ITACOSM 2019

FLORENCE 5-7 June 2019 | 6th ITAlian COnference on Survey Methodology

Optimal sample design with administrative sources for household finance surveys

Tiziana Tuoto

Joint work with: Giulio Barcaroli, Giuseppe Ilardi, Andrea Neri

Outline

1. Background and motivations
2. The previous sample design
3. The new available administrative sources
4. The new sample design
5. A comparative efficiency evaluation

Background and motivations

1. Banca d'Italia conducts since 1960 a **S**urvey on **H**ousehold **I**ncome and **W**ealth (**SHIW**).

2. The SHIW is the Italian component of the euro area **H**ousehold **F**inance and **C**onsumption **S**urvey (HFCS).

3. Measuring household wealth through surveys is challenging:
   - Wealth has a very skewed distribution;
   - Rich households are a rare population very difficult to contact and to enroll in the survey.

Objective of the study

- Modify the sample design of the SHIW using auxiliary information from population registers and personal income tax data.

The previous sample design of the SHIW

1. Two-stage sample design:
   - **First stage**: 350 municipalities (Stratified by region and demographic size of municipalities, PPS, some self-representative units);
   - **Second stage**: 8,000 households (SRS from population registers);

2. **Panel** component (about 4,000 hhs).

Strategies in the HFCS

Which sources are prefered in Europe for cachting the Wealth?

- Wealth tax data (individual level – FR, ES)
- Income tax data (individual level – FI, PL, HU, LU)
- Electric consumptions (individual level – CY)
- Average income or wealth (area level– BE, DE, AT)

(decreasing order of effectiveness)

Over-sampling largely used of for the right tail of the Wealth distribution

The new sample design: the role of the administrative sources

The administrative sources:

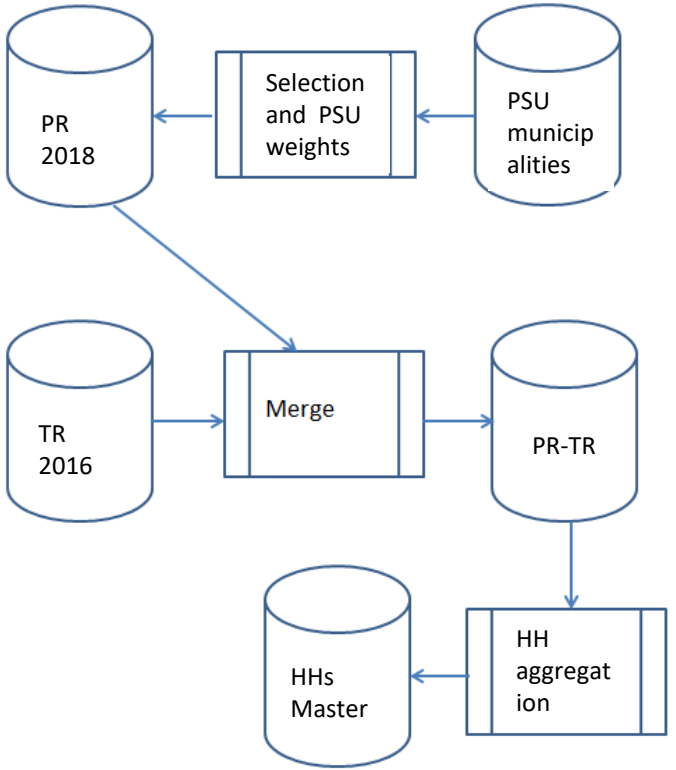
- Anagraphic Population Registers PR (latest version 2018)
- Personal Income Tax Register TR (latest version 2016).

The new sample design aims at modifying the allocation of the **non-panel** component at the **second stage**, via a stratification based on the auxiliary information available in TR.

Administrative sources and data integration

The steps of the process:

1. Municipalities selection from the PR and PSU weights assignment;
2. Merge with the TR (27.5 millions of individual records);
3. Aggregation of the **individuals** records into **households** records;
4. Definition of the optimal stratification and selection of the sample.



Stratification variables

1. total income **Y**
2. income from employment **YL**
3. income from self-employment **YM**
4. income from pensions (both work and social) **YTP**
5. revenues from real activities (rents and cadastral incomes) **YC**

The TR incomes used for stratification is a proxy of the incomes observed by the survey

Target and proxy variables for stratification

Correlation between fiscal and target income variables		
	Individuals	HHS
YL employement	71	77
YM self-employment	73	68
YTP pensions	80	84
YC real	50	50

Estimates of model parameters

type	beta	sig2
linear	0.8417096	1.194578e+04
linear	0.8229064	1.254771e+04
linear	0.5571044	1.863969e+04
linear	0.7665643	5.834692e+03
linear	0.1653843	5.436948e-01

Optimal stratification with SamplingStrata

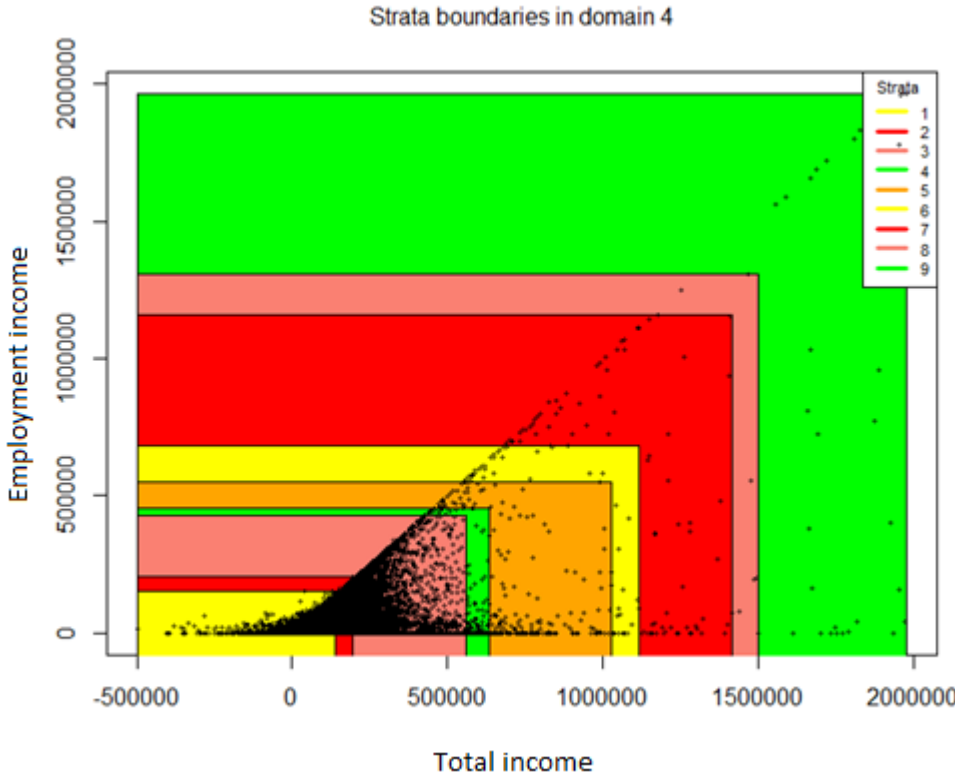
The R package SamplingStrata (Ballin & Barcaroli, 2014) is used for optimal stratification.

SamplingStrata uses a genetic algorithm that searches for the optimal solution exploring the universe of all the possible stratifications obtained by segmenting the five income variables.

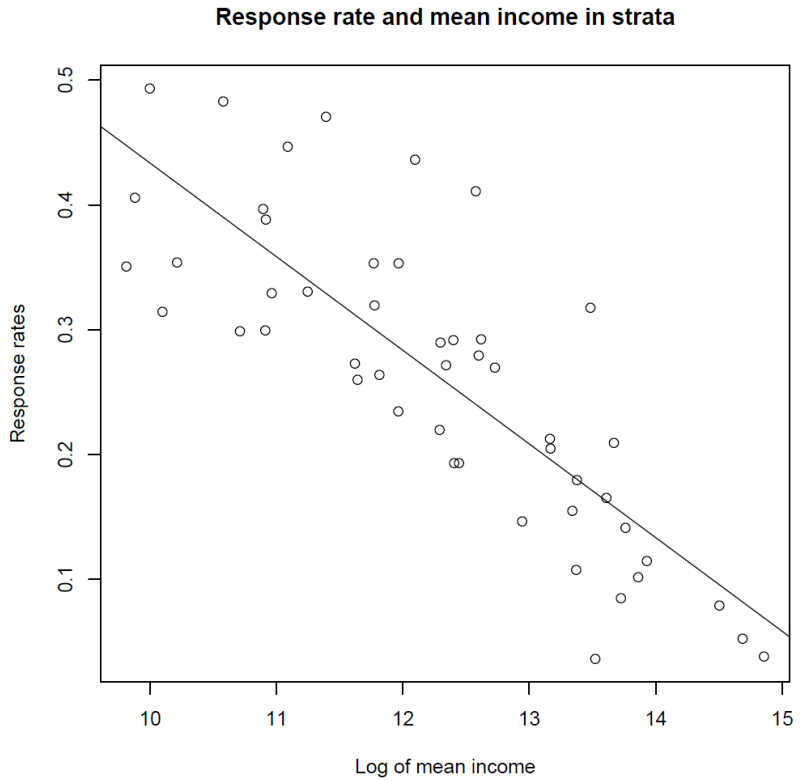
The solution is a partition in 9-10 stratification cells for each domain, characterized by different values of the 5 income variables.

Optimal stratification with SamplingStrata

Stratum	Pop	n	sampling rate	Limit_Y
1	2,180,619	1066	0.0005	140,908
2	122,074	117	0.0010	194,668
3	47,799	64	0.0013	559,077
4	28,413	60	0.0021	636,810
5	8,470	50	0.0059	1,028,955
6	2,200	50	0.0227	1,118,569
7	2,569	50	0.0195	1,416,904
8	276	50	0.1812	1,499,282
9	74	50	0.6757	6,015,764



Expected response rates in optimised strata



Comparative efficiency evaluation

Comparison between the efficiency of the average HT estimators shows that the new design is in general more efficient (especially for YM and YC) but also more stable across the domains than the previous design.

	Previous Design (CV%)					New Design (CV%)				
	Y	YL	YM	YTP	YC	Y	YL	YM	YTP	YC
1	8.0	12.7	26.0	6.6	22.6	3.9	7.1	11.8	8.5	10.5
2	4.0	4.5	16.9	5.5	13.8	4.1	7.5	11.9	8.5	12.3
3	10.9	10.0	53.9	12.9	28.9	5.2	8.9	12.1	8.6	10.8
4	4.9	7.5	16.1	10.1	16.3	4.5	7.6	12.8	9.7	11.0
5	4.6	6.8	25.5	7.7	25.5	4.9	7.2	15.3	9.5	12.1

ITACOSM 2019

FLORENCE 5-7 June 2019 | 6th ITAlian COnference on Survey Methodology

Thank you for your attention! Questions?

tuoto@istat.it
gbarcaroli@gmail.com
andrea.neri@bancaditalia.it
giuseppe.llardi@bancaditalia.it