Rivista Italiana di Economia Demografia e Statistica Volume LXIII nn. 1-2 – Gennaio-Giugno 2009

ANALYSIS OF GINI FOR EVALUATING ATTRITION IN ITALIAN SURVEY ON INCOME AND LIVING CONDITION

Claudio Ceccarelli, Giovanni Maria Giorgi

1. Introduction

European Community Statistics on Income and Living Conditions (Eu-Silc) is a set of statistical indicators of income, poverty and social exclusion which has been regulated by the European Parliament since 2003. In particular, the Regulation defines the responsibilities of Member States and Eurostat and lays down a set of rules to improve data quality, comparability and timeliness besides promoting a better integration of new surveys within national statistical systems.

In order to comply with all tasks entrusted by Eurostat and to deepen the analysis of income distribution, living conditions, inequality and poverty in Italy, the Italian National Institute of Statistics (Istat) has set up a survey on income and living condition (hereinafter It-Silc), which is substantially made of a cross-sectional and a longitudinal component.

Several studies show how the phenomenon of *selective attrition* may create a bias in the evaluation of results of analysis carried out by panel surveys due to non-random mechanisms generating non response from wave to wave.

According to Rendtel (2002, p.4), panel attrition "is defined by unit non-response of eligible persons or households that occurs after the first wave of panel".

Our paper aims at proving, by a decomposition of Gini concentration index, also known as Analysis of Gini (ANOGI) (see Frick et al., 2006), whether attrition introduces an element of bias in the analysis of income distribution at the regional level.

The work is organized as follows: in section 2, attrition is briefly described along with the main causes which may generate it in statistical surveys on households and individuals. In section 3, It-Silc main features are highlighted; section 4 gives a description of the methodology underlying the ANOGI; section 5 focuses on attrition patterns and after showing main results (section 6), the work ends with some concluding remarks and possible future developments (section 7).

2. Attrition

The substantial difference between repeated longitudinal and cross-sectional surveys lies in the sample of statistical units analyzed from time to time in the survey. In cross-sectional surveys, indeed, the initial theoretical sample is randomly drawn from population registers and is generally fixed, excluding those variations introduced during the survey design phase. In longitudinal surveys, from the second wave onwards, the sample size varies in function of previous wave respondents and of the different characteristics of the survey design which sets the mechanism generating the theoretical sample of the following wave. Within this context, the phenomenon of non-response assumes different connotations depending on whether it emerges during the first or the following waves.

From a merely theoretical and conceptual point of view, non-response at the first wave takes on characteristics which are absolutely similar to what happens in cross-sectional surveys. From the second wave onwards, non-response assumes different connotations. It may happen, indeed, that some statistical units, after responding to the first wave, choose not to participate in the survey anymore and drop out of the sample, albeit they are still eligible units: this is what is called attrition.

Attrition can be caused by different reasons which can be related either to fieldwork or to response behavior. Any mistake or lack which can inevitably occur during the various operational micro-phases of the survey may result in non participation. Among the others, it is worth mentioning the incapability to trace respondents with a high degree of mobility throughout the territory; non correct application of rules for the conduction of surveys; changes in survey technique or in questionnaire, which may induce refusal to participate; incorrect implementation of rules to trace sample units throughout the territory. Some additional causes, more or less subjective and depending on respondents' behavior and interaction with survey operators, can result in refusal to continue to participate in the panel. Impossibility to participate for health reasons or diffidence caused by change in interviewer from a wave to another or unconditioned refusal are some examples.

In panel surveys, sample units decrease in function of demographic exits due to individuals' death and migrations. Such units are no longer eligible units as they represent a part of population who is no longer considered as benchmark but defined out of scope.

There are also some cases in which attritors start participating again. Such participation pattern is called "temporary drop-out" and is due, for instance, to temporary impossibility to participate in the survey or to simple change of mind.

3. Italian survey on income and living condition (It-Silc)

The It-Silc is defined within the European Regulation no. 1177/2003 which outlines its main methodological, thematic and organizational aspects. In order to ensure the comparability of data collected by all Member States, common rules have been set for the following themes: sampling and tracing, definitions, list of primary variables, fieldwork aspect and imputation procedures, intermediate and final quality reports.

3.1 *Sample design*

The sample design planned and implemented in function of the main estimates which the survey has to produce and the planned study domains, is based on four independent longitudinal samples. Such design, called rotation design, provides that every year the longitudinal sample be closed after reaching the fourth wave and a new sample be started.

Each longitudinal sample is a two-stage sample: the primary sample units, municipalities, are stratified by region and demographic size, while the secondary sample units, households, are drawn from the population register of sampled municipalities.

Table 1 – *Rotational sample scheme in It-Silc.*

Samples	Years						
	2004	2005	2006	2007	2008	2009	2010
c1	$W_{(4)}$						
c2	$W_{(3)}$	$W_{(4)}$					
c3	$W_{(2)}$	$W_{(3)}$	$W_{(4)}$				
c4	W_1	W_2	W_3	W_4			
c5		W_1	W_2	W_3	W_4		
c6			W_1	W_2	W_3	W_4	
c7				W_1	W_2	W_3	W_4

In 2004, the first four longitudinal samples (*c1, c2, c3* and *c4* in Table 1) participate in the survey all for the first time. To start rotation, *c1* sample is assumed to be at its fourth and last wave ($W_{(4)}$ in Table 1), *c2* sample at its third wave ($W_{(3)}$), *c3* sample at its second wave ($W_{(2)}$) and *c4* sample at its first wave (W_1). The sample *c4* is the first longitudinal one which, started in 2004, will go on correctly over four waves, as per design, and in 2007 will allow the realization of

the first complete longitudinal sample (made up of W_1, W_2, W_3, W_4). The new longitudinal sample *c5* starts in 2005 and takes the place of *c1*, dropped in 2004. Generally, a new longitudinal sample is made up of the same first-phase units (municipalities) and new second-phase units (households).

The cross-sectional sample results every year from the union of the four longitudinal samples, each one for its specific wave: thus, each cross-sectional sample includes one fourth of households participating in the survey for the first time, one fourth of households participating for the second time, one fourth for the third time and one fourth for the fourth time.

The initial cross-sectional sample, relating to year 2004, is made up of about 32,000 households in all, that is 8,000 for each longitudinal sample.

For the year 2005, the cross-sectional sample size is given by the sum of the following items:

- number of households with individuals responding in the first wave for longitudinal samples *c2*, *c3* and *c4*;
- 8,000 newly drawn households belonging to the new longitudinal sample *c5*.

In this way, a household that has not been drawn for the first wave can enter the sample if joined with one sample member dropped out from the origin household.

The same procedure is used for the sample determination over the following years.

In the hypothesis of simple random sampling and given a level of sampling error, Eurostat fixes the minimum sample size; the definition of the sample size to be realized, according to which the whole survey is planned, comes from the hypothesis on design effect related to the sampling designs carried out by the various Statistical Agencies as well as from the supposed response rates by the survey. In longitudinal surveys the assessment of response rates requires, moreover, specification of an attrition trend.

3.2 *Cross-sectional weighting*

The cross-sectional weighting strategy develops through the following phases, which are usually used for the construction of estimators in various Istat's social surveys:

1) definition of design weight as the inverse of inclusion probability;
2) calculation of coefficients of correction for non-response bias;
3) determination of final cross-sectional weighting adjusted on according to known totals derived from external data relating to the distribution of households and persons in the target population.

The design weight is directly derived from the sampling design.

The second step is based on the hypothesis that the process generating non-response is not missing at random mechanism. In this case, a strategy is applied which follows the same criteria as weighting cells in order to single out sub-populations in which equal response behaviour may be assumed among those who have participated in the survey and those who have not. Sample households have been partitioned into cells through segmentation obtained using a *Chaid*-based decision tree (*Chi-squared automatic interaction detection*; Kass, 1980). Such a method consists in splitting the sample in sub-groups according to the relationship between ratio of response rate and explicative variables. The methodology underlying the *weighting cells* belongs to group of explicit modeling techniques to reduce non-response bias. In It-Silc, making use of both personal and fiscal data already available during the definition of the sample, a partition into homogeneous cells has been obtained in which is possible to adopt the hypothesis of missing at random non-response mechanism. To realize the cells, following data have been used: demographic size of the municipality; citizenship of the reference individual; region of residence; distribution of households by number of components; distribution of households by income group.

In order to calculate final cross-sectional weights, calibration estimators (see Deville and Särndal, 1992) are used. As provided for by Eurostat, each longitudinal sample at the first wave is bound to: resident population by geographical area, sex and age class[1], income reference year (31st December of year *t-1*), number of resident households by region as on date of survey (31st December of year *t*).

A brief mention should also be made of determination of final weight of the cross-sectional part of the survey for the years following the first. The weighting procedure, indeed, has to take into account that the cross-sectional sample comprises a longitudinal sample (e.g., *c5* for the 2005 sample) of households who have participated in the first interview and three samples (*c2*, *c3*, *c4*) who are already at their second interview. From a methodological point of view, the inclusion probability for households with sample individuals changes: three-fourths of the sample, indeed, who are not present in the first wave are no longer likely to be included for the calculation of design coefficient. For the households of individuals belonging to these three samples, however, the *weight share method* is used as if it were the design weight. The weight is defined as follows (see Istat, 2008):

[1] Age classes are: 0-15, 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75 and over.

$$\omega_h^t = \frac{\sum_{i \in s_0^t \cap h} \psi_i^t}{N_h^t} \tag{1}$$

where s^t is the entire sample, including new entrants, and $s_0^t \subset s^t$ is the longitudinal sample (individuals aged 14 and over belonging to the sample in the first year of the panel) of respondents in year *t*. ψ_i^t is the initial weight of the individual *i* in the year *t*, calculated as described above. N_h^t is the total number of sample and non-sample members of the household *h*. By construction, this household weight takes account of the correction for non-participation in the years following the first. Subsequently, after correcting non-response bias in the only new entrant sample *c5*, the whole cross-section undergoes the calibration procedure as described above.

3.3 *Longitudinal weighting*

A longitudinal sample produces estimates referred to the target population of the same year when the sample first participates in the survey (see Osier *et al.,* 2006). The longitudinal population in year *t+1*, includes individuals of the population in year *t* and excludes drop-outs between the year *t* and year *t+1* (OUT_{t+1}).

The target longitudinal population started in 2004 covering the years 2005 and 2006 assumes the following form:

$$P_{2005}^{(L)} = P_{2004} - OUT_{2005} \tag{2}$$

$$P_{2006}^{(L)} = P_{2005}^{(L)} - OUT_{2006} = P_{2004} - OUT_{2005} - OUT_{2006} \ . \tag{3}$$

In general terms, given a panel in year *n* starting in year *t=1*, the longitudinal target population is equal to:

$$P_n^{(L)} = P_1 - \sum_{t=2}^{n} OUT_t \ . \tag{4}$$

It is to be noted that the longitudinal population at time *t* ($P_t^{(L)}$) differs from the population at time *t* (P_t) in that it does not include individuals born or migrated into the reference population at time *t=1*.

Table 2 shows the values of the longitudinal population in respect of the whole population and people aged 16 and over in function of the various samples and respective waves. In year 2005, sample *c5* refers to the resident population in Italy on 31st December 2004, while sample *c4* refers to resident population in Italy on 1st January 2004 net of drop-outs (deaths and migrations) during the year 2004.

The strategy adopted for the weighting procedure develops through the following phases: determination of the design weight, calculation of attrition correction coefficients and determination of final weight (base weight). In the first wave, the theoretical sample drawn from municipal population registers, along with its design weight provides an estimate of the resident population.

Table 2 – *It-Silc longitudinal population by sub-samples and years (thousand).*

Sub-sample		ANNI		
		2004	2005	2006
c3	16 years and over	48,762	48,554	48,113
	Total	57,952	57,266	56,578
c4	16 years and over	48,762	48,554	48,113
	Total	57,952	57,287	56,594
c5	16 years and over	-	49,286	48,762
	Total	-	58,418	57,712

The "theoretical" sample to be interviewed in the second wave is composed of first wave respondents; coupled with its final weights (calculated as described in sub-section 3.2), it represents the resident population in year *t*.

Formally, let *c4* be the sample starting its longitudinal path in 2004 and Ψ_{2004} the vector of final weights of the year 2004. By construction, these weights make sample *c4* representative of the resident population in the year 2004.

$$(c4_{2004}, \Psi_{2004}) \rightarrow P_{2004} \ . \tag{5}$$

At time *t+1* (i.e. 2005) the longitudinal sample includes the initial sample net of drop-out from the reference population (out_{2005}) and those who do not participate in the survey, albeit still eligible (x_{2005})[2]:

$$c4_{2005}^{(L)} = c4_{2004} - out_{2005} - x_{2005} \ . \tag{6}$$

Assuming that sample drop-outs (out_{2005}), weighted by Ψ_{2004}, are an estimate of population drop-outs (OUT_{2005}), we have[3]:

$$\left(\left(c4_{2005}^{(L)} + x_{2005}\right), \Psi_{2004}\right) \rightarrow \left(P_{2004} - OUT_{2005}\right) \ . \tag{7}$$

In general terms, given a generic sample *j,* started in the year n_o, we have in the year *n*

$$\left(\left(cj_n^{(L)} + \sum_{t=n_0+1}^{n} x_t\right), \Psi_{n_0}\right) \rightarrow \left(P_{n_0} - \sum_{t=n_0+1}^{n} OUT_t\right) \ . \tag{8}$$

For the remaining sample to be representative of the initial population net of reference population drop-outs, weights have to be changed so as to take into account eligible units who stop participating in the survey. It follows that

[2] The longitudinal sample is also net of a third group of individuals, namely those who do not participate in the survey and, due to lack of information about them, it is not even clear whether they are still part of the target population or not. Therefore, each individual is to be assigned to one group or the other. In It-Silc surveys, a logistic regression model is used to estimate the propensity to stay in the population as a function of a set of explicative variables (age is obviously the most influential variable). This model is initially applied to the group of sample individuals for whom necessary information is available; the same parameters are then applied to individuals for whom no information is available for determining to which group to assign them.

[3] It is to be noted that such information cannot be obtained from external sources. Istat's demographic balance, indeed, gives the total of deaths and migrants (leaving out those who move into cohabitations, who are less numerous) but does not specify whether they belonged to the initial population. For instance, should an individual die during the period in question we could never know if he/she has just entered (by immigration, birth or following a move from an institution) or are already part of the target population.

$$\left(cj_n^{(L)}, \Psi_n\right) \rightarrow \left(P_{n_0} - \sum_{t=n_0+1}^{n} OUT_t \right) \tag{9}$$

where $\left(cj_n^{(L)}, \Psi_n\right)$ represents the initial population excluding drop-outs between n_0 and n.

The non-randomness of non-participation in the survey introduces an element of bias in estimates of aggregates. The basic idea to correct weights year after year was to work on the sample at the first wave in order to make it as representative of the initial population as possible, while taking into account those individuals who are still in the sample and inflating their weights in consideration of non randomness of attrition. From a practical point of view, an updating process has been set up which, starting from the individual weight of the sample unit and considering the various factors accounting for individuals remaining in a panel, leads to new individual weights. The method applied to inflate weights is the segmentation of the sample in homogeneous cells; the same as that described for the cross-sectional surveys with the only difference that in longitudinal surveys it is possible to use all information on individuals collected in previous years in order to determine the best partition possible. The so-called *base weight* is thus obtained, which is different over years even when it refers to the same individual, and by which it is possible to carry out longitudinal analyses: a real longitudinal weight, indeed, does not exist, but can be calculated starting from the base weight relating to individuals remained in the sample for the duration of the panel, but this basically depends on the nature of the analysis to carry out (see Osier *et.al.,* 2006; Ceccarelli and Cutillo, 2007).

4. Analysis of Gini (ANOGI): some methodological remarks

Among the indices used in the literature to investigate, for instance, the inequality of the income distribution, the Gini (1914) concentration ratio has again a role of primary and fundamental importance. Sometimes researchers have proposed different formulas from the original one with the purpose to fully exploit the application potentialities of the Gini index (G) in the most disparate fields[4].

[4] For a survey of the topical interest, new interpretations and extensions of the Gini index, see, e.g., Giorgi (1990, 1992, 1993, 1999, 2005).

In this context, in order to derive a useful decomposition by population subgroups, Lerman and Yitzhaki (1984) expressed G in terms of covariance[5] between a variable y (e.g. income) and its cumulative distribution function $F(y)$, that is:

$$G = \frac{2}{\mu}\mathrm{cov}\left[y, F(y)\right] \tag{10}$$

where μ is the mean of y.

Now, let us consider a population (P) divided in k subpopulations or groups $P = P_1 \cup P_2 \cup ... \cup P_k$ the Gini index can be written as (Yitzhaki, 1994, p.154)

$$G_u = \sum_{i=1}^{k} s_i G_i O_i + G_b \tag{11}$$

where

$s_i = p_i \bar{y}_i / \mu_u$ is the ratio between the mean of variable y in the subpopulation i weighted by the units presents in it (p_i) and the mean of y calculated on the whole population;
G_i is the Gini index within subpopulation i;
O_i is the *overlapping index* of subpopulation i with the entire population;
G_b is the between-subpopulations inequality.

The analysis based on formula (11) is known as Analysis of Gini or ANOGI (see Frick *et al*., 2006), that is similar to the Analysis of variance (ANOVA). In particular, the overlapping index is the element which conceptually distinguishes ANOGI and ANOVA, while G_b, albeit it can be negative (see Yitzhaki and Lerman, 1991, p. 322, note 9), is similar in its meaning to the between-group variance of ANOVA, i.e. it indicates the degree of inequality between subpopulations in terms of concentration.

4.1 *Overlapping index*

A brief mention should also be made of problems related to stratification and overlapping in the analysis of distribution of some variables (e.g income). Generally speaking, there is *stratification* (see, Yitzhaki, 1988, p.39; Yitzhaki and

[5] See also De Vergottini (1950, p. 453), Stuart (1954), and Piesch (1975, p. 39).

Lerman, 1991, p.319) when a group is isolated from the members of other groups. More specifically, there is perfect stratification when the members of a group occupy distinct range within an overall distribution and no member of a group is included in the same range of another group. A classical example is the subdivision of a population into income deciles. Each unit of a given decile belongs exactly to the range of the considered decile. In the absence of stratification, overlapping occurs.

In brief, being $\overline{F}_{ui}(y) = \int F_u(y) dF_i(y)$ the expected rank of the units belonging to group *i* within the distribution the entire population, and given that quantity $\mathrm{cov}_i(y, F_u(y)) = \int (y - \mu_i)(F_u(y) - \overline{F}_{ui}(y)) f_i(y) dy$ represents the covariance between *y* (income) and rank of units belonging to group *i ,* calculated on their position in the overall distribution, the overlapping index O_i may be expressed as:

$$O_i = O_{ui} = \frac{\mathrm{cov}_i(y, F_u(y))}{\mathrm{cov}_i(y, F_i(y))} \ . \tag{12}$$

In this case, the index measures the degree of overlapping between the distribution of the units belonging to group *i* with the distribution of the entire population. In other words, there is perfect stratification (as in the case of income deciles) when the units of the group *i* occupy the same relative position both in the population and in the group distribution.

With reference to the population partition in *k* groups, the overlapping index referred to a given group *i* is expressed by the following formula:

$$O_i = \sum_j p_j O_{ji} = p_i O_{ii} + \sum_{j \neq i} p_j O_{ji} = p_i + \sum_{j \neq i} p_j O_{ji} \tag{13}$$

where (Yitzhaki, 1994)

$$O_{ji} = \frac{\mathrm{cov}_i(y, F_j(y))}{\mathrm{cov}_i(y, F_i(y))} \ . \tag{14}$$

The formula (14) represents the overlapping index of group *j* by group *i* and provides a measure of the presence of group *j* units within the group *i*. In particular, the main properties are (Frick *et al*., 2006, p.437):

i. O_{ji}=0, no member of the *j* group lies in the range of distribution *i.* Group *i,* therefore, is a "perfect stratum", i.e. its range is not "contaminated" by members of the *j* group.
ii. O_{ji}=1, the distributions of group *i* and *j* are identical, being $O_{ii} = 1$.
iii. O_{ji} is not symmetrical, that is the higher O_{ji} the lower O_{ij}.
iv. $O_{ji} \leq 2$. If all observations of distribution *j* are in the range of *i* and are concentrated at the mean of distribution *i* then O_{ji} assumes the maximum value (Yitzhaki 1994, p.151).

4.2 *Between-group inequality (G_b)*

Another essential element of ANOGI is the measurement of the between-group inequality *(G_b)* defined as:

$$G_b = \frac{2\,\mathrm{cov}(\mu_i, \overline{F}_{ui}(y))}{\mu_u} \tag{15}$$

which is twice the covariance between the mean of variable *y* of each group and the groups' mean rank in the whole population, divided by the mean of *y* calculated on the whole population.

Pyatt (1976) introduced a type of between-group inequality measure (G_b^p) based on the assumption of perfect stratification. In this case, the covariance is calculated between the mean of each group and the groups' mean rank. From a strictly formal point of view, this is defined as:

$$G_b^p = \frac{2\,\mathrm{cov}(\mu_i, \overline{F}_i(y))}{\mu_u} \ . \tag{16}$$

From a conceptual point of view, it may be argued that G_b is not really a concentration index because, as mentioned earlier, it can be negative. As per formula (11), moreover, in case of perfect stratification – overlapping index equals to zero – the G_b indicator reaches its upper level as the quantification of the amount of total inequality is explained by between-group inequality.

It derives that (see Yitzhaki and Lerman, 1991, p. 322)

$$G_b^p \geq G_b \ . \tag{17}$$

With simple algebra (11) can be written as

$$\begin{aligned} G_u &= \sum_{i=1}^{k} s_i G_i + \sum_{i=1}^{k} s_i G_i (O_i - 1) + G_b^p + (G_b - G_b^p) = \\ &= IG + IGO + BG + BGO \end{aligned} \tag{18}$$

and these four components of ANOGI may be conceptually compared to ANOVA.

Frick *et al*. (2006, p.438-440) schematize the comparison between ANOGI and ANOVA as follow:

Component similar to ANOVA

Within $\quad IG = \sum_{i=1}^{k} s_i G_i \quad 0 \leq IG \leq G_u$

Between-Pyatt $\quad BG = G_b^p \quad 0 \leq BG \leq G_u$

Additional component respect to ANOVA

Within $\quad IGO = \sum_{i=1}^{k} s_i G_i (O_i - 1)$

Between $\quad BGO = (G_b - G_b^p) \quad -BG - IGO - IG \leq BGO \leq 0$.

5. Response pattern

On the basis of definitions given in section 2, three groups of individuals have been identified: respondents, those namely who have actively participated in a survey; *out of scope,* all individuals who, after responding to the previous wave*,* exit from the target population (moved abroad, moved to institutional household, deaths); attritors*,* all individuals, that for various reasons, have not participated in the survey even after responding to the previous wave*,* excluding those who exited from the sample and entered it again. These latter have been excluded as they are

not representative of a monotonic response pattern. In terms of response behaviour, those who have second thoughts cannot have the same approach to the survey as those who definitely exit it.

Table 4 – *Response pattern for wave and geographical area.*

Geographical area	Waves						
	2004	2005			2006		
North-West	Respondent	*Sample size*	*6,385*	*100.0*			
		Out of scope	117	1.8			
		Attritors	882	13.8			
		Respondent	5,386	84.4	*Sample size*	*5,386*	*100.0*
					Out of scope	52	1.0
					Attritors	805	14.9
					Respondent	4,529	84.1
North-East	Respondent	*Sample size*	*6,248*	*100.0*			
		Out of scope	96	1.5			
		Attritors	774	12.4			
		Respondent	5,378	86.1	*Sample size*	*5,378*	*100.0*
					Out of scope	78	1.5
					Attritors	595	11.1
					Respondent	4,705	87.4
Centre	Respondent	*Sample size*	*6,301*	*100.0*			
		Out of scope	104	1.7			
		Attritors	952	15.1			
		Respondent	5,245	83.2	*Sample size*	*5,245*	*100.0*
					Out of scope	61	1.2
					Attritors	588	11.2
					Respondent	4,596	87.6
South	Respondent	*Sample size*	*5,271*	*100.0*			
		Out of scope	66	1.3			
		Attritors	395	7.5			
		Respondent	4,810	91.2	*Sample size*	*4,810*	*100.0*
					Out of scope	50	1.0
					Attritors	319	6.6
					Respondent	4,441	92.4
Islands	Respondent	*Sample size*	*2,130*	*100.0*			
		Out of scope	46	2.2			
		Attritors	213	10.0			
		Respondent	1,871	87.8	*Sample size*	*1,871*	*100.0*
					Out of scope	13	0.7
					Attritors	117	6.3
					Respondent	1,741	93.0
Italy	Respondent	*Sample size*	*26,335*	*100.0*			
		Out of scope	429	1.6			
		Attritors	3,216	12.2			
		Respondent	22,690	86.2	*Sample size*	*22,690*	*100.0*
					Out of scope	254	1.1
					Attritors	2,424	10.7
					Respondent	20,012	88.2

The longitudinal component in It-Silc survey shows how the various types of non-response depend on different design characteristics. In Table 4, for instance, the response pattern of samples c3 and c4 (those who have reached the third wave) are analyzed in order to evaluate how the length of panel may affect attrition.

Table 4 shows the response patterns in Italy and by geographical area; in particular, it focus on the very low percentage of out of scope individuals (oscillating around 1.6%). It is, indeed, to be noted that the said percentage regarding both the second and the third wave is quite stable, with the only exception of the Islands, where it varies between 2.2% and 0.7%. The above values prove that the It-Silc sample is adequate to represent the drop-outs from the longitudinal population regardless of the wave.

The most interesting element is the trend of attritors. The total sample shows an attrition level after one year equal to 12.2%, which reduces to 10.7% after two years. With the only exception of the North-West, where the attrition level rises from 13.8% to 14.9%, the same downward trend is reported in other geographical areas, especially in the Islands.

Respondents' different behaviours may be accounted for by various factors, such as, for instance, the different structure of interviewers' network, or citizens' different awareness of the importance of official statistics, or the well-known difficulties to obtain high response rates in large cities. From the analysis of the pattern of those who have responded to all three waves, indeed, it emerges that the South (84.3%) has a higher rate of permanence in the sample than the North-West (70.9%).

These differences produce an increase of the variability of the final weights and a decrease of the accuracy of estimated parameters.

6. Results

This paragraph illustrates the results of the analyses carried out on the distribution of main individual income sources singled out in the It-Silc survey: equivalent[6], employee, self-employment and pension (after retirement from employment) income. The said analyses have been carried out on the national territory and by geographical area according to the following hypothesis.

The first hypothesis analyzes the effects of attrition between the first and the second wave and compares 22.690 respondents and 3.216 attritors. The second hypothesis compares the attritors between the second and the third wave with

[6] The individual *equivalent income* is the total household income assigned to each of its members equivalized by the OECD modified scale.

20.012 respondents vs. 2.424 attritors. The third hypothesis, resulting from the combination of the two previous ones, analyzes the attrition between the first and the third wave and compares 20.012 respondents and 5.640 attritors.

Table 5 – *Evaluation of the effects of attrition between the first and the second wave by income and geographical area.*

Geographical area	Respondent				Attritors			
	Mean	Fi	Oi	Gi	Mean	Fi	Oi	Gi
North-West								
Equivalent income	18,169.73	0.5054	0.9999	0.2924	17,651.65	0.4669	0.9888	0.2692
Employee income	15,987.20	0.5033	1.0005	0.2860	15,664.61	0.4802	0.9908	0.2652
Self-employment income	16,133.32	0.5021	1.0031	0.4840	15,976.78	0.4876	0.9802	0.4679
Pension income	12,275.14	0.4977	1.0045	0.3223	12,002.56	0.5164	0.9688	0.2779
North-East								
Equivalent income	18,835.15	0.5000	0.9983	0.2978	18,411.57	0.5000	1.0153	0.2770
Employee income	15,249.10	0.5016	0.9986	0.3007	15,042.79	0.4896	1.0103	0.3049
Self-employment income	16,678.46	0.4952	1.0123	0.5066	16,786.68	0.5302	0.9165	0.3933
Pension income	11,488.78	0.4976	0.9982	0.3129	11,479.61	0.5205	1.0230	0.2943
Centre								
Equivalent income	17,154.20	0.5002	0.9982	0.2967	17,380.28	0.4990	1.0102	0.3083
Employee income	14,700.26	0.4996	0.9985	0.3045	14,794.55	0.5020	1.0084	0.3116
Self-employment income	14,226.38	0.4988	1.0087	0.4709	13,891.87	0.5056	0.9579	0.4297
Pension income	12,408.85	0.4978	1.0042	0.3383	12,809.95	0.5140	0.9726	0.3432
South								
Equivalent income	12,608.92	0.5007	0.9974	0.3072	12,843.27	0.4918	1.0309	0.3398
Employee income	12,683.24	0.5036	0.9974	0.3389	12,661.84	0.4603	1.0241	0.3687
Self-employment income	11,981.14	0.5017	1.0018	0.4832	11,918.59	0.4808	1.0064	0.5233
Pension income	10,409.51	0.4989	1.0007	0.3288	10,215.15	0.5137	0.9908	0.3626
Islands								
Equivalent income	12,665.20	0.4971	1.0044	0.3298	12,750.99	0.5259	0.9649	0.2988
Employee income	13,413.71	0.4994	1.0072	0.3616	13,482.82	0.5049	0.9470	0.3305
Self-employment income	12,661.71	0.5065	0.9866	0.4352	12,301.97	0.4434	1.1145	0.5303
Pension income	*no attritors for this sub-sample*							
Italy								
Equivalent income	16,463.76	0.4989	1.0017	0.3119	16,494.46	0.5080	0.9880	0.2996
Employee income	15,406.36	0.5014	1.0017	0.3040	15,243.50	0.4886	0.9876	0.3052
Self-employment income	14,887.75	0.4985	1.0055	0.4889	14,500.56	0.5093	0.9639	0.4470
Pension income	11,669.70	0.4971	1.0031	0.3281	12,023.74	0.5238	0.9734	0.3139

When comparing two sub-populations, the methodology described in section 4 undergoes a significant simplification. In order to verify the hypothesis that the two sub-populations come from the same population, or in other words, that there is complete overlap between them, the following conditions have to occur (Frick *et al.*, 2006, p.442-443):

i. $\overline{y}_{resp} = \overline{y}_{attr}$, same average income;

ii. $\overline{F}_{resp}(y) = \overline{F}_{attr}(y) = 0.5$, mean rank equals to 0.5;

iii. $G_{resp} = G_{attr}$, same Gini coefficient;

iv. $O_{resp} = O_{attr} = 1$, overlapping index equals to 1.

Table 6 – *Evaluation of the effects of attrition between the second and the third wave by income and geographical area.*

Geographical area	Respondent				Attritors			
	Mean	Fi	Oi	Gi	Mean	Fi	Oi	Gi
North-West								
Equivalent income	18,572.58	0.4981	0.9999	0.2924	18,747.05	0.5105	0.9989	0.3215
Employee income	16,851.38	0.5048	0.9968	0.2769	16,557.57	0.4745	1.0171	0.3102
Self-employment income	16,283.53	0.4990	1.0048	0.4818	16,535.53	0.5051	0.9786	0.4788
Pension income	12,097.54	0.5001	0.9949	0.3090	12,229.29	0.4994	1.0268	0.3514
North-East								
Equivalent income	19,277.84	0.5004	1.0026	0.3004	18,935.48	0.4966	0.9801	0.2843
Employee income	15,850.99	0.5020	1.0047	0.2974	15,332.98	0.4848	0.9583	0.2668
Self-employment income	17,568.23	0.5003	0.9991	0.5152	18,021.41	0.4982	1.0071	0.5357
Pension income	11,437.16	0.4983	1.0004	0.2985	11,635.57	0.5158	0.9973	0.2973
Centre								
Equivalent income	17,530.29	0.5025	0.9954	0.2911	17,173.44	0.4805	1.0350	0.3076
Employee income	15,469.71	0.5024	0.9998	0.2919	15,179.04	0.4817	1.0031	0.2969
Self-employment income	14,165.08	0.4978	1.0123	0.4734	13,936.29	0.5162	0.9071	0.3899
Pension income	12,702.85	0.5038	0.9960	0.3408	12,539.87	0.4661	1.0310	0.3302
South								
Equivalent income	12,880.00	0.4994	0.9956	0.3009	13,214.94	0.5081	1.0391	0.3232
Employee income	13,509.82	0.5028	0.9991	0.3337	12,901.04	0.4634	1.0086	0.3573
Self-employment income	11,971.85	0.4987	0.9938	0.4591	11,913.98	0.5154	1.0399	0.6489
Pension income	10,138.28	0.4981	1.0026	0.3072	10,297.75	0.5390	0.9612	0.2587
Islands								
Equivalent income	12,916.80	0.5008	0.9916	0.3300	13,739.76	0.4871	1.1196	0.4091
Employee income	14,086.05	0.5019	1.0004	0.3518	13,664.81	0.4696	0.9868	0.3348
Self-employment income	13,221.28	0.5037	0.9861	0.4740	14,010.70	0.4632	1.1531	0.5789
Pension income	10,583.31	0.4961	0.9899	0.3337	10,420.67	0.5835	1.0377	0.2851
Italy								
Equivalent income	16,745.74	0.4972	1.0006	0.3110	17,067.68	0.5234	0.9891	0.3315
Employee income	15,406.36	0.5014	1.0017	0.3040	15,243.50	0.4886	0.9876	0.3052
Self-employment income	15,094.98	0.4951	1.0188	0.4982	14,832.95	0.5126	0.9658	0.5130
Pension income	11,604.07	0.4988	0.9998	0.3186	11,954.37	0.5109	1.0012	0.3258

Any significant difference may be interpreted as an indication that the two sub-samples do not represent the same population. It may be better said that attritors' non-random behaviour can cause a bias in the analysis of income distribution.

Table 7 – *Evaluation of the effects of attrition between the first and the third wave by income and geographical area.*

Geographical area	Respondent				Attritors			
	Mean	Fi	Oi	Gi	Mean	Fi	Oi	Gi
North-West								
Equivalent income	18,063.40	0.5032	1.0025	0.2912	17,963.17	0.4914	0.9929	0.2852
Employee income	16,013.00	0.5063	0.9972	0.2816	15,889.76	0.4836	1.0050	0.2876
Self-employment income	15,852.88	0.4987	1.0090	0.4827	16,262.28	0.5035	0.9785	0.4795
Pension income	12,271.80	0.4962	1.0020	0.3227	12,151.00	0.5117	0.9963	0.2994
North-East								
Equivalent income	18,954.44	0.5007	1.0019	0.3019	18,508.99	0.4975	0.9926	0.2709
Employee income	15,344.13	0.5045	0.9965	0.3006	14,818.04	0.4852	1.0108	0.3033
Self-employment income	16,884.64	0.4956	1.0184	0.5125	16,537.06	0.5138	0.9387	0.4211
Pension income	11,461.74	0.4946	0.9946	0.3146	11,600.94	0.5234	1.0016	0.2945
Centre								
Equivalent income	17,213.96	0.5023	0.9970	0.2959	17,112.32	0.4930	1.0095	0.3063
Employee income	14,701.10	0.5025	0.9949	0.2997	14,755.93	0.4928	1.0148	0.3229
Self-employment income	14,314.45	0.4977	1.0182	0.4789	14,700.21	0.5063	0.9478	0.4195
Pension income	12,497.96	0.5003	1.0016	0.3404	12,460.68	0.4990	0.9951	0.3342
South								
Equivalent income	12,548.35	0.4993	0.9953	0.3050	13,020.27	0.5042	1.0270	0.3378
Employee income	12,698.68	0.5052	0.9979	0.3373	12,021.77	0.4703	1.0104	0.3644
Self-employment income	11,897.72	0.4989	1.0040	0.4847	11,936.79	0.5059	0.9830	0.4991
Pension income	10,397.02	0.4969	1.0027	0.3303	10,999.12	0.5231	0.9777	0.3389
Islands								
Equivalent income	12,628.20	0.4957	0.9980	0.3269	12,919.76	0.5233	1.0171	0.3235
Employee income	13,524.38	0.5025	1.0054	0.3610	12,922.74	0.4880	0.9721	0.3426
Self-employment income	12,639.82	0.5048	0.9895	0.4405	12,403.42	0.4802	1.0496	0.4642
Pension income	10,684.04	0.4935	1.0117	0.3371	10,455.45	0.5484	0.9312	0.2555
Italy								
Equivalent income	16,389.32	0.4957	1.0047	0.3128	16,749.10	0.5156	0.9822	0.3013
Employee income	14,695.65	0.5021	0.9994	0.3093	14,557.87	0.4929	1.0021	0.3144
Self-employment income	14,843.96	0.4961	1.0119	0.4931	14,806.90	0.5124	0.9616	0.4514
Pension income	11,656.71	0.4949	1.0042	0.3299	11,922.03	0.5211	0.9826	0.3127

A first analysis of results obtained shows a certain homogeneity compared to the specified three work hypotheses; it emerges that the sample of respondents and

that of attritors come from the same population or, in other words, are representative of the same population.

Another element common to all the three work hypotheses is the substantial difference among income distributions that, although predictable, is unlikely to be observed both in the sample of respondents and in that of attritors.

7. Conclusion and remarks

The ANOGI is used on It-Silc data in order to analyze non-response behaviours and in the evaluation of the effects of attrition on the core survey variable (i.e. income). In particular, it proves the efficacy and simplicity of use of the ANOGI within the variability study of sub-populations.

Unlike several studies on attrition mainly aimed at determining the response probability in function of individuals' characteristics and of the context in which the survey is conducted, this study introduces a new perspective by the direct evaluation of the attrition effect on studied variables.

The results of this work represent a first step towards the realization of a larger project mainly aimed at studying methods for the analysis of income, living conditions and poverty.

To briefly sum up the results obtained, it can be said that the panel drop-out, analyzed separately between the first and the second wave, and between the second and the third wave and on the whole panel, results in an increase of sampling error.

The supposed bias due to sample self-selection produces only negligible effects on all types of income and geographical areas.

Acknowledgements

The present work has been realized within the PRIN-2007 project: "*Inequality, poverty and social exclusion: analysis and coherence of information sources, new measures and interpretative methods*".

References

Ceccarelli C., Cutillo A. 2007. *Il trattamento della mancata risposta totale nell'indagine Eu-Silc: una valutazione tramite una misura del cambiamento,* Congiuntura**,** 1° Trimestre 2007, CREF, Udine, 91-112.

De Vergottini M. 1950. *Sugli indici di concentrazione,* Statistica**,** 10, 445-454.

Deville J. C., Särndal C. E. 1992. *Calibration Estimator in Survey Sampling,* Journal of the American Statistical Association, 87, 376-382.

European Parliament and Council 2003. *Regulation (EC) No 1177/2003 of the European Parliament and of the Council of 16 June 2003 concerning Community statistics on income and living conditions (Eu-Silc)*, Official Journal of the European Union.

Frick, R. J., Goebel, J., Schechtman, E., Wagnar, G.G., and Yitzhaki, S. 2006. *Using Analysis of Gini (ANOGI) for Detecting Whether Two Sub-Samples Represent the Same Universe*. *The German Socio-Economic Panel Study (SOEP) Experience*, Sociological Methods and Research, 34, 427-468.

Gini, C. 1912. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche*. In: Studi Economico-giuridici della Regia Facoltà di Giurisprudenza, 3(2), Bologna: Cuppini.

Gini, C. 1914. Sulla *misura della concentrazione e della variabilità dei caratteri,* Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti, 73, 1203–1248. (English translation in Metron, 2005, 63, 3–38).

Giorgi, G. M. 1990. *Bibliographic portrait of the Gini concentration ratio*, Metron, 48, 183–221.

Giorgi, G. M. 1992. *Il rapporto di concentrazione di Gini: Genesi, evoluzione ed una bibliografia commentata*, Siena: Libreria Editrice Ticci.

Giorgi, G. M. 1993. *A fresh look at the topical interest of the Gini concentration ratio,* Metron, 51, 83–98.

Giorgi, G. M. 1999. *Income Inequality Measurement: The Statistical Approach*. In J. Silber ed., Handbook on Income Inequality Measurement, Boston: Kluwer Academic Publishers, 245-260.

Giorgi, G. M. 2005. *Gini's Scientific Work: An Evergreen*, Metron, 63, 493-503.

Istat 2008. *L'indagine europea sui redditi e le condizioni di vita delle famiglie (Eu-Silc),* Collana Metodi e norme, n.37.

Kass, G.V. 1980. *An explanatory technique for investigating large quantities of categorical data*, Applied Statistics, 29, 119-127.

Lerman, R. I. and Yitzhaki, S. 1984. *A Note on the Calculation and Interpretation of the Gini Index*, Economics Letters, 15, 163–68.

Osier G., Museux J.M., Seoane P., Verma V. 2006. *Cross-sectional and longitudinal weighting for EU-SILC rotational design,* contributed paper on Methodology of Longitudinal Survey (MOLS 2006), University of Essex.

Piesch, W. 1975. *Statistische Konzentrationsmasse,* Tübingen: J.B.C. Mohr (Paul Siebeck).

Pyatt, G., 1976. *On the interpretation and disaggregation of Gini coefficient,* Economic Journal, 86, 243–255.

Rendtel U. 2002. *Attrition in Household Panels: a survey*, Chintex working paper, 4.

Stuart, A. 1954. *The correlation between variate-values and ranks in samples from a continuous distribution,* British Journal of Psychology, 7, 37–44.

Yitzhaki, S. 1988. *On Stratification and Inequality in Israel,* Bank of Israel Economic Review, 63, 36-51.

Yitzhaki, S. 1994. *Economic Distance and Overlapping of Distributions,* Journal of Econometrics, 61, 147–159.

Yitzhaki, S. and Lerman, R. 1991. *Income Stratification and Income Inequality*, Review of Income and Wealth, 37, 313–329.

SUMMARY

The Italian National Institute of Statistics (Istat) has set up a survey on income and living conditions (It-Silc), mainly composed of a cross-sectional and a longitudinal component.

This paper aims at proving, by a decomposition of Gini concentration index, also known as Analysis of Gini (ANOGI), whether attrition introduces an element of bias in the analysis of income distribution.

Compared to other studies in the literature, it introduces a new perspective by the direct evaluation of the attrition effect on studied variables.

Claudio CECCARELLI, Senior Researcher of Statistics, Italian National Institute of Statistics, Division for Survey on Living Conditions and Quality Life, clceccar@istat.it

Giovanni Maria GIORGI, Full Professor of Statistics, University of Rome "La Sapienza", Department of Statistics, Probability and Applied Statistics, giovanni.giorgi@uniroma1.it