SAMPLING ISSUES IN EU-SILC SURVEYS

Vijay Verma, Gianni Betti

Working Paper n. 45, June 2004

SAMPLING ISSUES IN EU-SILC SURVEYS

Vijay Verma, Gianni Betti[1]

ABSTRACT

EU-SILC is the major new source of comparative statistics on income and living conditions in Member States of the European Union and some neighboring countries. As noted in the Regulation on the Collection of Statistics on Income and Living Conditions in the Community, EU-SILC has been developed as a flexible yet comparable instrument covering data and data sources of various types: cross-sectional and longitudinal; household-level and person-level; economic and social; from registers and interview surveys; from new and existing national sources. It envisages the creation of one or more micro-data base(s) in each country, to be used for the follow-up and monitoring of poverty and social exclusion at the EU and national levels. The present paper elucidates the structure and main characteristics of the EU-SILC surveys and the various technical considerations involved in the design and implementation of samples for EU-SILC. Many features of the EU-SILC survey structure are in fact based on recommendations originally made to Eurostat by one of the present authors. Noting the diversity of arrangements possible under EU-SILC, the paper presents a typology of EU-SILC data sources, taking into account the distinction between its cross-sectional and longitudinal components, as well as between income-related and non-income 'social' variables. Essential features of cross-sectional and longitudinal sample design and selection, including tracing rules involved in the follow-up over time of households and persons in the longitudinal sample, are discussed. Minimum effective sample sizes for country EU-SILC surveys have been specified in relevant European Commission Regulations; this paper explains the logic underlying these specifications. Finally, we discuss the standard 'integrated design', which has been recommended by Eurostat for countries starting new EU-SILC surveys, and has in fact adopted by a majority of the countries so far.

[1] V. Verma, e-mail: verma@unisi.it (corresponding author); G. Betti e-mail: betti2@unisi.it. Department of Quantitative Methods, University of Siena, P.za S. Francesco, 7, 53100, Siena, Italy.

1. Introduction

The European Community Household Panel (ECHP) has proven to be an extremely successful undertaking, as for instance demonstrated by the very large volume of policy and academic research published using ECHP data. It indeed became *the* reference source of comparable statistics on income and living conditions in the EU for the 1990s.

The ECHP was a standardised survey co-ordinated and supported by Eurostat. It began in 1994 in 12 Member States of the then European Union and subsequently expanded to include the additional members of EU-15. The survey involved annual interviewing of a representative panel of households and persons in each country covering a wide range of topics concerning income and living conditions. Its distinguishing characteristics included: a multi-dimensional coverage; a household panel design in which starting from an initial sample, households and individuals are followed-up over time; and a strong dimension of cross-national comparability (Verma and Clemenceau, 1996). With the development and dissemination of standardised microdata in the form of a users' data base, accessibility and extensive use of its data became another characteristic feature of the ECHP.

ECHP data collection was concluded in 2001 and it was decided develop a new instrument to replace the ECHP. This new instrument is called EU Statistics of Income and Living Conditions (EU-SILC). The rationale underlying the choice of the structure and design of this new instrument is the subject of this paper.

EU-SILC has been developed to overcome, or at least ameliorate the main shortcomings of ECHP. These included the following: the problem of sample attrition, loss of representativeness, and excessive respondent burden with increasing duration of the panel in the absence of any renewal of its sample; lack of flexibility in the design and content of the survey (again because of its panel structure) in response of changing needs and priorities; a certain lack of timeliness in the production of the data; and also a lack of sustainability of the survey in its present form for various institutional reasons.

EU-SILC represents a more flexible and sustainable alternative replacing the ECHP. Its objectives and contents are very similar to ECHP, but the context and structure differ. On the basis of common substantive requirements and technical standard, EU-SILC aims to utilise, in the interest of national circumstances and preferences, and above all in the interest of economy, diverse structures and data sources.

As noted in EU-SILC Regulation (Official Journal of the European Union, 2003a), depending on the country, micro-data could come from:

(1) one existing national source (survey or register);

(2) two or more existing national sources (surveys and/or registers) directly linkable at micro-level;

(3) one or more existing national sources combined with a new survey – all of them directly linkable at micro-level;

(4) a *new harmonised survey* (or survey system) to meet all EU-SILC requirements.

Indeed, EU-SILC aims to be a flexible yet comparable instrument covering data and data sources of various types: cross-sectional and longitudinal; household-level and person-level; economic and social; from registers and interview surveys; from new and existing national surveys or other sources.

Following pilot surveys in 2003, full-scale EU-SILC surveys were conducted in 15 countries in 2004, and 25 in 2005. Therefore, the number is expected to reach around 30 countries, including all EU Member States.

The present paper elucidates the structure and main characteristics of the EU-SILC surveys and various technical considerations involved in the design and implementation of samples for EU-SILC. Many features of the EU-SILC survey structure are in fact based on recommendations originally made by one of the present authors (Verma, 2001). Noting the diversity of arrangements possible under EU-SILC, the paper presents in Section 2 a typology of EU-SILC data sources, taking into account the distinction between its cross-sectional and longitudinal components, as well as between income-related and non-income 'social' variables.

Essential features of the cross-sectional and longitudinal sample design and selection, including tracing rules involved in the follow-up of households and persons over time in the longitudinal sample, are discussed in Section 3. Minimum effective sample sizes for country EU-SILC surveys have been specified in relevant European Commission Regulations; Section 4 explains the logic underlying these specifications. Finally, in Section 5 we explain the standard 'integrated design', which has been recommended by Eurostat for countries starting new EU-SILC surveys, and has in fact adopted by a majority of the countries so far.

2. EU-SILC Data Sources: A Typology

Most micro-level income statistics are confined to the population living in private households. This was the case with ECHP, and is so for EU-SILC. Excluded from the target population of private households are all persons living in collective households or in institutions on a permanent or long-term basis, and persons temporarily in collective households or institutionalised but not included as members of any private household on the basis of certain specified criteria.

Households form the basic units of sampling, data collection and data analysis. It is important to clearly define and consistently implement criteria for the grouping of individuals into households. This requirement is common to all surveys using households (or other such units) for sampling, so as to ensure that individuals in the population of interest are correctly covered in the survey, without omission or double-counting.

For income (as well as household budget and similar) surveys, this requirement is doubly important: how individuals are grouped into households directly determines the statistics which are measured and analysed. Given the specific purpose that EU-SILC *is the reference source of comparative income distribution statistics of households and persons at EU level*, a rigorous and harmonised definition of the household is necessary for all countries to ensure comparability of key indicators.[2]

The household is the conventional unit *defining* income, and is the recommended choice for this purpose in EU income analysis (Eurostat, 2004b). While most of the information on income may pertain to and be collected directly from individual persons, it is only at the level of the household that analytically meaningful variables on income can be constructed. True, the individual person rather than the household is often the preferred *unit of analysis* in the construction of income distribution and related statistics; however this is on the basis of household level measures ascribed to individuals on the basis of their membership of the household.

2.1. The substantive dimension: income versus 'social' variables

In terms of the substantive content, four types of data are involved in EU-SILC: (i) variables measured at the household level; (ii) information on household size and composition and basic characteristics of household members; (iii) income and other more complex variables measured at the personal level, but aggregated to construct household-level variables (which may then be ascribe to each member for analysis); and (iv) more complex non-income or 'social' variables collected and analysed at the person-level.

For set (i)-(iii) variables, a sample of households including all household members is required.

Households and household members

Among these, sets (i) and (ii) are normally collected from a single, appropriately designated respondent in each sample household – using a household questionnaire for set (i) and a household member roster for set (ii). Alternatively, some or all of these data may be compiled from registers or other administrative sources.

[2] In the standard EU-SILC definition, a private household means a person living alone or a group of people who live together in the same private dwelling and share expenditures, including the joint provision of the essentials of living. This general definition is supplemented by rules for treating particular categories of persons concerning household membership (Official Journal of the European Union, 2003a).

Household and personal income

Set (iii) – concerning mainly, but not exclusively, the detailed collection of household and personal income – must be collected directly at the person level, covering all persons in each sample household. Generally, these income and related variables are collected through personal interviews with all adults aged 16+ in each sample household. This collection will be normally combined with that for set (iv) variables, since the latter also must also be collected directly at the person level. These are the so-called 'survey countries'.

By contrast, in 'register countries', set (iii) variables are compiled from registers and other administrative sources, thus avoiding the need to interview all members (adults aged 16+) in each sample household for the purpose of collecting the income variables.

Personal 'social' variables

Set (iv) variables are normally collected through direct personal interview in all countries. These are too complex or personal in nature to be collected well by proxy, nor are they available from registers or other administrative sources. For the 'survey countries', this collection is normally combined with that for set (iii) variables as noted above. Consequently both are normally based on a sample of complete households, i.e. covering all persons aged 16+ in each sample household.

However, from the substantive requirements of EU-SILC, it is *not essential* that – in contrast to set (iii) variables – set (iv) variables be collected for all persons in each sample household. It is possible to do this collection on a representative sample of persons (adults aged 16+), such as by selecting one such person per sample household. This option is normally followed in 'register countries', since for these countries interviewing all household members for set (iii) is not involved. In countries which choose to do so, the sampling process involved will be the selection of *persons* (usually one adult member aged 16+ per household) directly, or optionally through a sample of households.

2.2. The temporal dimension: cross-sectional versus longitudinal data

Both cross-sectional and longitudinal data are required in EU-SILC. The cross-sectional component covers information pertaining to the current and a recent period such as the preceding calendar year. It aims to provide estimates of cross-sectional levels as well as estimates of net change from one period (year) to another. The longitudinal component covers information compiled or collected through repeated enumeration of individual units, and then linked over time at the micro-level. It aims at measuring gross (micro-level) change and elucidating the dynamic processes of social exclusion and poverty.

Both cross-sectional and longitudinal micro-data sets are updated on an annual basis. However, the first and clear priority is given to the production of comparable, timely and high quality *cross-sectional* data. Longitudinal data are limited in content and possibly also in sample size. Furthermore, for any given set of individuals, micro-level change is followed up only for a limited duration. A period of four years is taken as the minimum duration for longitudinal follow-up at micro level in EU-SILC.

The cross-sectional and longitudinal data can come from separate sources, i.e., the longitudinal dataset does not need to be "linkable" with the cross-sectional dataset at the micro-level, though such linkage is not precluded and would normally be possible when the two types of data come from the same source.

Sample rotation over time

This refers to the relationship between annual samples. The annual *cross-sectional component* can be based on independent samples, a rotational sample, or a long-term panel.

A rotational sample design appears to be the most suitable option for a cross-sectional survey, without precluding the other options in particular circumstances. Independent samples reduce respondent burden and permit more efficient cumulation of the data over time, if required; in certain circumstances they may also be easier to control and implement, and result in somewhat better response rates. However, their major drawback is the reduced efficiency in measuring trends or net change over time, which is an important consideration in EU-SILC. Fieldwork costs are also likely to be higher than repeated use of the same units. (See Section 5.2 for further details).

A long-term household panel can also yield cross-sectional estimates, as has been done in the case of the ECHP. The major drawback is the loss of cross-sectional representativeness due to cumulative sample attrition over time. Costs and complexity are also increased, often substantially.

For the annual *longitudinal component* the samples must of course be related over time, allowing only the last two options: a rotational sample, or a long-term panel.

Since flexibility is an essential feature of EU-SILC, the country datasets may comprise different types and combinations of data sources, with different designs. The next section describes a *typology of the structure and design* of EU-SILC data sources, concentrating on aspects pertaining to sampling. The development of such a typology is helpful in a number of ways:

(1) It clarifies the type of data-source structures which are possible and acceptable, given the EU-SILC objectives.
(2) Within each type of arrangement, it helps in identifying the basic design choices and providing a framework for their systematic elaboration.
(3) It also facilitates the documentation of the EU-SILC methodology in each country.

2.3. Defining basic components of EU-SILC

It is useful to summarise the essential features of EU-SILC data requirements as they determine the types of survey and sample structures permissible.

As noted, in each country EU-SILC involves the provision of cross-sectional and longitudinal data, each for income and social target variables as defined above. Combining these dimensions gives four basic data components in EU-SILC:

- (CI) Cross-sectional income component
- (CS) Cross-sectional social component
- (LI) Longitudinal income component
- (LS) Longitudinal social component

Substantive requirements of EU-SILC impose certain conditions on the samples for these components.

Firstly, as noted above, total income of the household can be collected only by obtaining detailed information on the income of each household member. This means that the income components, whether cross-sectional or longitudinal (CI, LI), must be based on a sample of *households and all household members potentially receiving an income* (all members aged 16+).

Secondly, while the social component may normally be collected on the same sample as the corresponding income component (i.e., CS=CI, LS=LI), this is not a *requirement* in EU-SILC. The social component must be included in the income component (it makes no sense to collect social information on units without collecting information on their income), but it can be a subsample of the latter in two respects: (i) it may be applied to a subsample of one or more persons in each sample household; and/or (ii) only to a subsample of households to which the income component has been applied.

Another point concerns the required micro-level linkage in EU-SILC data. It is required that all data (household, income, social) be 'linkable' at the micro-level for the cross-sectional component; and also separately for the longitudinal component. Linkage between the cross-sectional and longitudinal components is not included in the minimum EU-SILC requirements.

On the basis of the above requirements, the basic (essential, minimum) condition which must be satisfied by any data structure in EU-SILC can be expressed as:

$$\boxed{\begin{array}{l}(a)\,CS \subseteq CI\\(b)\,LS \subseteq LI\end{array}} \qquad \textit{… the basic condition of EU-SILC data structure.}$$

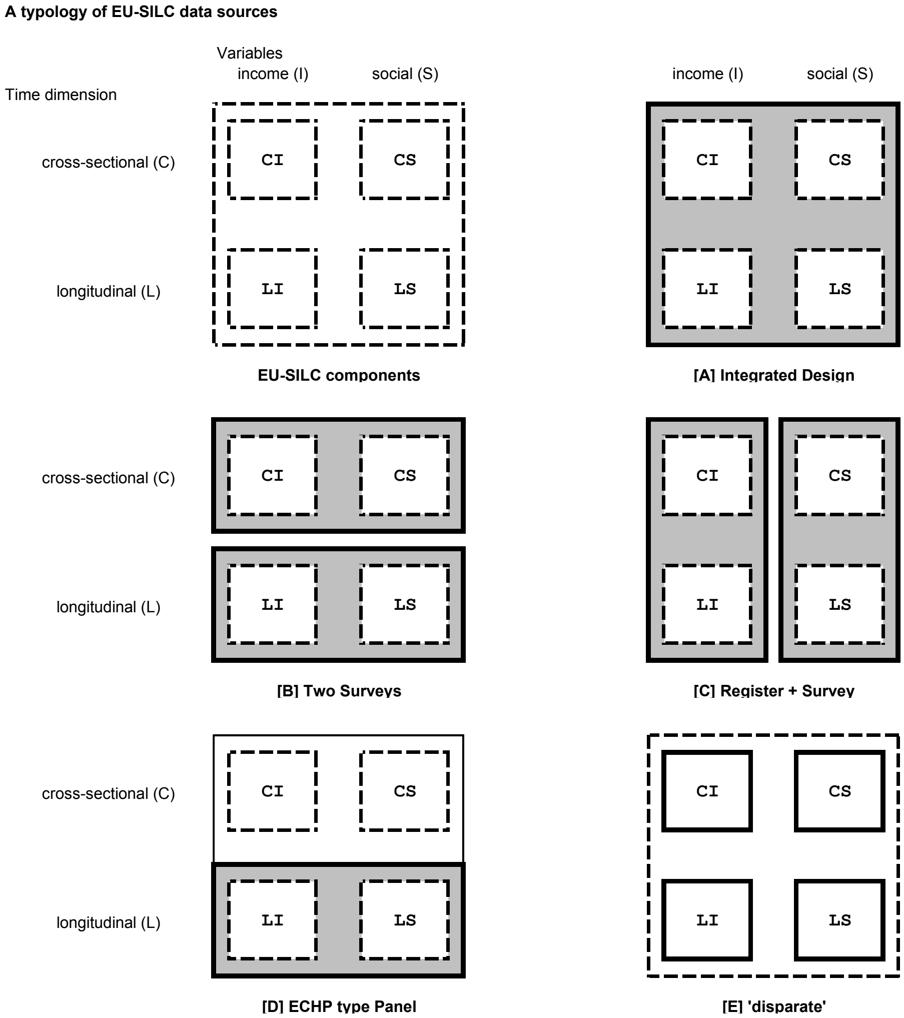
The basic condition means that the social data must be collected on the same sample as the income data, or on a subsample of the latter. The condition applies separately to both the cross-sectional and the longitudinal components.

Such a structure ensures that social data are linked to income data at the micro-level: CS to CI for the cross-sectional component; and LS to LI for the longitudinal component. However, no such relationship is mandatory between cross-sectional and longitudinal data. The two types of data can come from separate sources and need not be linked at the micro-level. Of course, as already noted, such linkage is not precluded, and would normally be possible when the two types of data come from the same source.

Different types of data structures (and the related sampling arrangements) are possible in EU-SILC depending on how the four components (CI, CS, and LI, LS) are combined.

Figure 1 illustrates a typology of possible data source structures. Arrangement [A], the integrated design, is by far the most common one adopted by countries up to now. A brief description of each type of design follows.

Figure 1. EU-SILC data source structures



Design [A] a single *integrated* source covering all components – cross-sectional and longitudinal, income and social.

This is the most suitable design in situations where a new survey is to be developed for EU-SILC. This design is described in detail in Section 5.

The basic idea is as follows. At any one time, the sample is made up of, say, 4 short-term panels or subsamples. Each year one new subsample is added to stay in the survey for 4 years, and then dropped to be replaced by another new subsample. Each susample provides a longitudinal sample of the chosen duration (say, 4 years). Movers from the original sample are followed up to their new location for up to the time the subsample remains in the survey.

The units present at a given time from all the subsamples are appropriately put together to constitute the cross-sectional sample.

An important advantage of this scheme is that both cross-sectional and longitudinal data are obtained from the same common set of units. This overlap is highly economical, and also maximises internal consistency between longitudinal and cross-sectional statistics produced from the survey.

Design [B] two separate surveys, one cross-sectional and the other longitudinal, each covering both income and social variables:

$$\boxed{C = (CI + CS); L = (LI + LS)}.$$

The 'basic condition of EU-SILC data structure' is automatically satisfied, and it is not necessary (though possible, and may be desirable) to have micro-level linkage between the two surveys (C and L).

This scheme may be used when one, or the other, or both surveys already exist in the country and can be adapted to meet EU-SILC requirements. Some countries may prefer to adopt this scheme, even when new surveys are being established, for the sake of simplicity and flexibility of designing the cross-sectional and longitudinal components separately.

The cross-sectional survey (C) may consist of independent (non-overlapping) annual samples, or may be rotational as in scheme [A]. However, the longitudinal sample (L), *in so far as it remains cross-sectionally representative*, can provide more precise estimates not only of gross (micro-level) change but also of net (macro-level) change. Hence, sample overlaps in the cross-sectional sample (useful for the measurement of net change as in scheme [A]) are not so critical.

The longitudinal survey (L) may be designed with a limited panel duration (as in scheme [A]), or may be a 'true' long-term panel (as in scheme [D] described below). In the latter case, periodic addition of new sample supplements may be required to retain cross-sectional representativeness (but less critically than in scheme [D] below).

A major concern with this design is the *consistency* between the cross-sectional and longitudinal sources for the same set of variables. Survey L also can be (and often is) used to produce cross-sectional estimates, which may differ from similar estimates produced from survey C. Procedures will have to be developed to make them consistent to the extent possible.

Design [C] two separate sources, one covering income variables and the other covering social variables, each comprising both cross-sectional and longitudinal components as in design [A]:

$$\boxed{I = (CI + LI);\ S = (CS + LS)}$$

Normally, the *income source* (I) will be based on registers – though in principle, it is also possible for it to be an interview survey, for instance an existing income survey with cross-sectional and longitudinal components. The *social survey* (S) almost always will be an interview survey.

The idea of this design is that for the income and the social parts, different sources may be used, possibly sources of different types. Note, however, that the 'basic condition of EU-SILC data structure' requires that S must be a subsample of I – micro-level linkage between social and income data is required for all units in S. This design differs from design [A] mainly in that sample I, so long as it fully incorporates the sample of S, can be based on a larger sample, for instance when income data are needed with higher precision. This option can be attractive in situations where income data can be collected more cheaply, such as from registers rather than from an interview survey.

Design [D] a single ECHP-type panel survey, providing all cross-sectional and longitudinal data, but primarily focused on the latter.

At least in technical terms, a panel survey of the ECHP-type with similar tracing rules and procedures can be designed to meet all the EU-SILC requirements.

The problems to be faced with this choice are practical rather than of principle: sample attrition and its effect, in particular on cross-sectional estimates; the lack of timeliness, again especially important in the case of producing cross-sectional estimates; difficulties in changing the size or content of the survey to reflect changing circumstances and needs.

None of these *preclude* the use of this type of design for EU-SILC in particular circumstance where steps can be taken to reduce practical problems of the above mentioned types. However, it has to be noted that while ECHP covered both cross-sectional and longitudinal aspects, its design was primarily determined by the *longitudinal* requirements. EU-SILC is also to cover both aspects. However, by contrast, its focus is on the *cross-sectional* aspects. In this sense, an ECHP-type long-term panel is not the optimal choice for EU-SILC.

Design [E] a separate source/arrangement for each component, CI, CS, LI and LS, or some combination of these not covered above.

It is possible to have different sources for each component of EU-SILC. This may arise if diverse sources already exist which can be adapted to meet the various EU-SILC requirements.

In using diverse sources, it is necessary to ensure that the micro-level linkage specified in the ‘basic condition of EU-SILC data structure’ is met, i.e. the income variables are collected at least for all the units for which the social variables are to be collected. This condition must be met separately for both the cross-sectional and the longitudinal components.

The second most important requirement is consistency (in terms of content, coverage, timing, etc…) of the data coming from different sources.

3. The cross-sectional and the longitudinal sample

3.1. Constructing a sample of households

As introduced in Section 2, there are four types of analysis units related to the sets of variables: (i) variables measured at the household level; (ii) information on household size and composition and basic characteristics of household members; (iii) income and other more complex variables measured at the personal level, but aggregated to construct household-level variables; and (iv) social variables collected and analysed at the person-level. Variables (iv) are confined to adult persons (taken as those aged 16+ in EU-SILC). An option exists between two scenarios here: sample comprising all adults in each sample household; or a direct sample of adults, normally no more than one selected per household.

For analysis units in sets (i)-(iii), as well in the first scenario of set (iv), the selection probability of all types of units is the same as that of their household. Hence the various analysis requirements are served best by having basically an equal probability ('self-weighting') sample of households. Variations in selection probabilities – by region, household size, or whatever – mostly results in reduced sampling efficiency for national-level estimates.

However, for the same reason, the second scenario of set (iv) with direct sampling of persons, it is desirable to aim at an equal probability sample of those persons, rather than an equal probability sample of households.

Consequently, in the situation pertaining to a majority of the countries (the 'survey countries'), the EU-SILC primary objective is served best by having an equal probability sample of households, taking all persons in a selected household into the sample for the personal interview, and hence obtaining an equal probability sample of persons as well. By contrast, in the ‘register countries’ it is more efficient to aim at a self-weighting sample of persons to be interviewed in detail on non-income variables; household variables and personal income data then may not be based on a equal probability sample, but that is less critical.

In any case, irrespective of details of the actual sampling procedure, each round of EU-SILC has to be based on a probability sample of households. This is obviously the case for the annual cross-sectional survey. This applies equally to the initial sample of each panel in EU-SILC longitudinal component; the longitudinal component will consist of selecting an initial sample of households, then following up (all or a random subsample of) individuals in that sample annually over time.

In different situations, the actual selection mechanism may vary. The ultimate sampling units may be addresses, households or persons. However, through rules of association between such units, the basic objective in all cases has to be the obtaining of a *valid sample of households*.

Sample of households from selection of addresses

Constructing a sample of households from a selected sample of addresses or similar units is normally straightforward. In most cases, there is in fact a one-to-one correspondence between the two types of units, so that, for instance, an equal probability sample of households is obtained by taking an equal probability sample of addresses. Exactly the same applies when all households are taken into the sample from any selected addresses containing more than one households.

However, in the presence of some addresses containing large numbers of households, it is desirable for technical and practical reasons to limit the maximum number of households which will be taken for the survey from any one selected address. This changes the probabilities of selection of the households. Procedures can be easily developed to control such variations in household selection probabilities.

Sample of households from selection of persons

Now consider the situation when a sample of households is constructed from a sample with individual persons as the selection units. A household is selected through its association with one or more individuals. Normally, the latter is selected from a list of adults. In so far as each eligible household contains at least one such person in the list, the household receives a non-zero probability of selection; consequently, a probability sample of persons will yield a probability sample of households. This will be the case for instance if the sampling frame (list) covers all persons aged 16+, in so far as it can be assumed that practically every household contains at least one adult. However, this might not be so in the case of certain other types of lists of persons. For instance, if an electoral roll is used as the frame, only those households which contain at least one eligible and registered voter will have any chance of being in the sample. The population of households not containing such a persons will not be covered in the survey.

In any case, the main consideration in constructing a sample of households through the selection of persons is that the selection probability of a household would vary in direct proportion to the number of persons in the list through which the household could have been selected. If, for instance, persons are selected with equal probabilities, larger households will be selected with higher probabilities. These differences in household selection probabilities have to be compensated for by applying weights to the data.

When the units of analysis are households, or all persons or all adults in the household, such departures from self-weighting tend to reduce efficiency of the sample (i.e. increase sampling variance). With essentially random weights i.e. weights not significantly correlated with target variables (as is likely to be the case in the situation being discussed), the increase in variance depends on the variability in the selection probabilities or the resulting design weights. This increase can be approximated as:

$$d_w^2 = \left(1 + cv^2\left(w_i\right)\right) \tag{1}$$

where *cv* is the coefficient of variation of the weights. The expression approximates the factor by which sampling variances are inflated, i.e. the effective sample size is reduced. This effect applies more or less uniformly across all types of estimates from the survey.

To reduce this loss in the efficiency of the sample of households (or of members or adults in each household when all of them are taken into the sample), it is desirable to vary the selection probabilities in the original sample of persons in inverse proportion to the person's household size (i.e. the number of such persons in the household). Such information may be available in the lists, registers or other administrative sources used as the sampling frame.

Alternatively, one may select a sample of persons larger than that ultimately required, compile or collect the necessary information on household size for the selected persons, and then subsample the initial sample in such a manner that the retention probabilities for the individuals finally in the sample vary in inverse proportion to the person's household size. The households of all persons finally in the sample would constitute an equal probability sample of households for the survey. Taking all members of these households would provide an equal probability sample of persons.

The increased efficiency obtained with the above procedure has to be balanced against the added cost of collecting the information required for its implementation. The cost may be small when the required information can be compiled from registers or other existing sources. However, it may be substantial if the information has to be collected in the field for the large initial sample.

The above considerations apply to all 'survey countries' in EU-SILC. However, the situation is the opposite in 'register countries', where the aim is to get an equal probability sample of adults, with one adult selected per household. Here it is preferable to have a sample of households in which the household selection probability varies in direct proportion to the number of adults in the household. That is achieved most conveniently if originally we select an equal probability sample of adults from a list of persons, enumerate the household associated with each selected person for household-level and income variables, and then select one adult per household for the personal interview survey on non-income variables.

Usually, the person selected for the final survey is the same as the one originally selected to bring the household into the sample. However, this need not be the case; a different individual from the household can be selected, so long as proper randomised procedures are followed at each stage.

3.2. Constructing the sample of persons

In the situation where no subsampling of persons within households is involved, the sample of households automatically gives a sample of persons, each receiving the same probability of selection as the household. As noted previously, the analysis requirements are served best by having basically an equal probability sample of households, and hence automatically of persons.

However, when the main personal interview requires a sample of persons, usually one person per household, it is desirable to aim at an equal probability sample of these units. This requirement conflicts with that of having a self-weighting sample of households, which is useful for collecting household data and income data from all household members. For instance, if one sample person is selected from each sample household, the probability of selection of the former in relation to that of the household is reduced in proportion to the number of eligible persons in the household.

Table 1. Constructing a sample of households and persons

Survey country (personal interview with all adults in household)

Original selection	**Household sample**	**Subsampling (if any)**		**Household and personal interview**
Equal probability Sample of addresses or households	Equal probability sample of households	→	1	Equal probability sample of households and persons
Equal probability sample of adults	Household selection probability proportional to size (adults in household)	→	[2]	Household and personal selection probability proportional to size
		Subsampling of households with probability inversely proportional to size	3	Equal probability sample of households and persons

Register country (personal interview with one adult per household)

Original selection	**Household and income data**	**Subsampling (if any)**		**Personal interview (social data)**
Equal probability Sample of addresses or households	Equal probability sample of households	→	[1]	Personal selection probability inversely proportional to household size
		Subsampling of households with probability directly proportional to size	2	Equal probability sample of adults
Equal probability sample of adults	Household selection probability proportional to size (adults in household)	→	3	Equal probability sample of adults

Option numbers in parentheses [ ] indicates options which are not desirable because of lost sampling efficiency.

Table 1 illustrates various schemes for the construction samples of households and persons with desirable properties in various situations. The first panel deals with 'survey countries', where all adults in each sample household are eligible for the personal interview. This was the situation in the case of the ECHP, and is for a majority of the countries in EU-SILC. Options 1 and 3 yield samples with desired characteristics. The first option has been followed in most EU-SILC (and also most ECHP) surveys. Where option 1 is not possible (e.g. due to the structure of the available sampling frame), option 3 can be used to obtain similar results. (We believe a variant was followed in ECHP Denmark).

Option [2] results in loss of sampling efficiency. This option should be avoided, though we have found two examples of its use in ECHP/EU-SILC.

The second panel of Table 1 considers the situation in 'register countries' where the personal interview is conducted with one selected adult per household. Option [1] results in loss of sampling efficiency, while options 2 and 3 yield samples with desired characteristics for the personal interview. Option 2 is optimal also for household and income data, but it assumes that the sample for these is larger than the number of households from which adults for the personal interview are selected – which may not be a major problem if these data can be compiled from registers without involving personal interviewing. Option 3 is not optimal for household and income data. However, in practice this option has been the most widely used one, given the low per unit cost of compiling such data from registers.

3.3. Tracing rules over time

Another important aspects in the definition of the statistical units in EU-SILC is the tracing of units over time. The tracing rules followed in the ECHP have formed the basis of the rules for the longitudinal component of EU-SILC. However, some changes have been introduced for practical reasons such as the following:

O Some simplifications were considered desirable on the basis of practical experience of the ECHP, for instance where certain requirements or rules proved unworkable or largely ineffective.

O EU-SILC requires the follow-up of individuals over only a limited duration of 4 years, unlike the indefinite duration implied in the ECHP design. This reduces the cumulative effect of some departures from what may be considered the 'ideal'.

O ECHP aimed to provide cross-sectional and longitudinal data from a single panel survey, with emphasis on the longitudinal (though in practice, much use has been made of ECHP as a source of comparable cross-sectional data). EU-SILC also aims to provide cross-sectional and longitudinal data, but with emphasis on the cross-sectional.

O The formulation of rules for EU-SILC has taken into account the expected diversity of arrangements and designs to be encountered within EU-SILC.

The considerations involved may not be identical when EU-SILC is based, for instance, on an integrated survey, or separate cross-sectional and longitudinal surveys, or when a sample of persons rather than of complete households is used for the personal interview survey.

The basic rules used in EU-SILC are as follows.

Original households:

The initial (Wave 1) sample consists of a probability sample of households in each country. All usual residents of those households who are over a certain age (stipulated to be no higher than 14 years) are initial sample persons; by contrast, in ECHP all the usual residents where sample persons, irrespective of the age. At any subsequent wave, the eligible population consists of:

O *Sample persons* i.e. all initial Wave 1 usual residents who are still alive and eligible for the EU-SILC. Any movers among these persons are followed up to their new address. Children in the original household as they reach the age of 16 become eligible for the personal interview. In this way the survey population is kept up-to-date for demographic changes except for immigrants into the original population.

O *Non-sample persons*: such persons are covered using the same procedures. These are persons who reside in the same household with one or more sample persons. However, the survey does not follow-up non-sample persons who move into households not containing any sample person.

Households covered in subsequent waves

The set of households covered in any wave consists of the following.

° Any household containing at least one sample person as defined above as a current resident. This includes newly formed households resulting from the movement of sample members since the last wave, as well as any new households added to the survey. Dropped are households which no longer contain a sample member (i.e. have become non-existent or contain only non-sample members).

° Households all of whose sample members are institutionalised, or have moved out of the EU are not followed up. This differs from the ECHP rules were those people were 'traced', thought not interviewed in detail.

° Persons moving out of the country are also dropped. In ECHP, an attempt was made to follow-up persons moving to another EU country.

° In ECHP persons moving into collective household as distinguished from an institution were each treated as a new one-person household in its own right. In EU-SILC, those persons are not followed up.

° For practical reasons, a limit is placed on the duration for which a household can remain un-interviewed for it to be retained in the sample for follow-up. This includes sample households not enumerated a single year due to the impossibility of locating its new address, lack of information on what happened to the household, or the household refusing to cooperate. Also excluded are households not contacted in the first year of the panel, or during any two consecutive years thereafter, due to inaccessibility, temporary absence, or the household's inability to respond due to incapacity or illness. The corresponding rules in ECHP were somewhat less restrictive.

4. Sample size consideration

4.1. Specified sample size requirements

The choice of sample size is a complex issue, involving compromises in several dimensions. These include: substantive requirements (scope of the information to be collected, precision requirements, required disaggregation and analyses of the results); cost constraints (budget, technical resources, response burden); and practical considerations (feasibility, sustainability, quality control, etc…).

A comparative, multi-country undertaking such as EU-SILC involves a number of additional factors. The data are needed not only for national analyses, but also for comparative analyses at the EU-level. Even for an individual country, judging its place in the collectivity of countries depends on the availability of sufficiently reliable and comparable information on all the other countries.

Yet, despite these multitude of statistical and practical considerations, one must ultimately decide on a single number as the target sample size. Let us begin by what has been agreed for EU-SILC surveys. This is summarised in Table 2 in terms of 'minimum effective sample size' requirements.[3] The reminder of this section aims to provide a clear interpretation of this table, and describe the rationale underlying this choice of sample sizes.

In EU-SILC, the distinction needs to be made, as already explained, (i) between income and complex 'social' variables, and (ii) between the cross-sectional and longitudinal components.

The income variables need to be measured on complete households, i.e. for all members of each household. The social variables may also be measured on complete households (as in the so-called 'survey countries'), or on a sample of adults (normally one per household, as in the 'registers countries'). In the former case, a sample of households is selected (the required minimum effective size of which is given in column A.1), and all persons aged 16+ are interviewed in detail covering both income and social variables. Column A.2 shows the expected number of such interviews in column A.1 households.

[3] These 'minimum effective sample size' requirements have been stipulated in the form of EU-SILC Commission Regulations (Official Journal of the European Union, 2003b), supplemented by technical descriptions in EU-SILC DOC 065 (Eurostat, 2004a), and EU-SILC Sampling Guidelines (Verma, 2001).

Table 2. Minimum effective sample sizes when (A) a sample of households, (B) a sample of persons is taken for the survey

	(A) households/addresses		**(B) persons**	
	(A.1) Households	**(A.2) Persons aged 16+ to be interviewed**	**(B.1) Households and selected respondents**	**(B.2) Persons aged 16+**
Belgium	4750	8750	6500	12000
Czech Republic	4750	10000	7500	15750
Denmark	4250	7250	5500	9500
Germany	8250	14500	11000	19250
Estonia	3500	7750	5750	12750
Greece	4750	10000	7500	15750
Spain	6500	16000	12000	29500
France	7250	13500	10250	19000
Ireland	3750	8000	6000	12750
Italy	7250	15500	11750	25000
Cyprus	3250	7500	5750	13250
Latvia	3750	8500	6500	14750
Lithuania	4000	9000	6750	15250
Luxembourg	3250	6500	5000	10000
Hungary	4750	10250	7750	16750
Malta	3000	7000	5250	12250
Netherlands	5000	8750	6500	11500
Austria	4500	8750	6500	12750
Poland	6000	15000	11250	28250
Portugal	4500	10500	8000	18750
Slovenia	3750	9000	6750	16250
Slovakia	4250	11000	8250	21250
Finland	4000	6750	5000	8500
Sweden	4500	7500	5750	9500
United Kingdom	7500	13750	10250	18750
Iceland	2250	3750	3000	5000
Norway	3750	6250	4750	8000

By contrast, column B applies to countries which chose to have a sample of individuals (one person aged 16+ per sample household) for the detailed person interview on social variables. For these, column B.1 gives the minimum effective sample size in terms of personal interviews, and hence also the number of households (for household-level data and basic data on household members) in the sample. In these countries, income data are compiled from registers and other administrative sources for all persons aged 16+ in those households; the estimated number of such persons in the households in column B1 is shown in column B2.

As to the longitudinal sample, according to EU-SILC Commission Regulations the required minimum effective sample sizes for the longitudinal component will be 75% of the corresponding cross-sectional sample sizes shown in Table 2. The same choice between columns A1 and B1 applies to the longitudinal component.

Why *minimum effective* sample size? By this is meant the required minimum sample size if the survey were based on simple random sampling (design effect = 1.0). The actual objective in EU-SILC regulations is to specify a minimum level of precision, at least for the most important estimates from the survey. Precision depends on three main factors: actual or achieved sample size (n_a), design effect ($deft^2$), and the (population) coefficient of variation (cv). Empirically, cv does not tend to be a major source of variation across EU countries.[4] In a multi-country undertaking such as EU-SILC, design effects cannot be specified at a central level since they depend on structure of the sample, which each country must be completely free to choose in the light of its own circumstances and requirements so long as common sampling standards are adhered to. Specification of the precision requirements in terms of effective sample size:

$$n_e = n_a / deft^2 \qquad (2)$$

removes the dependence on deft. Countries may choose the design, and hence the implied design effects for the variables of prime interest, and then determine the actual sample size to be achieved (n_a) so as to meet the minimum effective sample size requirement (n_e).

The *selected sample size* has to be larger still in order to take into account the loss due to non-response. The requirement is in terms of the minimum sample size needed; the *actual sample size* may be larger than this in particular countries, for instance to meet country-specific needs such as a more detailed regional breakdown. In fact, a number of countries have chosen to have effective sample sizes significantly larger than the stipulated minimums in order to meet country-specific needs.

4.2. The rationale

Now we turn to the rationale behind the particular choices of sample size in Table 2.

The cross-sectional component

In this choice, experience with similar surveys in the past - in particular the ECHP – has played a major role. ECHP experience shows that a range of 4000-6000 households can yield very useful results for the type of survey under consideration: providing sufficiently precise estimates of important variables on living conditions, income, poverty and social exclusion.

[4] For example, based on ECHP data, *cv* for the variable equivalised household income was mostly in the range 0.6 – 0.8. For statistics similar to a proportion, such as poverty rate, *cv* is of course insensitive to moderate variations in the statistic.

In fact a similar choice of sample size is indicated from the consideration of precision requirements for the most critical survey variables, as explained below.

The main household income measure for this purpose may be taken as the ***poverty rate*** (i.e. proportion of the population with equivalised income below 60% of the median).

In EU countries, based on the results of ECHP, the poverty rate is found to vary roughly in the range 5-25%. Taking poverty rate p=15% as the basis for computations, a (simple random) sample of 5,000 households is required to estimate this with *1 percentage point error*, i.e. with the 95% confidence interval as 14-16%.[5]

Another important variable is the mean equivalised income. With cv = 0.7, for instance, an effective sample size of 5,000 households gives a relative standard error of 1% in the estimated mean equivalised household income. The range of error is around 1.3%-0.7% corresponding to the effective sample size range of 3,000-10,000 households.

In the EU-wide context of EU-SILC, there are also practical constraints on the total sample size, summed over Member States. For reasons which need not be discussed here, it was decided to limit the total to around 80,000 households in EU-15, or around 120,000 households in the expanded EU.

Next issue is the allocation of the total across countries of the EU.

While different countries may require – despite differences in their population sizes - similar sample sizes for the same level of precision, there are many well-known reasons why it is meaningful and useful to have larger samples in larger countries.

The added reason for increasing the sample size with increasing population size (but of course much less than proportionately) is the requirement for reporting at the EU level. For such reporting, the ideal would be to sample at a uniform rate throughout, i.e. increase the national sample size in proportion to the population size. However, this will be unacceptable for the production of national level statistics (which require more equal sample sizes). A common and convenient compromise is to allocate the sample proportional to the square-root of the population size, modified by the imposition of minimum and maximum limits, for instance as Verma (1991):

$$n_i = n_0 . \sqrt{\left(k^2 + \left(1 - k^2\right) . M_i^{\alpha}\right)}, \tag{3}$$

where k is a parameter determined by the relative importance given to the national versus EU level estimation. M_i is a (relative) measure of the population size of

[5] For simplicity, this computation has been based on taking p as a simple proportion. In fact poverty rate is a more complex statistic since the poverty threshold defining it is based on the median income, which itself is subject to sampling variability. However, the impact of this approximation on the practical conclusion above is expected to be small, and, in any case, has been generally found to be favourable (reduced sampling error for the same sample size).

country i (normalised to average 1.0 over countries in equation (3)). Constant n_0 is determined to make the samples n_i add up to the desired overall total size, $\Sigma n_i=n$, say.

Note that the above imposes a minimum sample size (for M_i tending to zero, i.e. for a very small country) as = $k*n_0$. Parameter α has been introduced to impose a constraint on the maximum sample size which, with given k and n_0, is determined by the largest value of M_i encountered (i.e. the largest country).[6]

The above description is mainly in terms of the figures in column A1 of Table 2, which apply to a large majority of EU-SILC countries. For the minority of (register) countries using a sample of person (one selected respondent per household), column B1 applies. The required minimum sample sizes in terms of detailed personal interviews in column B1 are smaller than the corresponding number in column A2, because in the former case clustering of individuals within households is avoided. For attitudinal and other social variables, clustering within households can introduce significant design effects. On the basis of some empirical evidence, it has been taken that because of its higher efficiency, the required sample size for the 'social interview' under B1 can be 75% of the sample size under option A2. Note that while B1 in this context is an 'effective' sample size, A2 is not because it is subject to deft > 1 because of clustering of the interviews within households.

In terms of *households*, the required sample size under option B (column B1) is appreciably larger than that under option A (column A1), because in the former case less information on social variables is being collected per household (since such information is based only on one interview per household).

By contrast in terms of *persons* aged 16+ for whom detailed income information is being collected, the number under option B (column B2) is much larger than the corresponding number under option A (column A2). This larger sample size is acceptable since, with the complex income information obtained from registers rather than from personal interview, the cost per unit is much lower.

The logic of the figures in column B1 being larger than those in A1 (in terms of the number of households), and the corresponding figures in column B2 being larger than those in A2 (in terms of the number of persons for whom income data are collected), can be seen from the following.

The structures of the two samples – one a sample of persons for the personal interview to obtain social variables (column B1), and the other a sample of households for the collection of income variables from all adult members in each sample household (column B2) – will usually be different. Since the survey fieldwork costs mostly arise from the personal interview survey, it is important to

[6] For EU-15, the following values of the parameters were in fact used in the construction of Table 2: $k^2=0.25$, $n_0=5.785$, $\alpha=0.75$, giving $\Sigma n_i=80{,}000$, $n_{min}=k.n_0=3{,}000$. Subsequently, individual figures were adjusted or rounded as desired.

make this sample as efficient as possible. Basically this means opting for, at least approximately, an equal probability sample of persons. This would mean that the households associated with those persons are selected with unequal probabilities – *in direct proportion to the number of adults in the household* – and hence the sample is less efficient compared to a self-weighting sample of households and persons as would normally be the case under option A.[7]

The longitudinal component

The determination of the required sample sizes for the longitudinal component of EU-SILC is somewhat more complicated. The analytical objectives can be more varied and complex. Furthermore, in most situations, the longitudinal component will be linked to (even integrated with, see Section 5) the cross-sectional component, and the two cannot be determined independently. In any case, greater cost and sample size constraints apply to the longitudinal component since, as noted in Section 1, the first priority in EU-SILC is to obtain a large enough and representative cross-sectional component.

The basic approach in determining the minimum sample size requirements for the longitudinal component can be the same as that taken above for the cross-sectional component. The minimum precision requirements are expressed in terms of the *minimum effective sample size (households) required for the measurement of some critical longitudinal indicator(s) of the households' income situation*. The main household income measure for this purpose may be taken as the ***persistent poverty rate*** defined, say, as the proportion of the population which remains in poverty (i.e., with equivalised income below 60% of the median) continuously for at least a certain number of years.

On the basis of ECHP data, an average of around 10% of persons are in poverty consecutively for two or more years. Taking the same precision requirement as that for cross-sectional poverty rate discussed above, namely *1 percentage point error* in absolute terms, an effective sample size of a little under 4,000 households is required. On this basis, it has been decided in EU-SILC to take, in each country, the minimum effective sample sizes for longitudinal component as 75% of the corresponding sizes for cross-component given in Table 2.

It must be noted, however, that there is an added difficulty in moving from current poverty rates for determining the sample size of the cross-sectional component, to persistent poverty rates for the longitudinal component. It is that across countries, persistent poverty rates show a considerably wider range of variation in relative terms, compared to current poverty rates. Hence, using an EU-average value of persistent poverty rate (such as 10%) is an over-simplification, albeit the results are not sensitive to moderate variations in this

[7] See Section 3, Table 1 on selection probabilities in the various schemes for the construction of samples of households and persons.

rate. It is found empirically that in countries with *higher* current poverty rates, poverty also tends to be *more persistent* in relative terms (Betti, Cheli and Verma, 2005), thus accentuating national differences in persistent poverty rates.

The implication is that, with the minimum effective sample size determined as described above, relative error in estimating the persistent poverty rate tends to become larger in countries with low current poverty rates. Fortunately, a number of these countries are in fact 'register countries', where, if desired, the sample size for income-related variables (column B2) can be increased more easily.

In fact, an allowance has already been built into column B2 for this purpose: the sample size taken in this column is generally larger than what would be required merely to compensate for possible inefficiency of the sample, because of its non-self-weighting nature as noted earlier.

5. The integrated design

5.1. Introduction

The fundamental characteristic (hence advantage) of the integrated design is that the cross-sectional and longitudinal statistics are produced from essentially the same set of sample observations, thus avoiding unnecessary duplication which entirely separate cross-sectional and longitudinal surveys will involve.

This section describes the structure of an integrated EU-SILC survey (design [A], introduced in Section 2). As note in Introduction[8], depending on the country, micro-data could come from:

(1) one existing national source (survey or register);

(2) two or more existing national sources (surveys and/or registers) directly linkable at micro-level;

(3) one or more existing national sources combined with a new survey – all of them directly linkable at micro-level;

(4) a *new harmonised survey* (or survey system) to meet all EU-SILC requirements.

The objective here is to discuss aspects of the *structure of the sample over time* in view of the dual, cross-sectional and longitudinal, data requirements of EU-SILC.

The integrated design is the most appropriate one for data situation (4) involving the development of an entirely new interview survey instrument to meet all the EU-SILC needs - which incidentally has turned out to be the most common situation among the participating countries.

It is important to emphasise that the application of the integrated design is by no means confined to new surveys.

[8] EU-SILC Regulation (Official Journal of the European Union, 2003a).

Firstly, the integrated design applies also to the broader set of situations which design [A] of Section 2 represents. This including, for example, the situation where income, demographic, household and other data are obtained from registers to complement personal interview data, all within the structure of an integrated sample.

Secondly, the integrated structure also applies to other designs. Consider for instance design [C] of Section 2, in which separate income and social surveys are involved. Each of these surveys has to cover both cross-sectional and longitudinal components, and may therefore use the integrated design for each of them. Furthermore, the cross-sectional and longitudinal structures have to be the same or similar, since income and social data (each with cross-sectional and longitudinal components) must be linked to each other at the micro-level. The data source situation (3) above is included in this scenario.

Thirdly, at least some aspects of the integrated design may be incorporated in the adaptation of existing national sources (data source situations (1) and (2) above). For instance, when a new survey is developed for one of the two (cross-sectional or longitudinal) components while an existing national survey is used for the other component, either or both of those may be adapted so as to supplement each other, albeit to a more limited extent than would be possible with a truly integrated design being discussed here. Similar considerations may also arise in the more general situation involving the adaptation of existing national surveys for providing the whole of EU-SILC data. In EU-SILC, some interesting examples of such adaptations are already found in the surveys of Nordic countries. All these involve incorporating existing surveys to form parts of EU-SILC.

5.2. Rotational sample for cross-sectional data

Annual cross-sectional estimates can be produced from (i) independent samples from year to year; or (ii) retaining the same sample from one year to the next; or (iii) a rotational design - i.e. a combination of the above two - rotating a part of the sample from one year to the next and retaining the other part unchanged.

Cross-sectional estimates for a single year are essentially unaffected by the pattern of rotation (theoretically, modest improvements may be achieved with partial overlaps using special estimation procedures).

In principle, annual cross-sectional estimates can be produced using independent (non-overlapping) samples each year. The major consideration favouring independent samples is that such a system avoids cumulation of respondent burden which repeated interviewing of the same units would involve. The statistical advantage is that the data can be cumulated over survey years more efficiently to obtain larger sample sizes, permitting more detailed analysis and, even more importantly, greater spatial breakdown for the production of regional (subnational) estimates. Independent annual samples avoid cumulative effect of sample attrition over time. The last mentioned aspect also makes the control and implementation of the sample less complex. The main disadvantage of independent samples is the greater sampling error involved in the measurement of year-to-year net change and trends. Independent samples also tend to have higher fieldwork costs than overlapping samples re-using the same units: this is because of the higher costs of selecting and locating new sample units.

The other extreme would be to use a fixed panel, i.e. using the same sample from year to year. The advantages and disadvantages are just the opposite of those noted above for independent annual samples. Cumulative respondent burden and sample attrition, as well as the greater complexity in control and follow-up of the longitudinal sample, are major problems. The high positive correlations precludes efficient cumulation of the data over time. However, for the same reason, the system is efficient for the measurement of net change over time. The ECHP provides the most obvious example of such a design. Though primarily aimed at generating longitudinal data, the ECHP has been used extensively for providing cross-sectional statistics.

The appropriate pattern of rotation is determined primarily on the basis of a compromise between the two objectives:

(a) cumulation of data over time, so as to achieve increased sample size, which favours maximum rotation i.e. independent samples; and

(b) the measurement of change over time, which favours maximum overlap.

The effect of departures from these patterns depends on the correlation over time.

Consider two annual surveys based on similar sample sizes and design, and hence subject to similar magnitude of sampling variance. Let v be the variance of the estimated current level of a certain statistic from one annual sample.

O If the two samples are independent, the variance of the statistic estimated from the averaging over two samples will be approximately $\text{var}_0^{(c)} = v/2$. (This is because the available sample size is doubled.)

O Similarly, for an estimate of the net difference, the corresponding relationship is also $\text{var}_0^{(d)} = 2 * v$.

In the presence of sample overlaps, these relationships may be expressed in the following simple forms.

Cumulation

If $\text{var}_0^{(c)}$ is the variance which would be achieved with the aggregation of two independent samples (of the same design and size), then with overlapping samples the increased variance is approximately $\text{var}^{(c)} = \text{var}_0^{(c)} * (1 + PR) =$ $= (v/2) * (1 + PR)$, where P is the overlap (0-1), and R is the correlation coefficient for a particular statistic from the two samples (Kish, 1965, Section 12.4).

For instance with P=0.75 (i.e. a 75% overlap from one year to the next) and taking R (the correlation coefficient between successive years for some main variables of interest in the overlapping part of the sample) as 0.7, variance in estimating the two-year average will be increased roughly by 50% (1+P.R=1.5) due to the sample overlap.

Net change

If $\text{var}_0^{(d)}$ is the variance which would be achieved for the net difference between estimates from the two independent samples (of the same design and size), then with overlapping samples the reduced variance is $\text{var}^{(d)} = \text{var}_0^{(d)} * (1 - PR) = \; = 2 * v * (1 - PR)$.

For instance with P=0.75, and R (the correlation coefficient between successive years for some main variables of interest in the overlapping part of the sample) as 0.7, as in the above example, variance in estimating change will be roughly halved (1-P.R=0.5) due to the sample overlap. This means that, with the overlap and correlation as assumed above, variance in estimating net change (var) becomes of the same order of magnitude as the variance *v* in estimating annual cross-sectional levels.

In the case of EU-SILC, the measurement of trends (changes over time) is likely to be clearly more important than cumulating data over years, favouring large overlaps (P) from one year to the next. However, as noted above, there are practical limitation in continuing with the same sample, i.e. with having large overlaps (P close to 1.0) from year to year.

Special re-weighting to measure change more precisely

In principle, the gain in precision in estimating change can be enhanced by giving more weight to the overlapping part and correspondingly less weight to the rotating part in the estimation procedure. (This can be done because each of the two parts is a representative sample in itself.) With optimally determined weights, the theoretical gain can be shown to be $\text{var}^{(d)} = \text{var}_0^{(d)} * [(1 - R)/(1 - QR)]$, where Q=1-P. With P=0.75 and R=0.7 for instance, the variance is reduced to just over a third due to the sample overlap, compared with only halving of the variance in the previous illustration without reweighting.

This means that, with the overlap and correlation as assumed above and with optimal reweighting, variance in estimating net change ($\text{var}^{(d)}$) reduces to only 0.7 of the variance in estimating annual cross-sectional levels (*v*). Hence the advantage of such special reweighting in estimating changes can be very substantial. Similarly for cumulation over years, the loss in precision in estimating averages can be reduced by giving more weight to the rotating part and correspondingly less weight to the overlapping part in the estimation procedure.

With optimally determined weights, the theoretical loss can be shown to be $\text{var}^{(c)} = \text{var}_0^{(c)} * [(1+R)/(1+QR)]$.

The reweighting is much less effective in this case of cumulation: with P=0.75 and R=0.7 for instance, the inflation in variance due to overlapping is reduced only to 1.45, from 1.5 in the previous illustration without this special reweighting (Kish, *ibid*). In any case, cumulation over years is a less important issue in EU-SILC than estimating net change from year to year.

Note that the above comments apply to the *cross-sectional sample* being discussed here. Of course, by definition, the longitudinal component represents a 100% overlap in the sample from year to year. Cumulation over time has quite a different purpose for the longitudinal component: the interest being to increase the number of events and transitions observed. Such cumulation is useful, and also necessary, given the relatively small sample sizes likely to be available in EU-SILC.

Constructing overlapping samples

Below we describe a practical procedure for constructing samples which have a specified degree of overlap from one year (survey round) to the next.

Consider two successive years with partially overlapping samples. For the cross-sectional sample for each year to be separately representative requires each of the following three parts to be a representative sample:

(i) the dropped part to be representative of the population at year 1;

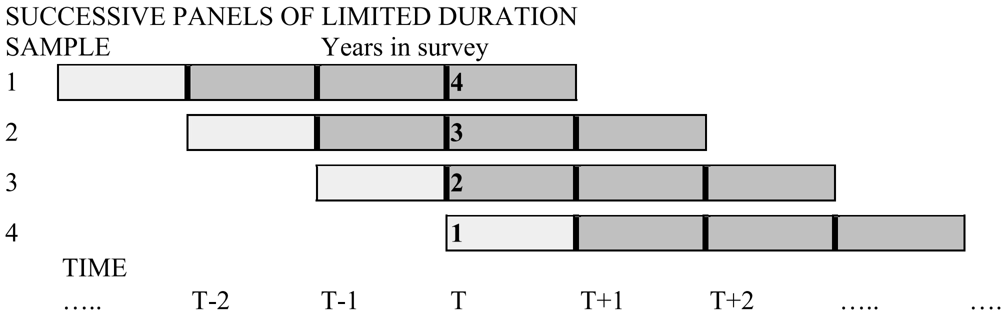
(ii) the added part to be representative of the population at year 2; and

(iii) the overlapping part to be representative of the population common to the two years.

Normally, the above is achieved in practice by selecting the total sample in the form of a number of replications. Each replication is in itself a representative sample, typically with the same design (structure, stratification, allocation, etc…) as the full sample, differing from the latter only in sample size. From one year to the next, some of the replications are retained, while others are dropped and replaced by new replications depending on the extent of overlap desired. The technique of selecting the sample in the form of independent replications, each similar in size and design and representing the whole population, offers flexibility and control over the pattern of sample rotation.

Figure 2 illustrates a simple rotational design (once the system is fully established). The sample for any one year consists of 4 replications, which have been in the survey for 1-4 years (as shown for 'Time=T' in the figure). Any particular replication remains in the survey for 4 years; each year one of the 4 replications from the previous year is dropped and a new one added, giving a 75% overlap from one year to the next. For surveys two years part, the overlap is 50%; it is reduced to 25% for surveys three years apart, and to zero for longer intervals. Generally with n replications, each kept in the survey for n rounds, the overlap between rounds declines linearly as the interval separating them increases. For two surveys i intervals apart the overlap is (n-i)/n, up to i=(n-1), after which (i>=n) the overlap becomes zero.

Figure 2. Illustration of a simple rotational design

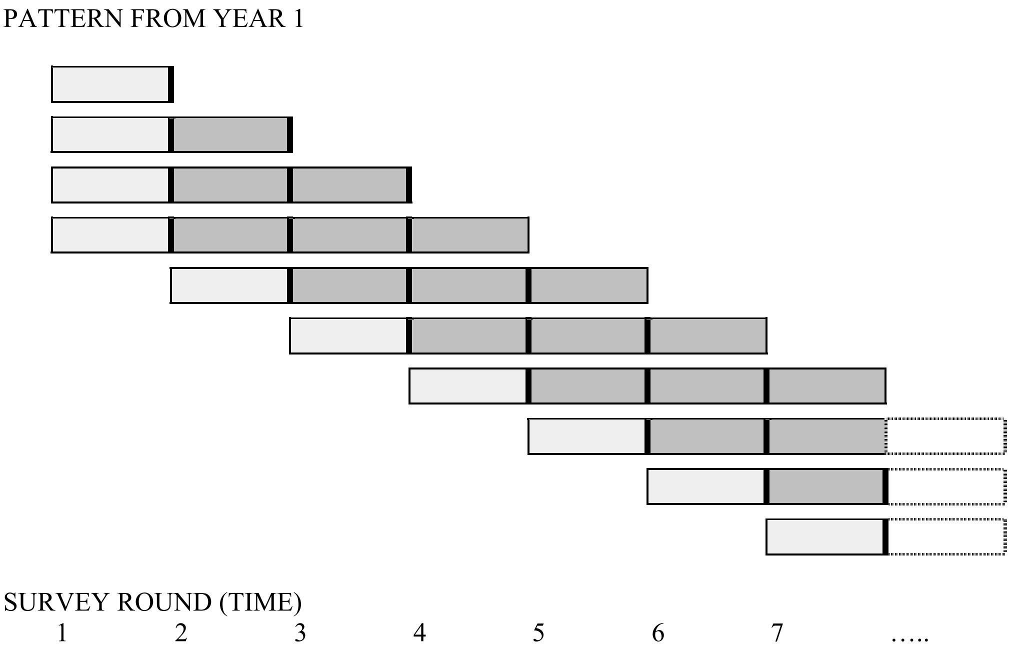


For EU-SILC, such a 'linear' rotation pattern is the simplest and most appropriate in so far as the main interest is in monitoring year-to-year changes. Indeed, it has been adopted by most of the countries to date. More complex patterns can be introduced to vary the degree of sample overlaps and how that changes over time, as for instance is done in some labour force surveys, but they are unlikely to be of interest for EU-SILC.

Starting the rotation pattern

Figure 3 illustrates how a rotation pattern may be started from year 1. To obtain the full sample with 4 replications for the first year, it is necessary to begin with all the 4 replications. These replications are treated differently over time. One of these is dropped immediately after the first year, the second is retained for only 2 years, the third for 3 years, and only the fourth is retained for the full 4 years. The pattern becomes 'normal' from year 2 onwards: each year one new replication is introduced and retained for 4 years. Alternative starting schemes are possible, but this illustration is perhaps the simplest one. Again, this structure has been adopted in most of the EU-SILC surveys to-date.

Figure 3. Illustration of the rotation pattern in the first years



5.3. The longitudinal sample

Initial sample of a panel

The longitudinal component of EU-SILC consists of selecting an initial sample, then following-up individuals in that sample annually over time. As in the case of the EU-SILC cross-sectional component, the original sample for each panel in the longitudinal component normally consists of a sample of *households*. The actual selection mechanism may vary; the ultimate sampling units may for instance be address, households or persons. However, irrespective of details of the actual sampling procedure, effectively each panel begins with a probability sample of households (see Section 3.1). This initial sample of households is used to define the samples for the different types of units of interest: households, all household members, all adults aged 16+ in the household; and possibly, a subsample of the above consisting of sample persons to be interviewed in detail. Hence the structure of the initial sample for the panel is the same as that of the cross-sectional component.

The difference between the two lies in the manner in which the sample is followed up over time. This is discussed in Section 5.4 below. It is useful first to consider how a rotational design of the type described in Figures 2 and 3 for the cross-sectional survey can also serve the longitudinal component in a combined design.

Basic structure of the combined cross-sectional and longitudinal design

Figures 2 and 3 above also illustrate the type of structure which may be suitable for meeting the combined cross-sectional and longitudinal requirements. Figure 2 illustrates the system once established, supplemented by Figure 3 displaying how the system may be started from EU-SILC year 1.

At the beginning, a cross-sectionally representative sample of households is selected. It is divided into, say, 4 subsamples, each by itself representative of the whole population and similar in structure to the whole sample (except for sample size). One subsample is purely cross-sectional and is not followed up after the first round. Respondents in the second subsample are requested to participate in the panel for 2 years, in the third subsample for 3 years, and in the fourth for 4 years. From year 2 onwards, one new panel is introduced each year, with request for participation for 4 years.

In any one year, the sample consists of 4 subsamples. In year 1 they are all new samples and there is no longitudinal data. In all subsequent years, only one is new sample. In year 2, three are panels in the second year; in year 3, one is a panel in the second year and two in the third year; in subsequent years, one is a panel for the second year, one for the third year, and one for the fourth (final) year.

While the above structure looks similar to that for a purely cross-sectional survey described in the previous section, it is important to appreciate differences in the follow-up ('tracing') rules between the two. These differences have important practical consequences and will be discussed more fully in Section 5.4. The essential point in this connection is the following.

In the purely cross-sectional survey with sample overlaps, the follow-up consists, at a minimum, of revisiting the same sample addresses, without following-up moving households and persons to their new location. Instead, we may follow-up originally selected households and persons to their new address if they move. This alternative is optional in a cross-sectional sample with overlaps. By contrast, such a follow-up procedure is essential in a longitudinal sample. Hence also in the combined cross-sectional and longitudinal structure, it is necessary to *follow-up individuals once selected*. The tracing rules and procedures for this purpose have already been described in Section 3.

5.4. Sample follow-up over time

Sample follow-up in the context of a cross-sectional survey

It is important to be clear about what 'overlapping' means in the context of a purely cross-sectional survey. The basic requirements are:

(i) that the data for the two years from the overlapping part are correlated; and

(ii) to the extent possible, the overlapping part is representative of population common to the two times.

Correlation does not necessarily require the samples to be identical in terms of the ultimate units (persons, or even households). A common procedure (as for instance used in most labour force surveys with rotational designs) is to have an overlapping sample of addresses or dwelling units. In each survey, all households and persons found at those addresses are taken into the sample. Households and persons which move are not followed-up to their new location, as the sample is defined by the locations (addresses) originally selected. For the measurement of net changes, it is not necessary to link the information for individual units over time: such linking is required only if the objective is to analyse gross or longitudinal changes at the micro level (as in panel surveys). With multi-stage sampling designs (e.g. the selection of areas followed by the selection of addresses) it is in fact not essential that the overlap be in terms of the ultimate sampling units (addresses); correlation, albeit reduced, can also be obtained by having common higher stage units (same areas but independent samples of addresses within each).

However, in most of the EU-SILC cross-sectional surveys, the year-to-year sample overlaps are in terms of *addresses or dwelling units*. In relation to overlap over time between cross-sectional samples the simplest option would be to retain in the sample households and persons found at the selected locations (addresses) even if they are different from those enumerated previously. In this way households and persons which have moved to a new location outside the sample can be replaced by households and persons which have moved into the sample locations in their place.

On the basis of rules of association between different types of units, sample overlaps in terms of fixed locations (addresses) are possible even when some different types of units such as households or persons, rather than addresses, have been used as the ultimate sampling units.

Follow-up in the longitudinal sample

As noted, the follow-up requirements in a longitudinal survey are quite different and more complex compared to those permissible in a cross-sectional survey based on a rotational sample.

The more complex follow-up procedures necessary in the longitudinal design, and hence also the combined design, compared to the simpler ones permissible for a purely cross-sectional survey, have some important practical consequences.

1. The main administrative complication is the need to *identify and trace individuals over time*. The success of the panel critically depends on uniquely and correctly identifying every person ever coming into the survey and linking his/her information over time.
2. The major technical complications arise from the more complex weighting, editing and imputation procedures to ensure longitudinal completeness, consistency and representativeness of the data, and also at the same time to

produce valid cross-sectional estimates making full use of all the available data.

3. A true panel of individuals is likely to suffer from somewhat higher attrition rates, compared to repeated enumeration of simply the same addresses which is an option in a rotational cross-sectional survey.
4. The main additional fieldwork cost arises from the need to trace movers. Assuming that on the average 2-3% of the population move address during any one year, the EU-SILC survey at any one time should not have more than 5% or so movers. (The four subsamples in it are subject to 0, 1, 2 and 3 years of moving, i.e. an average of 1.5 years.) Even though the proportion of movers is small, it can be taxing in terms of complexity, cost and duration of the fieldwork required.
5. The important point to note is that a vast majority (perhaps 95% or more) of households and persons will be re-enumerated at their original address, i.e. just as in a purely cross-sectional survey with sample overlaps. It is this fact which makes the incorporation of a panel component into a cross-sectional survey cost-effective: much of the longitudinal component is in fact enumerated as a by-product of the cross-sectional enumeration.
6. Finally we must note that there is also an added complexity in the cross-sectional component in a rotational design. At any one time the cross-sectional sample involves several panels, each with a different starting point, and hence a different population covered, especially concerning immigrants.

<u>A useful option: doubling the enumeration of movers in the cross-sectional sample</u>

The samples resulting from the different tracing rules in a purely cross-sectional sample and a combined cross-sectional/longitudinal design are illustrated in Figure 4.

Set (A+C) provides the only valid longitudinal sample for years 1-2 (with C enumerated at their original addresses at year 1 and at their new location at year 2). This also provides a valid sample for the cross-sectional component.

Suppose, however, that part B – movers into vacated address originally selected into the sample – is also enumerated in the survey. Now there are in fact three valid cross-sectional samples: (A+C); (A+B); and hence also their combination (A+B)+(A+C), which is the same thing as:

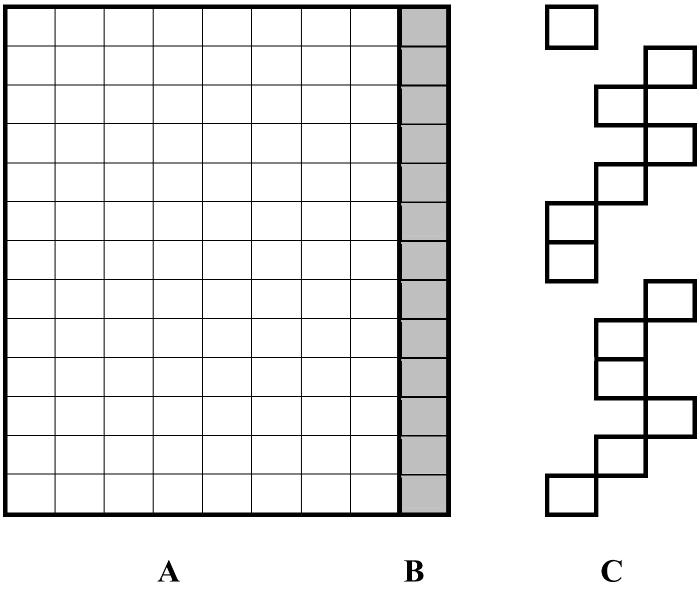
$$\left(A+\frac{B+C}{2}\right), \tag{4}$$

meaning a sample consisting of set A, plus B and C each with half their normal weight.

Each of the three alternatives have their advantages.

O Set (A+C) has the convenience of consistency in that in this case the longitudinal and the year 2 cross-sectional samples have the same base (A+C). No extra enumeration of B is involved.

Figure 4. Cross-sectional (A+B) and longitudinal (A+C) samples at year 2



Stationary	A	Set of persons residents at the same address at both years
Movers-in	B	New set of persons, found at year 2 at the original (year 1) sample addresses
Movers-out	C	Set of persons at the original year 1 sample addresses, who have moved to some new location at year 2

O Set (A+B) does not involve follow-up of movers to new addresses, and hence is likely to be completed earlier. The cross-sectional results can be produced in a more timely fashion, not affected by the normal delay in completing the longitudinal component (C).

O The recommended set [A + (B+C)/2] uses all the available information, which increases precision. The advantage is more than simply increased sample size. By bringing in both out-movers and in-movers into the sample, it doubles the available sample size for movers – the group likely to be of special interest in the analysis of change.

Unfortunately, this proposal of double enumeration of movers has not been taken up in any EU-SILC survey to-date.

5.5. Adjusting the relative sample sizes of the two components

The model described by Figure 2 is too rigid in one respect: it assumes a fixed relationship between the cross-sectional and longitudinal sample sizes. The relative size of the panel component can be increased (reduced) only by increasing (reducing) its duration. However, that duration (such as 4 years in EU-SILC) is given by the survey's substantive objectives, and is not a parameter which can be chosen on the basis of sampling considerations.

Greater flexibility can be achieved by supplementing the above structure by one or the other of the following: split panel and cross-sectional booster.

Split panel refers to addition to the basic structure (Figure 2) of a panel component of unlimited duration. This increases the available sample size of the panel part. (The term 'split panel' was introduced by Kish (1981) to describe such an arrangement.) The size of the split panel can be chosen flexibly to obtain a panel component of the required size.

Of course, the addition of a split panel of unlimited duration brings in new considerations and possibilities, beyond the stated basic (minimum) requirements of EU-SILC.

The size of the cross-sectional component can be increased by adding to the basic structure (Figure 2) a fully rotational cross-sectional booster. Again, the size of the booster can be chosen flexibly to obtain a cross-sectional component of the required size. These ideas in a modified form have been applied in a couple of EU-SILC surveys (such as Norway, Finland).

<u>Illustration</u>

Consider a rotational design with r replications or subsamples, each of size s. In the basic model, each subsample is retained in the survey for r years.

In any round,

- the cross-sectional sample is n_1=r*s;
- the longitudinal sample linked over two years is of size n_2=(r-1)*s (since all but the newly introduced panel provide such linkage with the previous year);
- the longitudinal sample linked over three years is of size n_3=(r-2)*s (since all but the two most recently introduced panels provide such linkage with year y-2);
- that linked over four years is of size n_4=(r-3)*s; and so on.

With the addition of a split panel of size p, each of the above is essentially increased by p, so that the longitudinal to cross-sectional sample size ratio, such as n_{i+1}/n_1 is increased from:

$$\frac{n_{i+1}}{n_1} = \frac{r-i}{r} \text{ to } \frac{n_{i+1}}{n_1} = \frac{r-i+(p/s)}{r+(p/s)}. \qquad (5)$$

With the addition of a cross-sectional booster of size x, the available cross-sectional sample is increased by x without affecting the longitudinal component. The cross-sectional to longitudinal sample size ratio is therefore increased from:

$$\frac{n_1}{n_{i+1}} = \frac{r}{r-i} \text{ to } \frac{n_1}{n_{i+1}} = \frac{r+(x/s)}{r-i} \qquad (6)$$

REFERENCES

BETTI, G., CHELI, B., VERMA, V. (2005), *On longitudinal analysis of poverty conceptualised as a fuzzy state*, paper to be presented to the First Meeting of the Society for the Study of Economic Inequality, Palma de Mallorca, 20-22 July 2005.

EUROSTAT (2004a), *Description of target variables: cross-sectional and longitudinal*. EU-SILC 065/04, Luxembourg.

EUROSTAT (2004b), *Technical document on intermediate and final quality reports*. EU-SILC 132/04, Luxembourg.

KISH, L. (1965), *Survey sampling*. John Wiley & Sons, New York.

KISH, L. (1981), *SCPR Survey Methods Newsletter*. Winter 1981.

OFFICIAL JOURNAL OF THE EUROPEAN UNION (2003a), 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 (2003b), Regulation (EC) No 1982/2003 of 21 October 2003 implementing 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) as regards the sampling and tracing rules.

VERMA, V., (1991), *Sampling Methods*. Training Handbook, Statistical Institute for Asia and the Pacific, SIAP, Tokyo.

VERMA, V., (2001), *EU Statistics on Income and Living Conditions (EU-SILC).* Sampling Guidelines. Luxembourg: Statistical Office of the European Communities.

VERMA, V., CLEMENCEAU, A. (1996), Methodology of the European Community Household Panel. *Statistics in Transition*, Vol. 2, No. 7, 1023-1062.