

NTTS 2009, Brussels 18-20 February 2009

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

Alessandra Coli

Department of Statistics and Mathematics Applied to Economics,
University of Pisa

e-mail: a.coli@ec.unipi.it

* Work supported by the project SAMPLE "Small Area Methodology for Poverty and Living Condition Estimates" awarded by the European Commission in the 7thFP

Objectives of the research

The aim of this research is to provide a method to reconcile micro and macro estimates on household income and consumption within the boundaries of National accounts (NA).

Main outcome: the estimate of propensity to consume by groups of households and categories of consumption.

Main indirect advantages:

- Improved international comparability of household income and consumption
- Better reconciliation of micro and macro data.
- Integration of micro data on households budgets from independent data sources

Data on household income and consumption

the state-of-the-art

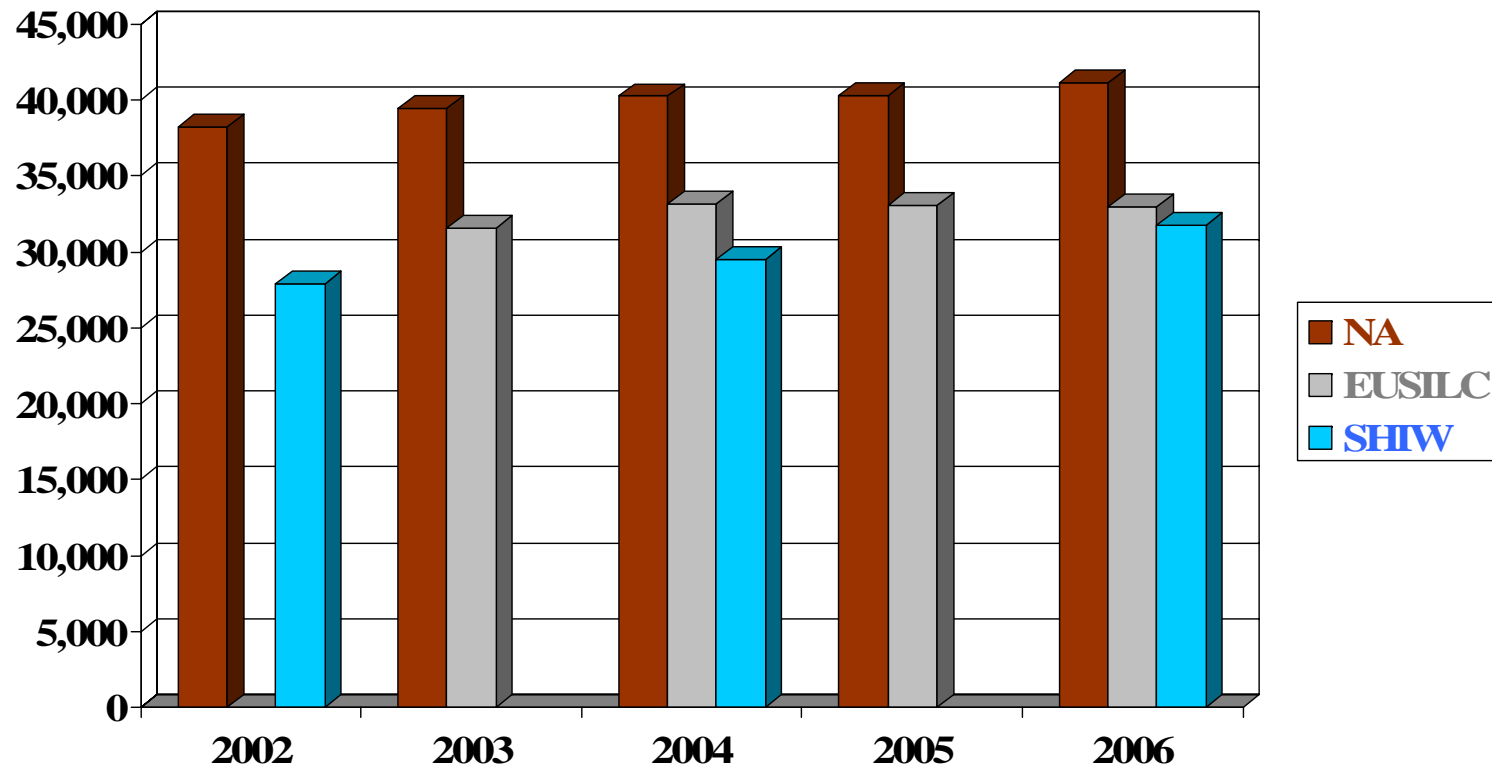
In statistically advanced countries, information on the economic behavior of households is provided by several data sources. Two main categories may be distinguished

National accounts (NA): NA describe the economic performance of households (the Households sector) from a *macro* perspective, showing the relationships between income, consumption and saving within a consistent and integrated framework.

Sample surveys: surveys provide insight on the economic behavior of single families (*micro* perspective). Frequently, surveys on consumption collect information also on income and surveys on income contain few general questions on consumption expenditures. It is unusual to have surveys with detailed and reliable enough information both on income and consumption.

Evidence from available data sources in Italy

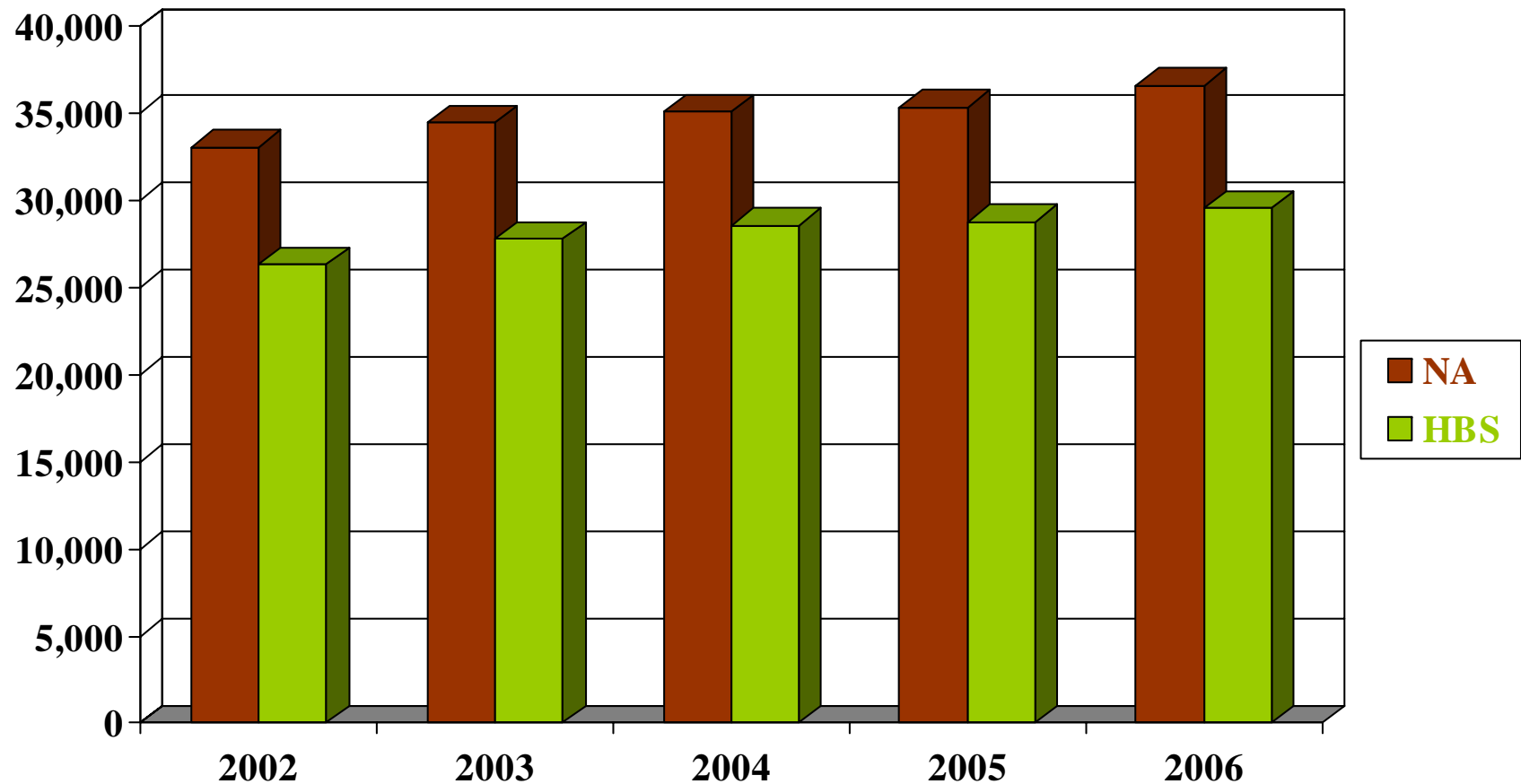
Household income (euro)



NA: National Accounts (Istat); EUSILC: European survey on income and living conditions (Istat); **SHIW: Survey on households income and wealth (Bank of Italy)**

Evidence from available data sources in Italy

Household consumption expenditure (euro)



NA: National Accounts (Istat); HBS: Household budget survey (Istat)

Average propensity to consume

From a statistical point of view propensity to consume reflects how income statistics relate to consumption statistics.

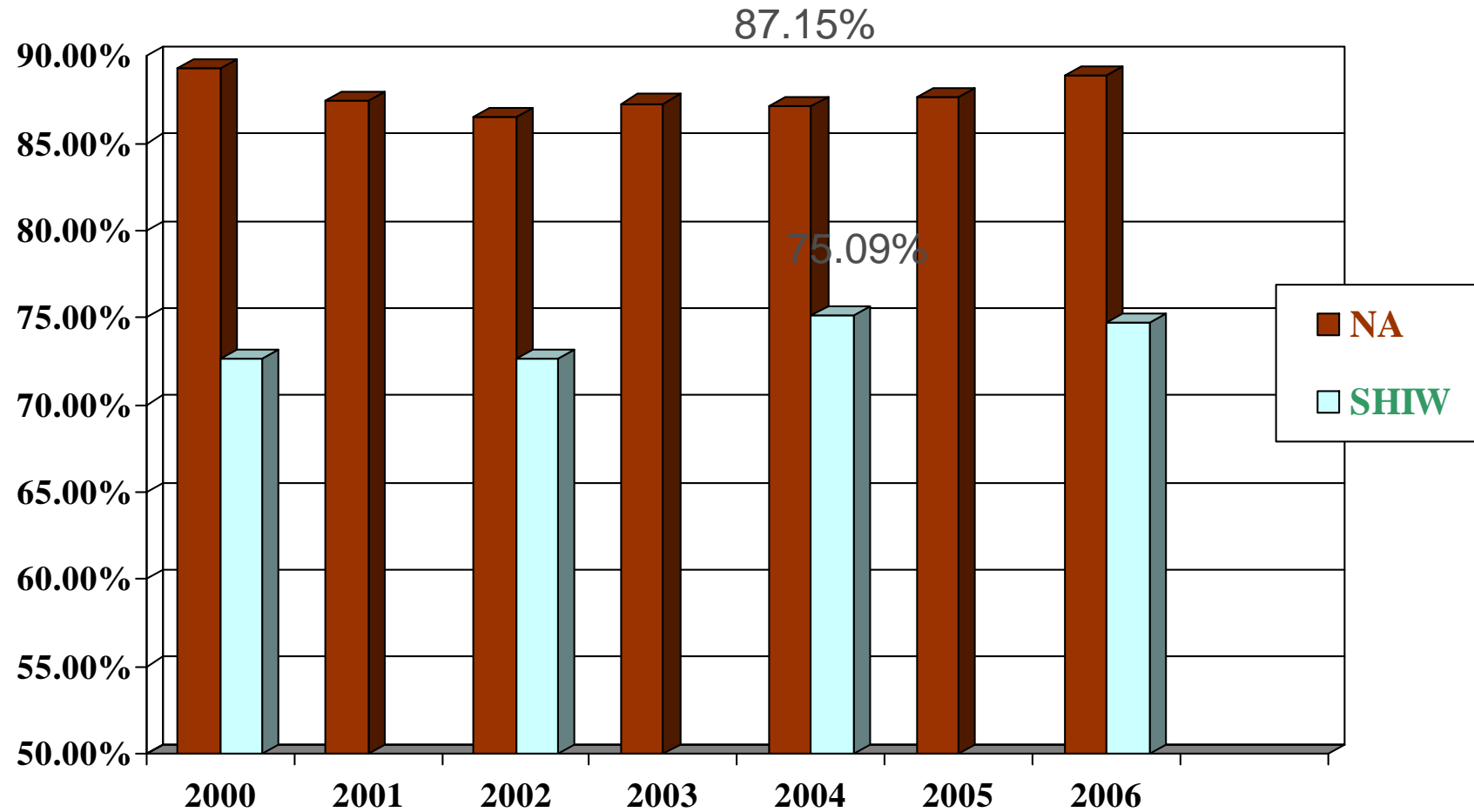
Consumption propensities are calculated on the basis of NA and sample surveys data (provided the surveys collect data both on income and expenditure).

Typically two problems may occur:

1. there are as many different estimates of consumption propensities as the number of surveys
2. the average consumption propensities calculated on the basis of surveys (micro-approach) largely differ from consumption propensity derived from NA (macro-approach).

Evidence from available data sources in Italy

Average propensity to consume



Follows

Furthermore:

- National accounts allow to break down the propensity to consume by consumption categories (food and beverages, clothing and footwear, transport, health,... ..) but only for the Household sector as a whole.
- SHIW allows to calculate consumption propensities of groups of households but not by consumption categories (only few macro categories are distinguished)
- In Italy, from 2003 onwards, the HBS does not provide any insight on income and as a consequence on consumption propensities.
- EUSILC does not collect data on consumption expenditure.

This work tries to estimate the following table

Groups of households	Y	Categories of consumption					
		C ₁	C ₂	C _j	..	C _m
h ₁	y ₁	c ₁₁	c ₁₂	c _{1j}	..	c _{1m}
.							
.							
h _i	y _i	c _{i1}	c _{i2}	c _{ij}	..	c _{im}
.							
h _n	y _n	c _{n1}	c _{n2}	...	c _{nj}		c _{nm}

y= household income

c= household consumption expenditure

Constraint: Row sums correspond to the NA aggregates

Follows...

In order to fill in the table, NA macro variables (row sums) are disaggregated according to proper indicators derived from survey micro data

Main issues

1) How to reconcile macro and micro data on income and consumption.

2) How to match income and consumption statistics at a micro level.

Beside inconsistencies between macro and micro data: sample surveys do not always provide unambiguous information on common monetary variables once these have been harmonized in definitions and classifications.

Propensity to consume – Italy 2004

	Consumption propensities (SHIW)	Consumption propensities (HBS - SHIW)
North-west	0.84	0.87
North-east	0.82	0.80
Centre	0.89	0.68
South	0.93	1.13

In this work we apply *statistical matching* in order to match income data (from SHIW and EUSILC) and consumption data (from HBS).

Statistical matching

- Statistical matching is a data integration procedure used to link independent samples of data, **A** and **B**, by means of some **variables common** to both data files.
- The samples are extracted from the same population. The units observed in the data sets are different.
- Some variables **Y** appear only in **A** whereas some variables **Z** appear only in **B**. A set of variables **X** can be observed in both samples.
- The objective is the construction of a synthetic data set **A∪B** which contains all the variables of interest (**Y,X,Z**)

Statistical matching can be regarded as a method to overcome a problem of missing values

The framework of the synthetic data set $A \cup B$

Sample	Y_1	...	Y_q	...	Y_Q	X_1	...	X_p	...	X_P	Z_1	...	Z_r	...	Z_R	
A	Y_{11}^A		Y_{1q}^A		Y_{1Q}^A	X_{11}^A		X_{1p}^A		X_{1P}^A						
						
	Y_{a1}^A		Y_{aq}^A		Y_{aQ}^A	X_{a1}^A		X_{ap}^A		X_{aP}^A						
						
	Y_{nA1}^A		Y_{nAq}^A		Y_{nAQ}^A	X_{nA1}^A		X_{nAp}^A		X_{nAP}^A						
B						X_{11}^B		X_{1p}^B		X_{1P}^B	Z_{11}^B		Z_{1r}^B		Z_{1P}^B	
					
						X_{b1}^B		X_{bp}^B		X_{bP}^B	Z_{b1}^B		Z_{br}^B		Z_{bP}^B	
					
						X_{nB1}^B		X_{nBp}^B		X_{nBP}^B	Z_{nB1}^B		Z_{nBr}^B		Z_{nBP}^B	

D'Orazio et al. (2006)

The framework in our case

Sample	Income	Common variables				Expenditure by consumption categories						
	Y	X_1	..	X_p	...	X_P	Z_1	...	Z_r	...	Z_R	
SHIW EUSILC	Y^{SHIW}_1	X^{SHIW}_{1l}		X^{SHIW}_{lp}		X^{SHIW}_{lP}						
						
	Y^{SHIW}_i	X^{SHIW}_{il}		X^{SHIW}_{ip}		X^{SHIW}_{iP}						
						
	Y^{SHIW}_n	X^{SHIW}_{nl}		X^{SHIW}_{np}		X^{SHIW}_{nP}						
HBS			X^{HBS}_{1l}		X^{HBS}_{lp}		X^{HBS}_{lP}	Z^{HBS}_{1l}		Z^{HBS}_{lr}		Z^{HBS}_{lP}
		
			X^{HBS}_{jl}		X^{HBS}_{jp}		X^{HBS}_{jP}	Z^{HBS}_{jl}		Z^{HBS}_{jr}		Z^{HBS}_{jP}
		
			X^{HBS}_{ml}		X^{HBS}_{mp}		X^{HBS}_{mP}	Z^{HBS}_{ml}		Z^{HBS}_{nBr}		Z^{HBS}_{nBP}

Conditional Independence Assumption (CIA)

- The application of traditional statistical matching implies the so called Conditional independence assumption between Y and Z given X (see especially Rodgers 1984). Conditional independence is produced for the variables not jointly observed even when such variables are actually conditionally dependent.
- CIA represents a strong limit to the application of traditional statistical matching (see Rässler 2002 for the debate on the pros and cons of statistical matching).
- According to the advocates' viewpoint CIA can be roughly satisfied by carefully selecting the common variables used to match the data sets.

The integration of SHIW and HBS data sets (2004)

SHIW: RECIPIENT file

HBS: DONOR file

The missing items of each record in the SHIW (expenditure by consumption categories) are imputed using records from HBS.

The Nearest neighbour hot deck procedure

Each record in SHIW is matched with the closest record in the HBS, according to a distance measure computed using the *matching variables X* (D'Orazio et al., 2006). The donor unit is the unit with the smallest distance. When more donors are identified, a random selection is performed

Main steps of the matching procedure

- Harmonizing common variables
- Selecting the matching variables (the common variables most strictly connected to household income and consumption)
- Performing statistical matching
- Assessing the accuracy of the Statistical matching procedure

The matching variables

- TM: income class (1,2,...,8) - only for 2002
- Qalim: quintile of food consumption expenditure
- Ncomp: numbers of members (1,2,3,4,5+)
- Nocc: numbers of members with a job (0,1,2)
- Ndip: number of employees (0,1,2+)
- Ndiploa: number of members with 11-13 years' schooling (0,1,2+)
- Nlaurea: Number of members with a university degree (0,1,2+)
- Nadul: number of members aged 40-64 (0,1,2+)
- Tipoanz: presence of at least one member aged ≥ 75
- Area: geographical area of resident (North-west, North-east, Centre, South)
- Tbtr: dwelling (owner/tenant)

Performing statistical matching

The nearest neighbour matching can be performed by selecting different subsets of the matching variables.

SHIW-HBS, 2002

	Maching variables	
Matched files	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

Accuracy of the matching

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	
μ	99.02	98.34	99.55	100.20
σ	101.94	96.53	104.73	101.30

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

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

Summary statistics SHIW-HBS 2004

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

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

NA disposable income and consumption expenditure by Households' subsectors

- NA disposable income and expenditure consumption have been broken down by household categories, 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 calculated propensity to consume (CP) for several groups of household.
- For most of the considered households subgroups, consumption propensities take more realistic values with respect to propensities calculated by using SHIW and HBS data without any previous matching process.

	QNC1	SHIW	HBS-SHIW
area			
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
nbam			
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
ndip			
0	81.0	89.3	76.9
1	90.9	87.2	97.2
2	92.0	83.8	90.5
nlaurea			
0	90.3	89.8	90.3
1	81.4	80.1	79.2
2	68.0	75.5	72.7

CP(QNC1)=imputed
consumption / SHIW
income

CP (SHIW) =SHIW
consumption/SHIW
income

CP (HBS-SHIW) = HBS
consumption/SHIW
income

Concluding remarks

- The matching of micro datasets is an essential step in order to estimate NA income and consumption by groups of households.
- Statistical matching gives good results only if the Conditional Independence Assumption holds. In order to respect CIA it is essential to recover as much information as it is possible on the relationship between income and consumption (auxiliary information). Moreover the comparability of surveys needs to be improved.
- The use of income micro data is essential for estimating income by Households sub-sectors. The best would be probably to introduce income micro statistics in the integration process underlying the building of NA.