Some reflections on the comments and recommendations on the project "Data fusion of EU-SILC and HBS at ISTAT"

One of the main issues in the project can be summarized by the following question. Given the two samples EU-SILC, with detailed and accurate information on household income, and HBS, that accurately collects a diary on household consumption, and given that the two samples are independent and the presence of the same household in the two sample is an extremely rare event, is it possible to estimate the joint distribution of household income and expenditure?

The presence of a specific question on household income in HBS gave us the possibility to use the following trick: even if household income on HBS is not very accurate, and usually underestimated, it is a good source of information for the household ordering in terms of their income. In other words, all the households tend to give inaccurate figures on their income in HBS, but the higher income households tend to give higher figures then the lower income households. Hence, the household ordering on income as detected by HBS seems to be a reliable source of information. We claimed that it is also a very correlated source of information with respect to the actual household income. This fact suggested its use as a matching variable. Under this condition, the data fusion problem can be solved by a plug-in estimator. In fact, we claimed that income Y and expenditures Z are independent given the ordered income X. Hence, the joint distribution $f(y,z|x)$ can be decomposed as the product $f(y|x)f(z|x)$. The plug-in estimator estimates each factor respectively on EU-SILC and HBS.

As a matter of fact, the data fusion problem solved along the previous lines relied on an untestable assumption, the conditional independence of Y and Z given X.

As part of the revision of the EU-SILC within the new Framework Regulation on Social Statistics (IESS which will most likely be in force from 2021), Italy implemented the ESS Agreement by testing the rolling module on Consumption & Wealth (C&W) in EU-SILC 2017.

The module collected five consumption target variables:

- Food at home
- Food outside home
- Public Transport
- Private Transport
- Regular Savings

Italy decided to continue to collect the most relevant variables of the Consumption & Wealth module also in EU-SILC 2018 and 2019 to have consolidated and useful proxy variables for statistical matching purposes.

The variables collected in the C&W module, together with the already available housing costs, should represent a significant part of total consumption, such as to allow us to estimate a total consumption variable also in EU-SILC (in the Italian HBS, food, housing and transport costs correspond to 70% of the overall expenditures). Italy and some other countries already have the C&W 2017 data.

As claimed at the beginning, in order to match HBS and EU-SILC data, the information in the Italian HBS ad-hoc section on income and savings was used to estimate a synthetic income variable that was able to reduce the large discrepancy compared to SILC. This synthetic variable was then used as one of the matching variables for imputing consumption classes into SILC.

Then it is possible to construct in EU-SILC data an ordered income variable that has the same support (the 7 classes) and a similar definition than the HBS synthetic income.

With respect to these EU-SILC data, it is achievable to test if the independence assumption between income and expenditures given the ordered income class is an appropriate assumption

In this example, the matching variables are: geographical macro areas (in 5 classes), number of owned durable goods (5 categories), ordinal income classes (7 categories).

Geo. Ripart.	N. goods	Ordinal inc. cla.	Partial correlation	Number of obs.
1	4	1	0,02	217
1	4	2	0,12	201
1	4	3	0,22	120
1	4	4	0,14	60
1	4	5	0,15	34
1	4	6	0,30	15
1	4	7	1,00	2
1	5	1	-0,18	186
1	5	2	0,15	241
1	5	3	0,01	158
1	5	4	0,01	122
1	5	5	-0,10	92
1	5	6	-0,09	61
1	5	7	0,06	19
1	6	1	0,01	193
1	6	2	-0,01	244
1	6	3	0,06	287
1	6	4	-0,06	206
1	6	5	0,09	173
1	6	6	0,07	142
1	6	7	-0,07	51
1	7	1	-0,06	115
1	7	2	0,14	143
1	7	3	0,05	274
1	7	4	0,17	208
1	7	5	0,17	272
1	7	6	0,15	335
1	7	7	0,06	177
1	8	1	0,06	40
1	8	2	-0,12	55
1	8	3	-0,03	134
1	8	4	0,06	138
1	8	5	-0,05	250
1	8	6	0,16	400
1	8	7	0,11	349
2	4	1	-0,04	151
2	4	2	-0,11	148
2	4	3	-0,01	95
2	4	4	0,22	26
2	4	5	0,09	11
2	4	6	0,02	8

2	4	7	1,00	3
2	5	1	0,07	152
2	5	2	0,04	145
2	5	3	0,08	132
2	5	4	0,02	84
2	5	5	-0,14	71
2	5	6	0,14	38
2	5	7	-0,12	13
2	6	1	-0,04	147
2	6	2	0,15	217
2	6	3	0,05	283
2	6	4	0,03	169
2	6	5	0,06	186
2	6	6	0,04	142
2	6	7	0,30	57
2	7	1	-0,10	112
2	7	2	0,12	149
2	7	3	0,18	266
2	7	4	0,00	226
2	7	5	0,18	272
2	7	6	0,20	413
2	7	7	0,04	188
2	8	1	-0,31	30
2	8	2	-0,03	61
2	8	3	0,17	114
2	8	4	0,05	148
2	8	5	0,11	250
2	8	6	0,11	443
2	8	7	0,14	391
3	4	1	0,03	210
3	4	2	0,12	183
3	4	3	0,31	74
3	4	4	0,02	48
3	4	5	0,25	23
3	4	6	-0,58	12
3	4	7	0,85	3
3	5	1	-0,08	205
3	5	2	0,18	216
3	5	3	0,05	155
3	5	4	-0,01	75
3	5	5	0,24	46
3	5	6	-0,23	40
3	5	7	0,27	10
3	6	1	0,01	200
3	6	2	0,09	228
3	6	3	0,00	264
3	6	4	0,01	168

3	6	5	0,19	158
3	6	6	0,04	130
3	6	7	-0,06	38
3	7	1	0,03	148
3	7	2	0,07	174
3	7	3	0,09	266
3	7	4	0,05	223
3	7	5	0,08	322
3	7	6	0,11	338
3	7	7	0,01	149
3	8	1	-0,07	44
3	8	2	0,22	69
3	8	3	0,24	138
3	8	4	0,17	158
3	8	5	0,21	243
3	8	6	0,18	380
3	8	7	0,01	292
4	4	1	-0,14	383
4	4	2	0,32	202
4	4	3	0,02	113
4	4	4	0,26	50
4	4	5	-0,22	31
4	4	6	0,30	15
4	4	7	-1,00	2
4	5	1	-0,03	208
4	5	2	0,14	181
4	5	3	0,18	119
4	5	4	-0,06	84
4	5	5	0,26	46
4	5	6	0,01	33
4	5	7	1,00	2
4	6	1	0,02	196
4	6	2	0,09	211
4	6	3	0,09	214
4	6	4	0,03	136
4	6	5	-0,02	114
4	6	6	0,15	81
4	6	7	-0,12	22
4	7	1	0,02	123
4	7	2	-0,01	141
4	7	3	0,13	180
4	7	4	-0,09	192
4	7	5	0,05	149
4	7	6	-0,06	146
4	7	7	0,00	69
4	8	1	0,19	38
4	8	2	0,25	49

4	8	3	0,07	101
4	8	4	0,04	93
4	8	5	-0,08	124
4	8	6	0,22	176
4	8	7	0,03	101
5	4	1	0,07	107
5	4	2	-0,12	52
5	4	3	0,22	27
5	4	4	0,18	12
5	4	5	-0,92	5
5	4	6	0,53	4
5	4	7	NA	1
5	5	1	0,09	109
5	5	2	0,15	76
5	5	3	0,06	49
5	5	4	0,32	30
5	5	5	-0,23	15
5	5	6	-0,02	16
5	5	7	1,00	3
5	6	1	0,03	104
5	6	2	0,01	95
5	6	3	0,18	99
5	6	4	0,19	39
5	6	5	0,21	46
5	6	6	-0,26	41
5	6	7	0,23	13
5	7	1	-0,03	54
5	7	2	0,18	59
5	7	3	0,17	84
5	7	4	0,03	67
5	7	5	-0,23	51
5	7	6	0,17	67
5	7	7	0,24	21
5	8	1	0,45	18
5	8	2	0,05	21
5	8	3	-0,13	46
5	8	4	-0,06	34
5	8	5	-0,03	47
5	8	6	0,09	69
5	8	7	0,19	35

Apart those classes with a small number of observations, partial correlations seems to confirm the adequacy of the conditional independence assumption. Note also that this result could be expected, given that the matching variable “ordinal income classes” plays the role of “proxy variable” as suggested by Zhang (2015): it has the same support (the 7 classes), with a similar definition (ordered, but without the observed income figures).

Comments

As suggested by the committee members, National Statistical Institutes are in a very lucky position: they can decide how to design a sample and what questions can be included in a questionnaire. The exercise performed for household income and expenditures can consequently be exported to other possible data fusion problems of interest. For instance, this exercise can be extended to any pair of variables (income, W) where W is any other variable observed in a social survey (not only HBS, but also labour force or multipurpose surveys). Furthermore, the census sample can be a file where this additional question is included. The important warning to add is that this income variable" is not useful for direct analysis: it should only be used as a strong glue for the data fusion of interest.

A further warning is that this approach is consistent with the use of categorized income (so that the ordered income classes can be considered as a proxy). Furthermore, given the ordered income classes, independence holds only between any of the variables in HBS and income. This model does not hold if income is substituted by any other variable in EU-SILC.

As suggested by the committee, micro data should not be the primary focus of data fusion, and could be used only for internal purposes or as experimental statistics only after careful analysis of the quality and accuracy of the estimates.

As a last not, ecological inference models have not been applied in the case of data fusion, yet. When proxy variables cannot be reconstructed or included, ecological inference seems to be a very appealing approach.