratum. As a first step the algorithm<sup>7</sup> identifies in each stratum the HBS units with the lowest distance from the SHIW recipient units where the distance is a function of the values assumed by the matching variables. Whenever more than one donor is identified, a random selection is run.

Different combinations of the matching variables can be chosen in order to calculate distance. The following tables describe the performed combinations. The first column records the name of the data set which results from the matching process.

Tab 4 The combination of matching variables – SHIW-HBS matching- 2002

Matched files	Maching variables	
	Strata variables	Distance matching variables
QNC	qalim,ncomp	nocc,ndiploma,nlaurea,ndip,nadul,tipoanz,tabt,area
TMNC	TM,ncomp	nocc,ndiploma,nlaurea,ndip,nadul,tipoanz,tabt,area

Table 5 The combination of matching variables – SHIW-HBS matching- 2004

Matched files	Maching variables	
	Strata variables	Distance matching variables
QNC1	qalim,ncomp	nocc,ndiploma,nlaurea,ndip,nadul,tipoanz,tabt,area
QNC2	qalim,ncomp	nocc,ndiploma,nlaurea,ndip,nadul
QNC3	qalim,ncomp	nocc,nadul,ndiploma, ndip
QNO1	qalim,nocc	ncomp,ndiploma,nlaurea,ndip,nadul,tipoanz,tabt,area
QNO2	qalim,nocc	ncomp,ndiploma,nlaurea,ndip,nadul
QNO3	qalim,nocc	ncomp,nadul,ndiploma, ndip

<sup>7</sup> We use a software developed by Giuseppe Sacco (Istat); see Coli et al. 2006 for details.

The statistical matching procedure generates matched files which have the same dimension as the SHIW (recipient dataset). In order to choose the best matched file it is customary to compare the distributions of imputed variable in the donor and in the matched data set (see Rässler 2002, D’Orazio et al. 2006). As our main objective is to impute HBS consumption expenditure, we compare summary statistics on total consumption expenditure calculated with HBS data and with each matched file data. The summary characteristics of total consumption expenditure are quite well preserved in the 2002 matched files. The best result is obtained by the TMNC file (weighted values). In this file households are stratified by income category <sup>TM</sup> and by number of members variable (ncomp).

Estimated correlations between imputed consumption ( $\tilde{C}$ ) and observed SHIW income (Y) show clearly the effects of the Conditional independence assumption. Values are quite low if compared to the SHIW inner correlation between income and consumption which is around 0.60 in SHIW. The highest correlation is recorded for the TMNC file, followed by QNC, QNC1 e QNO1.

Table 6 Comparisons between consumption expenditure statistics computed on each matched file data and on the HBS data (HBS statistic=100) – year 2002

Summary statistics	Matched data sets			
	QNC	TMNC	QNC	TMNC
	Unweighted values		Weighted values	
$\mu$	99.02	98.34	99.55	100.20
$\sigma$	101.94	96.53	104.73	101.30

Table 7 Comparisons between consumption expenditure statistics computed on each matched file data and on HBS data (HBS statistic=100) – year 2004

Unweighted values						
	QNC1	QNC2	QNC3	QNO1	QNO2	QNO3
$\mu$	102.55	102.72	104.06	103.18	104.39	103.66
$\sigma$	91.52	94.19	93.81	92.85	96.44	92.54
Weighted values						
	QNC1	QNC2	QNC3	QNO1	QNO2	QNO3
$\mu$	106.95	105.79	108.58	108.57	106.83	107.30
$\sigma$	101.88	103.31	103.26	108.05	106.17	99.68

Substituting the income category (TM) for the quintile of food consumption (qalim) clearly leads to an increase of the correlation between imputed consumption and SHIW income.

Table 8 Correlations between imputed consumption ( $\tilde{C}$ ) and SHIW income (Y) - 2002

	Matched data sets	
	QNC	TMNC
${}_Y \rho_{\tilde{C}}$	0.329	0.390

Table 9 Correlations between imputed consumption ( $\tilde{C}$ ) and SHIW income (Y)

	QNC1	QNC2	QNC3	QNO1	QNO2	QNO3
${}_Y \rho_{\tilde{C}}$	0.308	0.273	0.255	0.310	0.296	0.267

### 3.2 The merging of EU-SILC and HBS data sets

Matching variables have been selected only among the social and demographic characteristics of the household, since there are not comparable monetary variables in the surveys. EU-SILC does not provide any information on households total or food consumption expenditure and HBS does not supply any information on income. We made two trials: in the first households are stratified by geographic area (Area) and number of members (Ncomp), in the second, households are stratified by number of members (Ncomp) and number of members with a job (Nocc). The trials give similar results both in terms of preservation of the HBS consumption summary statistics and in terms of correlation between imputed consumption and EU-SILC income. The correlation index is around 0.23 for both the matched. Results are undoubtedly worse if compared to the SHIW-HBS matching.

Table 10 Comparisons between consumption expenditure statistics computed on each matched file data and on HBS data (HBS statistic=100) – year 2004

Summary statistics	Matched data sets			
	ARNC		NONC	
	Unweighted	Weighted	Unweighted	Weighted
$\mu$	113.86	114.56	115.32	115.55
$\sigma$	101.6	105.61	107.14	109.01

In order to fulfil CIA assumption as much as possible, it is essential to consider matching variables capable of explaining most of income and consumption variability. Given the available data, we have to rely on rather weakly explaining variables. The possibility of including variables such as the quintile of food consumption or the income categories considered for the SHIW – HBS matching would increase the quality of the matched file.

#### 4. Estimate of NA disposable income and consumption expenditure by Households' subsectors

In this section we provide estimates of 2004 NA disposable income and consumption expenditure by households subgroups. NA variables are subdivided according to indicators derived in the SHIW-HBS matched file (QNC1 file). In order to validate our estimates from an economic point of view we have compared consumption propensities, computed on different kinds of income and consumption. Some results are given in table 11 where three kinds of consumption propensities (CP) are compared. Columns names indicate the data source which provides the indicators to calculate computation propensities. Consumption expenditure and income are grossed up to the NA values. Propensities are computed as follows:

$$CP_{(QNC1)} = \text{imputed consumption} / \text{SHIW income}$$

$$CP_{(SHIW)} = \text{SHIW consumption} / \text{SHIW income}$$

$$CP_{(HBS-SHIW)} = \text{HBS consumption} / \text{SHIW income}$$

Table 11 Consumption propensities for groups of households Italy, 2004

	QNC1	SHIW	HBS-SHIW
<b>area</b>			
1	89.0	84.4	86.9
2	86.2	82.3	79.9
3	80.3	89.0	68.2
4	92.5	93.2	112.6
<b>nbam</b>			
0	83.5	86.1	82.8
1	98.8	91.7	107.7
2	99.2	86.9	94.0
3	105.6	100.4	84.5
<b>npens</b>			
0	91.9	88.6	104.5
1	82.7	87.9	70.0
2	79.7	80.9	61.9
<b>ndip</b>			
0	81.0	89.3	76.9
1	90.9	87.2	97.2
2	92.0	83.8	90.5

For most of the considered households subgroups QNC1 consumption propensities take more realistic values with respect to propensities calculated by using SHIW and HBS data without any previous matching process. Besides, propensities computed on the basis of the matched file values are more in line with those calculated with SHIW values.

## References

- Brandolini, A., 1999. The Distribution of Personal Income in Post-War Italy: Source Description, Data Quality, and the Time Pattern of Income Inequality, Banca Italia - Servizio di Studi.
- Brandolini A., Cannari L., D'Alessio G., Faiella I. (2004), Household Wealth Distribution in Italy in the 1990s, Banca d'Italia, Temi di discussione n. 530.
- Brandolini A., Cannari L., D'Alessio G., Faiella I. (2004), Household Wealth Distribution in Italy in the 1990s, Banca d'Italia, Temi di discussione n. 530.
- Coli A., S. Colombini, M. Di Zio, M. D'Orazio, I. Faiella, I. Siciliani, G. Sacco, M. Scanu, F. Tartamella (2006), La costruzione di un Archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi delle famiglie italiane e l'Indagine Banca d'Italia sui bilanci delle famiglie italiane, Istat, Contributi.
- Coli A., F. Tartamella (2008) Income and consumption expenditure by households groups in National accounts –IARIW 30th General Conference - working paper -
- D'Orazio M., M. Di Zio, M. Scanu (2005) Statistical Matching and the Likelihood principle: Uncertainty and logical Constraints, Istat, Contributi.
- D'Orazio M., M. Di Zio, M. Scanu (2006), Statistical matching: Theory and Practice, Wiley & Sons, New York.
- EUROSTAT(2002), Handbook on Social accounting matrices and labour accounts
- Filippello R., U. Guarnera, Jona Lasino G., Use of Auxiliary Information in Statistical Matching, SIS, Riunione Scientifica - Bari 2004 – Comunicazioni spontanee.
- Gower, J. C. (1971) A general coefficient of similarity and some of its properties. *Biometrika*
- Istat (2000) Le nuove stime dei consumi finali delle famiglie secondo il SEC 95, Metodi e norme n. 7)
- Istat (2005) I conti economici nazionali per settore istituzionale:le nuove stime secondo il Sec95, Metodi e norme n.23
- Istat (2007) Conti economici nazionali per settore istituzionale, Statistiche in breve
- Istat (2008) La misura dell'economia sommersa nelle statistiche ufficiali, Statistiche in breve
- Kadane Joseph B. (2001) Some statistical problems in merging data files – *Journal of official statistics*, vol. 17, No. 3. Reprint (1978).
- Moriarity C., Sheuren Fritz (2001) Statistical matching: a paradigm for assessing the uncertainty in the procedure – *Journal of official statistics*, vol. 17, No. 3.
- Prem K. Goel, T. Ramalingam (1989), The matching methodology: some statistical properties, *Lecture notes in statistics*, 52 –Springer, New York.
- Rodger W. L. (1984) An evaluation of statistical matching – *Journal of business and economic statistics*, vol.2, no. 1.
- Rosati Nicoletta (1998) Matching statistico tra I dati Istat sui consumi e dati bankitalia sui redditi per il 1995. Rapporto tecnico.
- Richard Ruggles, Nancy D. Ruggles (1974) A strategy for merging and matching microdata sets, *Annals of economic and social measurement*, 3, 1974.
- Susanne Rässler (2002), Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches, *Lecture notes in statistics*, 168 – New York Springer.
- Sutherland Holly, Rebecca Taylor and Joanna Gomulka (2001) Combining household income and expenditure data in policy simulations, Working paper Department of Applied economics, University of Cambridge.