# Cross-sectional and longitudinal weighting for the EU-SILC rotational design

Guillaume Osier, Jean-Marc Museux and Paloma Seoane[1]

*(Eurostat, Luxembourg)*

Vijay Verma[2]

*(University of Siena, Italy)*

## 1. THE EU-SILC INSTRUMENT

### 1.1. Aim of the project

European Statistics on Income and Living Conditions (EU-SILC) is a European project launched in 2003 by the European Parliament and the Council (Regulation 1177/2003 of 16th June 2003). It aims to collect every year timely and comparable micro-data on income, poverty and social exclusion at European level. EU-SILC is a household survey, which means data are collected at household level on all the current household members. In this framework, both household and individual indicators are estimated in cross-sectional and longitudinal dimensions.

### 1.2. The rotational design

In order to meet both the cross-sectional and longitudinal requirements, Eurostat has recommended a rotational design based on four rotational groups. At the first year of EU-SILC, four panels of individuals are drawn. Each subsequent year, one panel is dropped out and a new one is substituted for.

Figure 1: the rotational design

Panel introduced in year:

| | | | | |
|---|---|---|---|---|
| | X | | | |
| | X | X | | |
| | X | X | X | |
| Y-3 | **X** | X | X | |
| Y-2 | | **X** | X | |
| Y-1 | | | **X** | |

[1] Guillaume.Osier@cec.eu.int, Jean-Marc.Museux@cec.eu.int

[2] Verma@unisi.it

Such a structure is interesting because it enables both longitudinal and cross-sectional estimation:

- Longitudinal inference is clearly possible over four consecutive years through the following-up of one panel since the first year of its selection. In the same way, longitudinal following-up can be done over three (respectively two) years by combining two panels (respectively three panels).
- Cross-sectional inference is possible because of the refreshing of the sample at each year (a new panel is inserted).

In short, the rotational design can be regarded as a good balance between two opposite estimation strategies:

- Independent samples drawn every year and which are recommended for cross-sectional estimation but cannot be used for longitudinal inference.
- A pure panel which is clearly the most natural way of longitudinal inference but becomes outdated for cross-sectional purposes.

Table 1: three estimation strategies

| | **Cross-sectional estimation** | **Longitudinal estimation** |
|---|---|---|
| **Independent samples** | Recommended | Impossible |
| **Pure panel** | Problem of outdating | Recommended |
| **Rotational design** | Suitable | Suitable |

### 1.3. Scope and outline of the document

The objective of the present document is to propose a unified structure for the whole weighting procedure for the standard integrated EU-SILC design, covering the initial sample, and its cross-sectional as well as longitudinal development. Such an integrated structure is possible and desirable, given that different parts of the EU-SILC design are inter-related.

The document is outlined as follows:

1. Weighting for the first year of each subsample (panel).
2. Computation of base weights.
3. Cross-sectional weights, year 2 onwards.
4. Longitudinal weights.

## 2. WEIGHTING FOR THE FIRST YEAR OF EACH SUBSAMPLE

### 2.1. The sample design

In most of the situations, it consists of a multi-stage selection of households. Then all the household members are exhaustively interviewed (cluster sampling with households as clusters of individuals).

## 2.2. The weighting procedure

### 2.2.1. Household design weights

They are defined for all selected households, and not only for those which respond to the survey. Household design weights are calculated by taking the inverses of the household inclusion probabilities.

Design weights ensure unbiased estimates for totals in the ideal case of full response. However, because of non-response, they have to be corrected in order to reduce bias burden at the estimation stage.

### 2.2.2. Adjustment for non-response at the first wave.

In a panel, the largest loss of the sample due to non-response generally occurs at the first wave when the household is introduced into the survey. Good and efficient procedures to re-weight the responding cases are therefore a critical requirement. However, the possibilities are often constrained by lack of information: non-response adjustment has to be based on characteristics which are known for both responding and non-responding households.

There are two commonly used procedures for non-response weighting. The first is to modify the design weights by a factor inversely proportional to the response rate within each "weighting cells" (appropriately determined grouping of units). It is common to use sampling strata or other geographical partitions as weighting cells. The response rates should be computed with data weighted by the design weights:

$$R_k = \frac{sum\ \ of\ \ design\ \ weights\ \ of\ \ responding\ \ units\ \ in\ \ cell\ \ k}{sum\ \ of\ \ design\ \ weights\ \ of\ \ selected\ \ units\ \ in\ \ cell\ \ k}$$

Numerous, very small weighting cells can result in a large variation in $R_k$ values, and should be avoided. On the other hand, if only a few broad classes are used, little variation in the response rates across the sample may be captured – making the whole re-weighting process ineffective. On practical ground, cells of average size 100-300 units may be recommended.

The other alternative is to use a regression-based approach. Using an appropriate model such as logit regression, response propensities can be estimated as a function of auxiliary variables, which are available for both responding and non-responding cases. When many auxiliary variables are available, this approach is preferable to the first one.

### 2.2.3. Integrative calibration of household weights

At this step, household weights are adjusted so that they reproduce the totals of external variables. This procedure is performed in an "integrative" way. This means both household and individual external information can be used in a single-shot calibration at household level. Individual variables are added up at household level and then used under that aggregated form. This will allow using two different levels of data while keeping household and individual weights equal (see next).

### 2.2.4. Individual weights

Considering the sample design and particularly that all household members are interviewed, individual weights shall be equal to the corresponding household weights. At this stage, no further calibration is needed (individual calibration variables have been already used in 2.2.3) and even not desirable because it would break down the essential requirement of having household and individual weights equal.

## 2.3. Some technical issues

### 2.3.1. Non-response correction

It may be useful to apply the adjustment in two steps:

i. For non-contact (of households and/or of selected individuals)
ii. For non-response, once a contact with the households or the person concerned has been made.

For both steps, especially for i., area level characteristics provide a main part of the auxiliary variables explaining non-response.

In dealing with the effect of non-response, it is of crucial importance to identify responding and non-responding units correctly.

- Selected units which turn out to be non-eligible or non existent must be excluded and not counted as non-responding.
- Imputation has to be made for units with unknown status, i.e. when it is not clear whether they are non-eligible or non-respondents. Every unit has to be assigned uniquely to one category or the other.
- In surveys where substitution has been allowed, non-responding original units for which successful substitutions have been made are to be considered as responding units in the computation of response rates for the purpose of determining non-response weights.
- Note also that by “respondent” is meant “final interview accepted”.

### 2.3.2. Trimming

This refers to recoding of extreme weights to more acceptable values. The objective of trimming is to avoid excessive increase in variance due to weighting, even though the process introduces some bias. The aim is to seek a trimming procedure which reduces the mean squared error.

- At each step of the weighting procedure, the distribution of the resulting weight adjustments should be checked.
- In principle, the results of every step should be subject to the trimming procedure. This applies to weighting for non-response as well.

There is no rigorous procedure for general use for determining the limits for trimming. While more sophisticated approaches are possible, it is desirable to have a simple and practical approach.

Such an approach may be quite adequate for the purpose if the permitted limits are wide enough. The following simple procedure is recommended with:

- $\omega_i^{(HD)}$ household design weight
- $\omega_i^{(HN)}$ the weight determined after adjustment (non-response or calibration)
- $\overline{\omega}^{(HD)}, \overline{\omega}^{(HN)}$ their respective mean values

any computed non-response weights outside the following limits are recoded to the boundary of these limits:

$$1/C \le \frac{\omega_i^{(HN)} / \overline{\omega}^{(HN)}}{\omega_i^{(HD)} / \overline{\omega}^{(HD)}} \le C .$$

A reasonable value for the parameter is C=3.

Since trimming alters the mean value of the weights, the above adjustment may be applied iteratively, with the mean re-determined after each cycle. A very small number of cycles should suffice normally.

## 3. COMPUTATION OF BASE WEIGHTS

The aim is to produce a set of sample weights for each panel independently. At wave t = 1, we define the "base" weight as: $\omega_1^{(B)} = \omega^{(RC)}$ where $\omega^{(RC)}$ designates the individual subsample weight, calculated on the basis of the procedure outlined at the previous section.

In order to determine base weight $\omega_t^{(B)}$ from known $\omega_{t-1}^{(B)}$ (t ≥ 2), we use the following procedure. Consider the set of persons enumerated at (t-1) who are still in-scope at t. For each person j in this set, we can define a binary variable $r_j$:

- $r_j = 1$ if the person is successfully enumerated at t.
- $r_j = 0$ otherwise, i.e. the person is not successfully enumerated at t.

Using a logit model, for instance, we can determine the response propensity $p_j$ of each person in the above set as a function of a vector of auxiliary variables $V_j$:

$$p_j = Pr(R_j = 1 \,|\, V_j)$$

where $R_j$ is a random indicator of response, whose realisation in $r_j$.

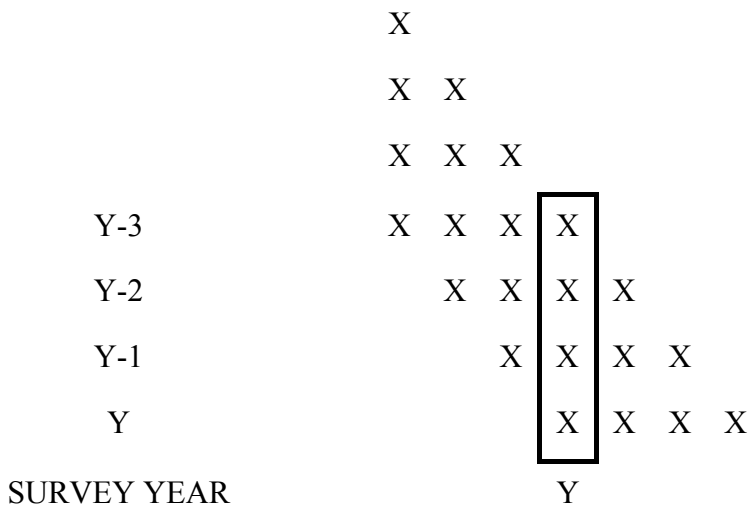
Hence, for any person j with ( $r_j = 1$ ) the required base weight is: $\omega_{t,j}^{(B)} = \frac{\omega_{t-1,j}^{(B)}}{p_j}$.

In so far as most non-response occurs at the household level, a majority of the relevant auxiliary variables ($V_j$) will be geographical and household level variables (region, household size and type, tenure) and also constructed variables (household income, household work status …).

Some personal variables are also likely to be useful (gender, age, employment status…). The main difference from similar adjustment for non-response at wave 1 is that a great deal is known about non-respondents at subsequent waves, in so far as those persons have already been enumerated before.

# 4. CROSS-SECTIONAL WEIGHTS, YEAR 2 ONWARDS

Figure 2: Representation of the cross-sectional sample




The following specifies the sample weights and the cross-sectional population estimated by the various panels.

Table 2: Inference populations estimated by each panel at year Y

| Panel introduced in year | Sample and weight | Population |
|---|---|---|
| *Y* | $(s_{1,}\omega_1^{(B)})$ | $P_Y$ |
| *Y-1* | $(s_{2,}\omega_2^{(B)})$ | $P_Y - IN_Y^{(new)}$ |
| *Y-2* | $(s_{3,}\omega_3^{(B)})$ | $P_Y - (IN_Y^{(new)} + IN_{Y-1}^{(new)})$ |
| *Y-3* | $(s_{4,}\omega_4^{(B)})$ | $P_Y - (IN_Y^{(new)} + IN_{Y-1}^{(new)} + IN_{Y-2}^{(new)})$ |

- $P_Y$ is the target cross-sectional population at Y.
- $IN_Y^{(new)}$ is the population entering the target population during the year preceding Y.
- $s_k$ is the panel at k-th year.
- $\omega_k^{(B)}$ is the corresponding base weight at k-th year of the specified panel.

To put the four cross-sections together we first start dividing the base weights as follows:

- $P_Y - (IN_Y^{(new)} + IN_{Y-1}^{(new)} + IN_{Y-2}^{(new)})$ by 4
- $IN_{Y-2}^{(new)}$ by 3
- $IN_{Y-1}^{(new)}$ by 2
- $IN_Y^{(new)}$ by 1

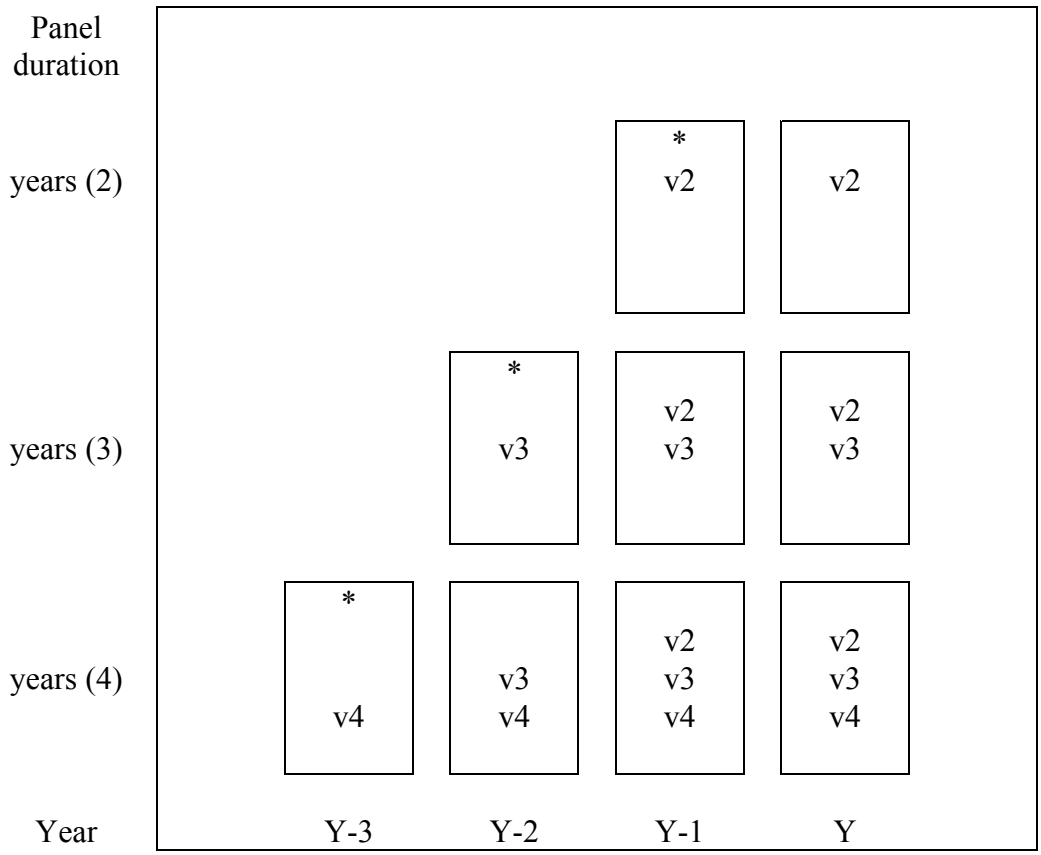
This specific treatment for "immigrants" needs to trace them out. Let $\omega_j$ be the weight of unit j after the above mentioned modification. Each household member j has been assigned a weight $\omega_j$, except for "co-residents" (i.e. the people not belonging to the panel) for whom $\omega_j$=0. Average of these weights over all household members is then assigned to each member, including co-residents.

# 5. LONGITUDINAL WEIGHTS

## 5.1. Description of the longitudinal samples

Consider the longitudinal data set delivered each year, after EU-SILC year 2, when the normal rotational system has been established. The set consists of three panels of duration 2, 3 and 4 years as shown below. We will refer to each panel by its current duration.

Figure 3: Representation of the longitudinal samples



* Panel selected. Each square represents an annual data set. V2-V4: longitudinal variables to be defined

If Y is the most recent year for which the data are included in the longitudinal data set, panels 2, 3 and 4 were selected, respectively, in years (Y-1), (Y-2) and (Y-3). These are three longitudinal data sets of different durations which are of interest:

- Longitudinal set of two year duration, involving annual data from year (Y-1) and Y. All the three panels 2, 3 and 4 contribute to this set. In the above figure, V2 stands for the required longitudinal weight to be used in the analysis of these data. The diagram also shows the annual data sets for which this variable is required.
- Longitudinal sets of three year duration, involving annual data from years (Y-2) to Y. Panels 3 and 4 contribute to this set. V3 is the required longitudinal weight for the analysis of this set. The annual data sets for which this variable is required is shown in the diagram.
- Longitudinal set of four year duration. Only panel 4 with data from years (Y-3) to Y contributes to this set. V4 is the required longitudinal weight for its analysis.

There are also other sequences of longitudinal data embedded in the data set shown in the diagram: the 3 year longitudinal sample from (Y-3) to (Y-1) in panel 4; and three 2 year samples (Y-3) to (Y-2) in panel 4, and (Y-2) to (Y-1) in panels 3 and 4.

Looking at the components of longitudinal samples (1), (2) and (3) defined above, two types can be identified:

A. Panels starting from their time of selection (t=1):

A.1: a 2 year longitudinal sample of panel 2, covering years (Y-1) to Y
A.2: a 3 year longitudinal sample of panel 3, covering years (Y-2) to Y
A.3: a 4 year longitudinal sample of panel 4, covering years (Y-3) to Y

B. Panels which are included from a later time (t>1):

B.1: a 2 year longitudinal sample from panel 3, covering years (Y-1) to Y
B.2: a 2 year longitudinal sample from panel 4, covering years (Y-1) to Y
B.3: a 3 year longitudinal sample from panel 4, covering years (Y-2) to Y.

## 5.2. Construction of longitudinal weights

In all cases of type A above, the weights involved are also identical to base weights defined earlier. We may write this as:

- $\omega^{(A1)} = \omega_2^{(B)}$, for a unit in panel A.1
- $\omega^{(A2)} = \omega_3^{(B)}$, for a unit in panel A.2
- $\omega^{(A3)} = \omega_4^{(B)}$, for a unit in panel A.3

The left hand side represents the longitudinal weight, with the superscript (A1) etc. while the right hand side specifies the base weight for the unit, the subscript indicating the wave concerned. For instance for a unit in A.1, the reference is to its base weight in wave t=2.

Let us consider now the three longitudinal data sets of durations 2, 3 and 4 years defined in the first paragraph.

*1. Longitudinal set of two year duration, for the most recent period (Y-1) to Y*

| Sample from panel | weight | population not represented * |
|---|---|---|
| (2) | $\omega_2^{(B)}$ | - |
| (3) | $\omega_3^{(B)}$ | $IN_{Y-1}^{(new)}$ |
| (4) | $\omega_4^{(B)}$ | $IN_{Y-1}^{(new)} + IN_{Y-2}^{(new)}$ |

* IN : entrants in the year preceding Y, forming separate households.

To ensure proper representation of the special groups identified in the last column, we firstly multiply the weights assigned to cases in:

- $IN_{Y-2}^{(new)}$ by 3
- $IN_{Y-1}^{(new)}$ by 3/2.

Then the required target variables can be computed as follows:

$$V2_j = \omega_j / 3$$

where $\omega$ is the weight for any unit j as defined above.

*2. Longitudinal set of three years duration, for (Y-2) to Y*

| Sample from panel | weight | population not represented * |
|---|---|---|
| (3) | $\omega_3^{(B)}$ | |
| (4) | $\omega_4^{(B)}$ | $IN_{Y-2}^{(new)}$ |

After multiplying the weights assigned to cases in $IN_{Y-2}^{(new)}$ by 2 and the required target variable for all the longitudinal units of interest can be computed as:

$$V3_j = \omega_j / 2$$