Appendice 1: L'indagine EU-SILC

Uno degli obiettivi principali verso il quale tende, da numerosi anni, l'Unione Europea è quello della lotta all'esclusione sociale e alla povertà all'interno dei Paesi membri. A tale scopo è stata definita una strategia comune a livello europeo che si esplicita principalmente attraverso la definizione di Piani di Azione Nazionali, attraverso l'applicazione del metodo aperto di coordinamento e mediante il confronto tra le esperienze maturate nei singoli Paesi membri. In particolare, a partire dal Consiglio Europeo tenutosi a Lisbona nel 2000, sono stati definiti alcuni obiettivi di lungo periodo tra i quali quello di ridurre drasticamente la povertà dei cittadini europei entro il 2010. Da qui nasce la chiara esigenza di disporre di indicatori statistici robusti, tempestivi e comparabili a livello europeo che permettano il monitoraggio delle performance degli Stati membri su questi temi. Questo rende possibile la valutazione sia dell'impatto delle politiche implementate nei singoli Paesi, che della possibilità di estendere le "buone pratiche" così emerse agli altri Stati membri. In particolare, questo processo è culminato con il Consiglio d'Europa tenutosi a Laeken nel dicembre 2001, dove sono stati individuati gli indicatori introdotti nel Capitolo 3.

App. 1.1. Che cosa è il regolamento EU-SILC

L'indagine *European Union – Statistics on Income and Living Conditions* (EU-SILC) risponde al regolamento dell'Unione Europea n°1177/2003 (European Community, 2003), elaborato in seguito alla crescente domanda di informazioni da parte delle istituzioni nazionali ed europee, della comunità scientifica e degli stessi cittadini.

Il progetto ha come obiettivo principale la produzione sistematica di statistiche comunitarie su reddito, povertà ed esclusione sociale, sia a livello trasversale che longitudinale, puntando all'armonizzazione di un insieme di indicatori statistici. Questo costituisce una delle principali fonti di dati per i rapporti periodici dell'Unione Europea sulla situazione sociale e sulla diffusione della povertà nei paesi membri.

Si tratta di una indagine campionaria sulle famiglie il cui *core* informativo è essenzialmente incentrato attorno alle tematiche del reddito e dell'esclusione sociale. Il

progetto è ispirato da un approccio multidimensionale, con una particolare attenzione agli aspetti di deprivazione materiale.

Nel regolamento sono definite delle norme comuni al fine di migliorare la qualità, la comparabilità e la tempestività dei dati, e favorire una migliore integrazione tra i sistemi statistici nazionali.

Gli aspetti metodologici di questo strumento sono stati sviluppati in cinque regolamenti della Commissione; inoltre ogni anno, un nuovo regolamento definisce la lista delle variabili target per un modulo *ad hoc* scelto tra una rosa di tematiche di interesse.

I quattro temi toccati finora sono stati: la trasmissione intergenerazionale della povertà (attraverso la rilevazione di informazioni sulla famiglia di origine), la partecipazione sociale, le condizioni abitative, l'indebitamento e l'esclusione finanziaria. Il modulo di approfondimento del 2009 è dedicato, invece, alla misura della deprivazione materiale.

EU-SILC è stato creato come strumento flessibile e confrontabile, comprendendo dati e fonti di dati di diverso tipo: *cross-sectional* e longitudinali, a livello di nucleo familiare e di singolo individuo, economici e sociali, provenienti dai registri o da indagini campionarie, derivanti da indagini nazionali nuove ed esistenti, o da altre fonti.

Si tratta di un'indagine a carattere obbligatorio per i Paesi membri dell'UE.

Il progetto EU-SILC è stato sperimentato nel 2003 in sette paesi (Belgio, Norvegia, Grecia, Lussemburgo, Austria, Danimarca e Irlanda). Il lancio ufficiale si è avuto, invece, nel 2004 in tredici degli originali Stati Membri, compresa l'Italia (non hanno partecipato Paesi Bassi, Germania e Regno Unito), e in dieci nuovi Stati Membri (eccetto Estonia), oltre che in Norvegia e Islanda. Nel 2005, EU-SILC ha raggiunto la sua piena estensione con venticinque Stati Membri, più Norvegia e Islanda (EU-SILC è in preparazione anche in Turchia, Romania, Bulgaria e Svizzera).

App. 1.2. Le principali differenze tra EU-SILC ed ECHP

Questo progetto ha sostituito *l'European Community Household Panel* (ECHP), una indagine campionaria che dal 1994 al 2001 è stata effettuata con cadenza annuale basata su un questionario standardizzato somministrato alle famiglie e agli individui residenti in quattordici Paesi europei. Il disegno dell'indagine prevedeva che ogni componente

della famiglia, facente parte del campione, venisse intervistato per otto anni consecutivi (Verma e Clemenceau, 1996); anche questa indagine era coordinata da Eurostat.

L'EU-SILC è stato introdotto con l'obiettivo di inserire uno strumento migliorativo, soprattutto per far fronte a diversi difetti presentati dall'ECHP.

Nell'ECHP le famiglie facenti parte del campione estratto venivano seguite per tutta la durata dell'indagine e re-intervistate nelle otto "ondate" successive: ciò comportava un fenomeno noto come "attrito", ossia l'inevitabile decremento della numerosità campionaria nel susseguirsi degli anni, imputabile sia ai rifiuti che all'impossibilità di rintracciare nel tempo coloro che fanno parte del campione, con conseguenti problemi di rappresentatività del campione stesso specie nelle stime trasversali. L'ECHP era inoltre caratterizzato da un alto tasso di rifiuto a partecipare sin dall'inizio all'indagine, specie in alcuni Paesi; in più c'erano tempi eccessivamente lunghi nel rilascio dei dati.

ECHP era uno strumento rigido e le cui articolazioni venivano stabilite in sede europea; al contrario il regolamento EU-SILC ha consentito ai singoli paesi una certa elasticità rispetto all'impiego di differenti fonti di dati (indagine campionaria/archivi), al periodo di riferimento del reddito (fisso/mobile), alla modalità di raccolta delle informazioni sui redditi lordi (indagine/archivi/micro simulazione) e alla struttura dei questionari nazionali.

I paesi possono scegliere i dati che ritengono provenire dalle *"best source(s)"*, questo può ridurre il grado di armonizzazione delle metodologie fra i paesi e conseguentemente la comparabilità, ma migliorare la qualità dei dati rilevati a livello nazionale.

Molta flessibilità è stata lasciata anche per la definizione del disegno campionario, per i metodi di imputazione e per il calcolo degli stimatori.

L'obiettivo della comparabilità si configura come un processo di convergenza graduale che dovrà considerare non solo le metodologie utilizzate dagli Istituti nazionali, ma anche le differenze che caratterizzano i sistemi di *welfare* dei paesi membri e la loro evoluzione nel corso del tempo.

Uno dei principali obiettivi del nuovo strumento di rilevazione è stato riconosciuto nella tempestività dei dati. Infatti, mentre per l'indagine panel europea i dati *cross-sectional* e longitudinali venivano raccolti e trattati nello stesso momento, nel caso di EU-SILC è stato previsto che i dati trasversali e longitudinali, che possono derivare da fonti separate, siano rilasciati secondo un diverso calendario.

App. 1.3. Obiettivi dell'indagine e contenuti informativi

Le indagini sui redditi e le condizioni di vita hanno due tipi di utilizzo: scientifico e politico.

Tra gli impieghi di carattere scientifico, i più importanti sono l'analisi della distribuzione dei redditi, della disuguaglianza e della povertà. Nell'EU-SILC è previsto che la stima degli indicatori inerenti l'esclusione sociale sia corredata dai relativi errori standard così da poter costruire, per ogni indicatore, degli intervalli di confidenza all'interno dei quali si troverà la relativa stima puntuale. Questo è un importante elemento che non era presente nell'ECHP.

Nonostante infatti l'obiettivo e i contenuti dell'EU-SILC siano molto simili a quelli dell'ECHP, quello che cambia sono il contesto e la struttura.

EU-SILC ha la caratteristica di basarsi su diverse fonti e strutture di dati, e consente di confrontare dati di diversa origine.

Infatti come risulta dai regolamenti europei, a seconda dei paesi, i micro-dati possono provenire da:

1) Una fonte nazionale già esistente (indagine o registro);
2) Due o più fonti nazionali esistenti (indagini e/o registri) direttamente collegabili a livelli micro;
3) Una o più fonti nazionali esistenti unite con nuove indagini (tutte collegabili a livelli micro);
4) Una nuova indagine armonizzata (o un sistema di indagini) che soddisfi tutti i requisiti previsti da EU-SILC.

App. 1.4. Definizioni di reddito nell'EU-SILC

Secondo la definizione prevalente nei manuali di economia, e come abbiamo avuto la possibilità di osservare nel Capitolo 1, il reddito è semplicemente la somma dei consumi e dei risparmi:

$$Y = C + S \qquad \text{(App. 1.1)}$$

Si noti che il risparmio S è uguale alla variazione della ricchezza. Se in un periodo Y>C (ovvero, se il reddito supera i consumi), il risparmio sarà positivo e si avrà un aumento

di ricchezza. Quando invece Y<C, il risparmio sarà negativo e tale situazione corrisponderà ad una diminuzione di ricchezza. In quest'ultimo caso, in effetti, l'individuo finanzia i suoi consumi attraverso la vendita di beni e attività finanziarie oppure indebitandosi (in tutti e due i casi la ricchezza si riduce).

Da questa semplice premessa discende la definizione di reddito accettata internazionalmente, ed accolta dal progetto EU-SILC nelle sue linee fondamentali: "*il reddito è la quantità massima di moneta che un individuo può spendere per consumi senza diminuire la propria ricchezza, cioè senza vendere parte del proprio patrimonio e senza fare nuovi debiti*".

Nell'indagine EU-SILC, il reddito viene osservato come un insieme di entrate ricavate da fonti diverse, secondo lo schema seguente:

1. *Reddito guadagnato sul mercato*
    - 1.1 *Redditi da lavoro*
        - dipendente
        - autonomo
    - 1.2 *Redditi da capitale*
        - reale (affitti e rendite di terreni e fabbricati)
        - finanziario (interessi, dividendi, utili)
        - intellettuale (diritti d'autore)
2. *Reddito da trasferimenti*
    - 2.1 Trasferimenti pubblici
        - pensioni
        - altri trasferimenti pubblici in denaro (per esempio assegni familiari)
    - 2.2 Trasferimenti privati
        - aiuti in denaro di familiari ed amici, assegni di ex-coniugi
        - aiuti in denaro di istituzioni private (per esempio da associazioni religiose).

Accanto a queste componenti, misurate in moneta, si considerano anche altre risorse "non-monetarie" che concorrono al benessere familiare:

- salari in natura (*fringe benefits*): come l'uso gratuito di una abitazione, l'auto aziendale per usi privati, i buoni pasto (dal 2007), l'asilo nido aziendale (dal 2007);

- affitti imputati dalle case occupate dai proprietari: che è pari al valore del servizio che queste abitazioni rendono a chi ne è proprietario. Per convenzione, è "come se" i proprietari affittassero la casa a sé stessi;
- autoconsumi (dal 2007): cioè il valore stimato dei beni che la famiglia ha eventualmente prodotto per il proprio consumo, come per esempio frutta, vino e ortaggi.

Sono invece escluse dalla definizione di reddito adottata per l'indagine EU-SILC, per difficoltà di rilevazione e/o di stima del valore monetario corrispondente, alcune componenti che pure concorrono a determinare le condizioni economiche delle famiglie:

- trasferimenti pubblici in natura, come per esempio i servizi sanitari e scolastici forniti gratuitamente o a prezzi agevolati dalla pubblica amministrazione. È indiscutibile che tali beni, se utilizzati, possano contribuire al benessere delle famiglie. Tuttavia è praticamente impossibile, in assenza di un mercato privato indipendente, stimare il valore corrispondente a questo tipo di benefici pubblici ricevuti in natura dalle famiglie. Anche quando esistono servizi privati in concorrenza di quelli pubblici, infatti, il loro prezzo di mercato è "residuale" rispetto alla politica di offerta dell'operatore pubblico. La valutazione al costo di produzione, a sua volta, ignora la qualità dei servizi e può non riflettere la disponibilità a pagare degli utenti;
- i beni e i servizi in natura ricevuti da parenti e amici (per esempio, la cura dei figli da parte di una parente non coabitante), per la difficoltà di valutarne sia la quantità, sia il valore "figurativo";
- per difficoltà di rilevazione, sono anche escluse tutte quelle attività lavorative effettuate dai membri della famiglia in sostituzione di analoghi servizi di mercato, come per esempio la riparazione di elettrodomestici, la manutenzione di mobili, eccetera. La difficoltà in questo caso riguarda la vasta gamma coperta dalla produzione domestica: in effetti, anche le pulizie di casa e la preparazione dei cibi sostituiscono servizi acquistabili altrimenti sul mercato.

Sono, infine, escluse alcune entrate eccezionali che sono considerate come variazioni "istantanee" della ricchezza:

- le vincite alla lotteria;

- le eredità e le donazioni una tantum;
- i guadagni in conto capitale, cioè gli aumenti del valore del patrimonio posseduto (case, terreni, gioielli, azioni ed altre attività finanziarie).

A partire dall'edizione del 2007, l'indagine EU-SILC ha incluso anche i redditi lordi individuali e familiari. Si tratta di un'importate innovazione metodologica, che ha consentito di valutare in che misura la disuguaglianza nella distribuzione dei redditi dipende dalle opportunità di mercato (distribuzione primaria), per esempio dai livelli di occupazione e di salario, e quale sia l'effetto redistributivo delle imposte e dei trasferimenti pubblici.

Nel progetto EU-SILC, il reddito lordo totale è uguale alla somma dei redditi netti, delle imposte personali sui redditi, delle imposte patrimoniali e dei contributi sociali a carico dei lavoratori. Su tale argomento fondamentale è l'apporto del modello di microsimulazione SM2, adottato ufficialmente da Eurostat, e ampliamente descritto nel Capitolo 6 della dispensa.

App. 1.5. Le strategie di rilevazione dei redditi

La rilevazione campionaria dei redditi pone numerosi problemi, dovuti a due ordini di motivi:

- scarsa conoscenza da parte degli intervistati:
  - delle definizioni di reddito
  - degli importi esatti percepiti
- scarsa disponibilità a rispondere all'intervista:
  - per diffidenza (soprattutto timore di controlli fiscali)
  - per sfiducia nelle istituzioni e nell'utilità delle indagini statistiche.

Il primo problema è stato affrontato attraverso una formulazione il più possibile semplice e precisa del questionario, l'accurata formazione dei rilevatori e prevedendo, per chi non ricorda un importo esatto, la possibilità di dare risposte approssimate. Per superare il secondo problema, è stata determinante la condivisione da parte delle famiglie dello scopo dell'indagine, insieme alla reputazione di Eurostat a garanzia dell'assoluta riservatezza delle informazioni raccolte.

In questo contesto assumono un'importanza cruciale sia le modalità di contatto e di sensibilizzazione delle famiglie, sia le tecniche di rilevazione che consentono di ridurre al minimo le mancate risposte (in particolare il disegno del questionario e la formazione dei rilevatori), sia infine le metodologie di imputazione delle mancate risposte sui redditi.

Il coinvolgimento dei rispondenti ha mostrato buoni risultati, infatti nella maggioranza dei casi le famiglie hanno compreso l'importanza e lo scopo dell'indagine, e collaborato attivamente all'intervista, consultando quando possibile i documenti a loro disposizione (come la busta paga).

Questo ha avuto una ricaduta positiva sia sul contenimento delle non risposte che sulla qualità delle informazioni raccolte.

Tradizionalmente, il reddito monetario viene utilizzato per valutare il benessere delle famiglie e degli individui che ne fanno parte. Esistono però recenti studi, che hanno messo in luce come l'analisi delle condizioni delle famiglie sulla base di una misura soltanto monetaria a volte non consenta di valutare aspetti importanti della qualità della vita. L'indagine EU-SILC, proprio a questo scopo, considera un insieme ampio di rilevazioni delle condizioni di vita. A livello individuale, sono rilevate le condizioni lavorative, i livelli di istruzione e il grado di salute. A livello familiare vengono evidenziate le caratteristiche della casa e della zona di abitazione ed una serie di indicatori soggettivi sulle difficoltà economiche della famiglia. Sono presenti poi altre variabili non monetarie che rilevano i funzionamenti (o meglio, i mancati funzionamenti) legati alla disponibilità di moneta. Per esempio viene chiesto alle famiglie se possono riscaldare adeguatamente l'abitazione, se riescono a sostenere senza difficoltà le spese scolastiche, ecc…

App. 1.6. Redditi lordi, imposte, contributi sociali e comparabilità internazionale

A partire dal 2007, il progetto EU-SILC ha previsto la disponibilità, accanto ai redditi netti, di microdati sui redditi lordi, imposte e contributi sociali. Si è trattato di un obiettivo importante che ha consentito agli studiosi e ai *policy makers* europei di valutare gli effetti delle politiche tributarie e sociali sulla distribuzione dei redditi.

Un altro obiettivo ritenuto fondamentale dal progetto EU-SILC, è la comparabilità internazionale dei dati, ottenuta grazie ad opportuni studi metodologici. Inizialmente gli sforzi si sono concentrati sulla definizione accurata di variabili obiettivo, eliminando dove possibile le ambiguità concettuali legate alle terminologie nazionali. Rimangono però dei problemi legati soprattutto al fatto che alcuni paesi utilizzano dati di reddito campionari ed altri soltanto dati amministrativi; inoltre la definizione del reddito da lavoro autonomo non è ancora sufficientemente comparabile a livello internazionale a causa dell'eterogeneità delle fonti a disposizione dei diversi paesi; infine risulta complessa la definizione di reddito disponibile e la possibilità di osservare redditi negativi.

App. 1.7. Il disegno di indagine

Una delle maggiori innovazioni introdotte con la nuova indagine sono il passaggio da un'indagine Panel "classica" ad una che prevede la presenza sia di una componente trasversale che di una longitudinale non necessariamente collegabili tra di loro. La scelta della durata della parte longitudinale dell'indagine è stata fissata in almeno quattro anni per permettere la misurazione della popolazione a rischio di povertà persistente. Tale indicatore ricopre un'importanza fondamentale nella misurazione della coesione sociale.

Il campione relativo a ogni occasione d'indagine è costituito da quattro gruppi rotazionali (ognuno di dimensione pari a un quarto della numerosità campionaria complessiva). Ogni gruppo rimane nel campione per quattro anni consecutivi e ogni anno il campione si rinnova con l'entrata di un nuovo gruppo.

Ovvero viene stabilito uno schema di rotazione che fa sì che le famiglie estratte vengano intervistate in un numero limitato di ondate (quattro) per poi uscire dal campione, mentre nuove famiglie entrano a far parte del campione stesso, sostituendole. L'adozione di uno schema di rotazione ha un duplice vantaggio: da un lato permette di ridurre l'attrito tipico delle indagini Panel, dall'altro, l'introduzione di nuove famiglie nel campione, consente l'arricchimento informativo dei dati trasversali rispetto a quello che si avrebbe con un'indagine Panel classica. Allo stesso tempo tale disegno integrato, consigliato dallo stesso Eurostat per quei Paesi dove l'EU-SILC viene espletata

attraverso la progettazione di una nuova indagine, permette di ridurre l'errore di campionamento delle stime trasversali e di evitare la duplicazione di informazioni.

Gli eventi demografici vengono osservati e rilevati nel momento in cui si manifestano (in trasversale), inoltre però, sono presenti anche le "variabili risultato" di processi che si manifestano nel tempo lungo il ciclo di vita (in longitudinale).

Verma e Betti (2006) hanno illustrato una possibile tipologia della struttura dei dati per l'indagine EU-SILC. La loro proposta fa riferimento ad un'unica fonte integrata che copra tutte le componenti, trasversali e longitudinali, di reddito e sociali.

L'idea di base si fonda sul fatto che in un certo istante sono presenti 4 sottocampioni, o panel, di breve periodo. Ogni anno viene aggiunto un nuovo panel che deve rimanere nel campione per una durata pari a 4 anni, dopo di che viene eliminato e rimpiazzato da un nuovo panel.

I soggetti del campione originario che si trasferiscono, vengono seguiti presso la loro nuova locazione fin tanto che il loro panel rimane nel campione[1]. Ogni panel fornisce un campione longitudinale di durata prestabilita.

Le unità che sono presenti in tutti i panel, in un certo istante, costituiscono il campione trasversale. Il vantaggio maggiore di questo schema è costituito dal fatto che sia i dati longitudinali che quelli trasversali sono ottenuti dallo stesso insieme di unità statistiche.

Questa sovrapposizione è conveniente sia in termini economici, sia in termini di massima coerenza tra le statistiche longitudinali e trasversali computate.

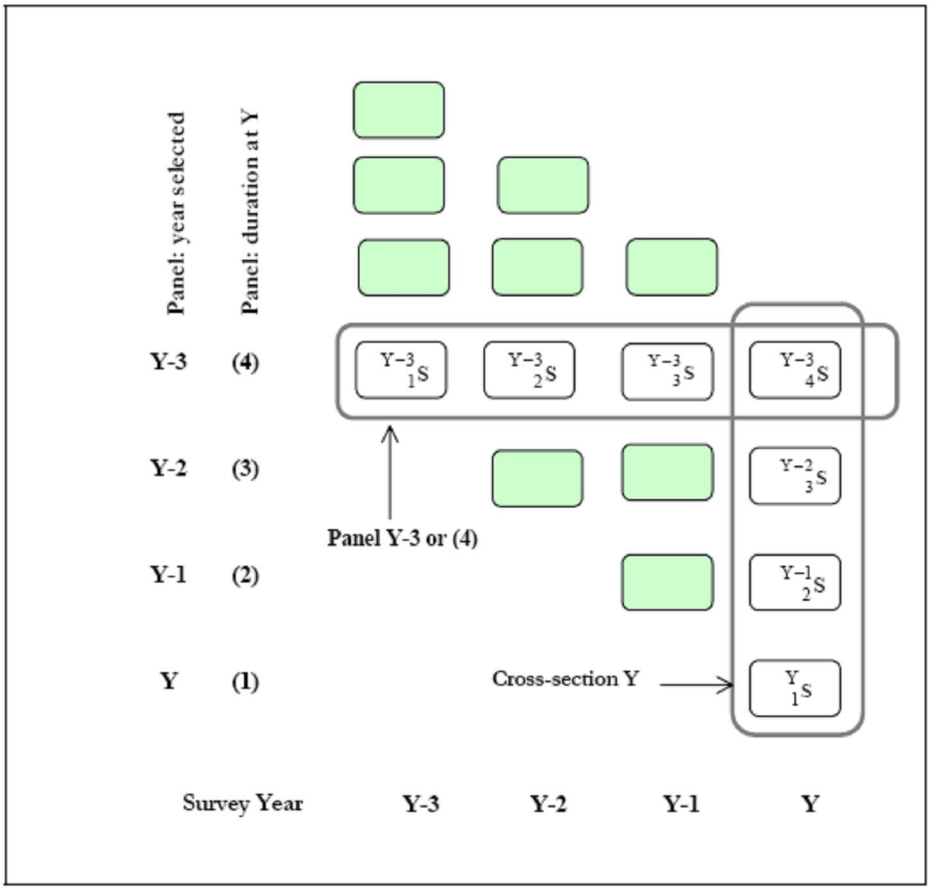
Lo schema di rotazione è illustrato nella Figura App. 1.1.

Nel diagramma seguente ${}^{Y-3}{}_{1}S$ è un tipico panel introdotto nel campione nell'anno Y-3; questo campione, più precisamente il campione ottenuto seguendo precise regole di tracciabilità, è enumerato per un periodi di 4 anni, da Y-3 a Y. I campioni longitudinali di durata da 2 a 4 anni, sono costruiti mettendo insieme panel diversi di questo tipo. Il campione trasversale completo all'anno Y è composto da quattro panel ${}^{Y-t+1}{}_{t}S$, t=1,…,4, come mostrato in figura (App. 1.1).

Da un anno al successivo, alcune replicazioni vengono conservate, mentre altre vengono eliminate e sostituite da quelle nuove.

[1] Secondo la *Commission Regulation* (EC) N°1982/2003 *as regards the Sampling and Tracing Rules*, il disegno rotazionale si riferisce alla selezione del campione basata su un numero di sottoccampioni o replicazioni, ognuno di questi con dimensioni simili e rappresentativi dell'intera popolazione.

Figura App. 1.1. Il panel e il campione trasversale



Fonte: Verma, Betti e Ghellini (2007)

App. 1.8. Costruire un campione con sovrapposizione

Di seguito viene descritta in pratica la procedura per costruire un campione che presenta un certo grado di sovrapposizione tra un anno e quello successivo.

Consideriamo due anni di seguito con campioni che parzialmente si sovrappongono.

Per il campione trasversale, ogni anno, affinché il campione sia rappresentativo, è necessario che ognuna delle tre componenti sotto riportate sia un campione rappresentativo:

- la parte del campione cancellata, deve essere rappresentativa della popolazione per il primo anno;
- la parte del campione che si aggiunge deve essere rappresentativa della popolazione del secondo anno;

- la parte in comune deve essere rappresentativa della popolazione comune ai due anni.

Generalmente, questo si ottiene selezionando l'intero campione attraverso una serie di replicazioni. Ogni replicazione deve essere un campione rappresentativo, solitamente con lo stesso disegno (struttura, stratificazione, allocazione, eccetera) del campione intero, differenziandosi da quest'ultimo solo per la numerosità.

Da un anno all'altro, alcune delle replicazioni vengono conservate, mentre altre vengono sostituite da nuove replicazioni, a seconda dell'entità della sovrapposizione desiderata.

La tecnica di selezionare il campione con una serie di replicazioni indipendenti ma simili tra di loro per la numerosità, il disegno, e il fatto di rappresentare l'intera popolazione di riferimento, fornisce flessibilità e controllo del campione ruotato.

La figura di seguito riportata illustra un semplice disegno di rotazione. Per ogni anno, il campione è costituito da replicazioni, che sono state all'interno dell'indagine per 1-4 anni. Ogni particolare replicazione rimane nell'indagine per 4 anni; ogni anno, una delle 4 replicazioni viene rimossa e sostituita da una nuova, generando così una sovrapposizione del 75% da un anno all'altro. Considerando invece due anni dell'indagine, la sovrapposizione è del 50%, e si riduce al 25% se consideriamo tre anni, e zero per intervalli più lunghi.

Per l'EU-SILC viene applicato questo modello che risulta anche il più semplice e il più appropriato per monitorare i cambiamenti di anno in anno.

App. 1.9. Strategia di campionamento e precisione delle stime

Per quanto riguarda il disegno campionario, la sua progettazione è stata guidata dai requisiti, stabilita da Eurostat, relativamente ai diversi aspetti della rilevazione, ovvero: la tipologia dei parametri che l'indagine deve produrre, la cadenza e il periodo di riferimento dei quesiti, nonché la precisione di alcune stime obiettivo di tipo trasversale o longitudinale. Per la comparabilità delle stime dei diversi paesi in termini di precisione, Eurostat impone una numerosità campionaria minima sotto l'ipotesi di campionamento casuale semplice; la definizione della numerosità campionaria da realizzare, sulla base della quale viene pianificata l'indagine, deriva poi dalle ipotesi sul

design effect connesso con il disegno di campionamento attuato dai diversi Istituti, nonché dai tassi di risposta attesi per la rilevazione. Nel caso di un'indagine come quella EU-SILC, dove la rilevazione ha anche natura longitudinale, la valutazione dei tassi di risposta richiede anche la specificazione di un andamento dell'*attrition*[2].

Definiamo adesso più precisamente la popolazione trasversale e quella longitudinale. La prima si riferisce alle unità esistenti ad una prefissata data di ciascun anno di rilevazione, la seconda invece è riferita ad un intervallo di tempo, ed è quindi necessario tenere in considerazione gli aspetti dinamici della popolazione, ossia tutti i fenomeni di entrata e uscita di unità nell'intervallo di tempo considerato.

Un aspetto importante per la specificazione della rilevazione riguarda l'interesse di EU-SILC per variabili di natura differente, ossia il reddito e le condizioni socio-economiche. Queste due tipologie di variabili possono provenire da fonti diverse, purché tra loro abbinabili. Ad esempio, la rilevazione sui redditi potrebbe provenire da archivi amministrativi, qualora disponibili, mentre un'indagine campionaria potrebbe essere effettuata solo per le variabili del secondo tipo.

Un altro aspetto rilevante è quello relativo alla scelta di un disegno di indagine che fosse in grado di produrre stime corrette ed efficienti, sia dei parametri trasversali che di quelli longitudinali. Anche in questo caso, data l'assenza di vincoli da parte di Eurostat, era possibile adottare uno dei seguenti disegni alternativi:

1) Due rilevazioni separate: una prima rilevazione trasversale a campioni indipendenti o basata sul campionamento ruotato (l'opzione con campionamento ruotato era maggiormente auspicabile per migliorare la precisione delle stime di variazione netta), un seconda rilevazione longitudinale basata su un panel ruotato con o senza sovrapposizione;

2) Un'unica rilevazione integrata, basata su un panel ruotato annualmente con sovrapposizione, mediante la quale ottenere sia le stime di natura trasversale sia quelle di tipo longitudinale.

La scelta di numerosi paesi, tra cui l'Italia, è stata quella di utilizzare un'indagine integrata per la componente longitudinale e trasversale, anche per ridurre i costi complessivi della rilevazione.

Consideriamo adesso le numerosità campionarie.

[2] Ossia l'assenza di risposta dell'unità campionaria a partire da un'occasione di indagine.

Per quanto riguarda la definizione della numerosità campionaria per la componente trasversale, Eurostat ha fissato la numerosità minima (in termini di famiglie) dei campioni nazionali sulla base della precisione della stima della percentuale di famiglie povere a livello nazionale, nell'ipotesi di campionamento casuale semplice.

La definizione del numero minimo stabilito in generale da Eurostat è basato sui seguenti elementi:

- la percentuale di famiglie povere nei paesi EU varia nell'intervallo 5-25 per cento ed è stato pertanto considerato come valore di riferimento una percentuale intermedia pari a p=15 per cento;
- si è supposto un effetto del disegno pari a 1, ossia si è ipotizzato un disegno di campionamento casuale semplice di unità finali;
- si è definita la precisione della stima prefissando l'ampiezza dell'intervallo di confidenza della stima di p pari a 2 punti percentuali; a tale ampiezza corrisponde l'intervallo (14-16 per cento) e un errore relativo del 3,4 per cento.

Ogni paese ha dovuto stabilire la numerosità campionaria effettivamente da selezionare sulla base della valutazione dell'effetto del disegno ($deft^2$) relativo al disegno di campionamento prescelto. Questo tiene conto dei tassi attesi di mancata risposta totale e dell'impatto sull'efficienza delle stime indotto dalla stratificazione, dal *clustering* e dalla ponderazione.

Sia n_e la numerosità campionaria minima effettiva e sia n_a la numerosità campionaria necessaria per garantire la precisione prefissata delle stime, ottenuta tenendo conto del disegno di campionamento adottato; tra le due quantità sussiste la seguente relazione:

$$n_e = n_a / deft^2 \qquad \text{(App. 1.2)}$$

Per scegliere il numero di famiglie campione da selezionare, n_s, è necessario tenere conto anche del tasso di risposta atteso R; risulta quindi:

$$n_s = n_a / \mathrm{R} \qquad \text{(App. 1.3)}$$

Relativamente alla numerosità longitudinale questa è stata definita da Eurostat come il numero di famiglie intervistate con successo in ogni anno e di cui la maggior parte dei membri sono intervistati anche nell'anno successivo. Per la definizione di tale numerosità si è proceduto analogamente alla definizione della numerosità trasversale.

App. 1.10. Strategie di controllo e correzione delle informazioni

Come è facile intuire, l'abbondanza di informazioni, ed anche la complessità di alcuni percorsi di intervista, richiedono un accurato ed esteso controllo incrociato delle informazioni; i dati infatti vengono sottoposti a particolari procedure di identificazione e correzione degli errori e, là dove necessario, all'imputazione dei valori mancanti[3] (solitamente attraverso modelli di regressione multipla disponibili nel software *IVEware*).

Progettato come modulo aggiuntivo di SAS, *IVEware* consente di trattare con relativa semplicità un insieme complesso di dati correlati fra loro, nei casi in cui sarebbe troppo oneroso esplicitare un modello completo di relazioni multivariate. In sintesi la tecnica consiste nella generazione, per ogni valore mancante di ciascuna variabile, di una predizione condizionale ai valori di tutte le altre variabili. In un primo passo preliminare, il modello imputa in sequenza i valori mancanti di ognuna delle variabili, iniziando da quella che ne presenta una minore percentuale. Una volta che ne siano stati imputati i valori mancanti, ciascuna variabile entra come covariata nel processo di imputazione di tutte le altre. Nei passi successivi, dal secondo fino alla convergenza dei valori imputati, il modello è costituito da tante equazioni quante sono le variabili da imputare. Quando è raggiunta la convergenza delle stime, ogni equazione predice i valori mancanti della variabile dipendente. Il principale vantaggio del software *IVEware* consiste nella flessibilità d'uso: per ogni variabile da imputare è possibile sia selezionare le covariate rilevanti, sia definire il modello di regressione appropriato (lineare per le variabili continue, logit per le dicotomiche, log-lineare per le variabili discrete). Inoltre, è possibile vincolare le imputazioni al rispetto di vincoli di coerenza (per esempio, imputare soltanto un particolare sottoinsieme di casi) e assegnare limiti superiori ed inferiori ai valori da imputare.

Quest'ultima caratteristica è utile per trasformare in valori puntuali le cifre approssimate indicate dagli intervistati che non ricordano l'importo esatto dei redditi. In questi casi, i valori puntuali sono imputati con *IVEware* all'interno di una banda centrata sul valore approssimato.

[3] La tematica delle tecniche di imputazione per dati mancanti è approfondita nel Capitolo 8 della dispensa. In questo paragrafo vengono brevemente descritte le azioni intraprese per l'indagine EU-SILC.

Nel caso delle variabili qualitative dell'indagine EU-SILC, può essere impiegato un algoritmo di ricerca del “donatore” più simile all'individuo che presenta informazioni mancanti o errate. La similitudine fra individui è valutata con riferimento alle principali caratteristiche individuali (età, sesso, titolo di studio...) e alle variabili correlate con il dato da imputare. Per esempio, oltre ai dati anagrafici si considera anche la somiglianza nelle condizioni lavorative (settore, qualifica professionale, anni di anzianità...) quando si deve imputare tramite donazione una mancata risposta relativa al tipo di contratto di lavoro (a tempo determinato o indeterminato). Le informazioni dell'individuo più somigliante vengono in pratica copiate al posto dei valori errati o mancanti.

Per l’indagine EU-SILC, gli effetti delle imputazioni vengono controllati attraverso il confronto dell’intero profilo distributivo, prima e dopo l’intervento di correzione (cioè tutti i percentili, attraverso la sovrapposizione dei grafici delle distribuzioni di frequenza semplici e cumulate). Inoltre, gli effetti delle imputazioni sono valutati comparando i principali parametri della distribuzione (media, mediana, primo e terzo quartile, minimo, massimo, *standard error*) disaggregati per sesso, età, area geografica, titolo di studio, condizione professionale ed altre covariate significative.

App. 1.11. Strategie di ponderazione per le stime

Eurostat ha inoltre definito delle linee guida sia per il calcolo dei pesi di riporto dei dati campionari all’universo di riferimento, sia per le procedure della gestione delle mancate risposte.

Con riferimento a queste ultime è previsto un sistema di imputazione e/o assegnazione di opportuni pesi. In particolare il regolamento stabilisce quali sono le caratteristiche auspicabili dei modelli statistici atti ad imputare il valore assunto dalle variabili target (come ad esempio il reddito percepito dagli individui che concorre a formare il reddito disponibile della famiglia), qualora questo non sia disponibile.

L’indagine, nella sua componente trasversale, deve produrre sia le stime riferite al numero di individui (o famiglie) che nella popolazione di riferimento possiedono una certa caratteristica, sia il livello di una quantità misurata sugli individui (o famiglie), come ad esempio il reddito.

Una tra le metodologie utilizzate in questa fase, è caratterizzata dall'uso di una famiglia di stimatori noti in letteratura come *calibration estimator* (stimatori di ponderazione vincolata), consente la determinazione di un unico coefficiente di riporto all'universo in grado di produrre stime coerenti a totali noti, desunti da fonti esterne, sia per individui che per famiglia assegnando cioè, lo stesso coefficiente di riporto all'universo a tutti gli individui della stessa famiglia (calibrazione integrata).

I dati campionari devono essere pesati affinché siano il più rappresentativi possibile della popolazione che riproducono. È necessario un sistema di pesi integrato in quanto sono presenti diversi tipi di dati provenienti annualmente dal panel rotato. Devono essere costruiti dei pesi iniziali da applicare ad ogni nuovo campione introdotto nell'indagine. Per la procedura dei pesi è stato introdotto un concetto innovativo, il concetto dei pesi base. Cominciando dal peso che ogni individuo ha nel campione iniziale, il peso base dell'individuo viene costruito per ogni anno successivo al primo per compensare l'attrito. L'obiettivo finale, consiste comunque nella costruzione di pesi trasversali e longitudinali da utilizzare nell'analisi dei dati. La costruzione di questi due tipi di pesi si basa sullo stesso peso base, e ciò fa sì che l'intero sistema dei pesi sia integrato.

Consideriamo adesso le stime trasversali che si basano su dati provenienti da un campione composto dall'unione di quattro campioni longitudinali, ognuno per la sua specifica *wave*. In tal modo, ogni campione trasversale è composto da un quarto di famiglie che partecipano per la prima volta, un quarto di famiglie che partecipano per la seconda volta, un quarto di famiglie che partecipano per la terza volta e infine un quarto di famiglie che partecipano per la quarta volta all'indagine.

Il principio su cui è basato ogni metodo di stima campionarie è che le unità appartenenti al campione rappresentino anche le unità della popolazione di riferimento che non sono incluse nel campione stesso. A tale scopo, ad ogni unità campionaria viene attribuito un peso, coefficiente di riporto all'universo, che indica quante unità della popolazione sono rappresentate, rispettivamente, da ogni unità presente nel campione.

App. 1.12. Il sistema dei pesi longitudinali in EU-SILC[4]

WEIGHTING

I. Introduction

The objective of the present section is to outline 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.

According to the Commission Regulation on sampling and tracing rules (EC No 1982/2003, §7.4): *Weighting factors shall be calculated as required to take into account the units' probability of selection, non-response and, as appropriate, to adjust the sample to external data relating to the distribution of households and persons in the target population, such as by sex, age (five-year age groups), household size and composition and region (NUTS II level), or relating to income data from other national sources where the Member States concerned consider such external data to be sufficiently reliable.*

On the other hand, the Framework Regulation (No 1177/2003) pointed out in its article 10 that: *Member States shall transmit to the Commission (Eurostat) in the form of micro-data files weighted cross-sectional and longitudinal data which has been fully checked, edited and imputed in relation to income.*

II. Weighting for the first year of each sub-sample

1. Design weights (Household weights DB080 and "Selected respondent" weights PB070)

These weights are of methodological interest, but are not used in substantive analysis. Actually, the design weights need to be defined for all selected units, and not only for responding units. DB080 is computed as follows:

- In case of households are sampled (or addresses or other units containing households):

$$DB080_h = = 1 / (\text{probability of selection of h})$$

[4] Questo paragrafo costituisce un **approfondimento**. Si tratta di parte di un documento ufficiale di Eurostat, ampiamente tratto da Verma, Betti e Ghellini (2007).

- <u>In case of persons are sampled</u>:

$$DB080_h = 1 / \Sigma \text{ (probabilities of selection of eligible persons in h)}$$

"Eligible persons" are persons who are given a non-zero probability in the selection procedure, such as persons aged 14+ or 16+. In the particular situation where the probability of selection is the same for all eligible persons in each household, the denominator is simply the number of such persons in the household multiplied by that probability of selection.

When households or persons are selected from lists which contain "blanks", i.e. non-existing or unoccupied structures, the unit not being a private household, the household or person listed does not exist…, it is important to ensure that the selection probabilities are computed correctly. For instance, if there are N listings which contain N' actual units, and an equal probability sample of n listing yielding n' actual units is selected, then the selection probability is:

$$\pi_i = n'/N' \quad \text{if } N' \text{ is known}$$
$$= n/N \quad \text{otherwise (the most common case)}$$

PB070 is defined only when a sample of persons is used, for the selected respondent k:

$$PB070_k = 1 / \text{(probability of selection of the selected respondent k)}$$

2. Adjustments for non-response

The principle is to adjust the household design weights to allow for the bias which is caused when all measured variables are missing for some of the sample households. The main reasons for not having information are the household refuses to cooperate or is temporarily away. However, other factors may cause household non-response: the data collected are of insufficient quality, the questionnaire has been lost… Non-response is particularly critical when the non-responding households over-represent survey characteristics, which may create substantial bias in the estimates. For instance, one can admit that households with high incomes would be less willing to cooperate to income surveys than households with low incomes.

This step involves estimating response rates or propensities to response as functions of characteristics available for responding and non-responding households, and also characteristics of the areas where the households are located. Basically, the design weights have to be inflated by the inverse of the response propensities in order to compensate for the loss of units in the sample.

A classical procedure consists in modifying the design weights by a factor inversely proportional to the response rate within each "homogeneous group", wherein the response probabilities are assumed to be equal:

$$DB080_h^{(N)} = DB080_h \cdot \frac{1}{R_K}$$

Where R_K denotes s the (weighted) response rate in the group k the household h belongs to:

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

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.

An alternative to estimate response probabilities 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. A very important point when using the regression approach is to ensure that weights assigned are confined to be within reasonable limits.

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.

When a sample of persons is involved, exactly the same non-response adjustment as above for the household level applies to the selected respondent as well. Let $DB080^{(N)}$ the household weight after final non-response adjustment; then the selected respondent non-response weight is given by:

$$PB070^{(N)} = PB070 \cdot \left[\frac{DB080^{(N)}}{DB080} \right]$$

This follows from the fact that household and personal non-responses always occur together according to EU-SILC interview acceptance procedures. The household interview or data compilation is accepted only when the personal interview has been accepted for the selected respondent.

3. Adjustment to external sources (calibration): SILC target variables DB090 and PB060

The key feature of this step is the modification of the household weights $DB080^{(N)}$ to reproduce from the sample population characteristics, namely totals and category frequencies. For example, in a human population survey, age and sex are natural ancillary variables. The distribution of the human population by age and sex is often known from other statistical sources such as a census or a population register and by proper modification of the survey weights, the population structure may be exactly reproduced by the sample. For variables in the survey correlated with the ancillary information, higher precision in estimates is usually obtained on application of the new calibrated weights.

More precisely, suppose that there exist J auxiliary variables $x_1 \ldots x_j \ldots x_J$, called calibration variables, with known population totals (for the numerical variables) or marginal counts (for the categorical variables). Without loss of generality, we can assume that all the calibration variables are numerical (otherwise, we consider the 0/1 variables for each category).

We seek new household weights DB090 that are "as close as possible" (as determined by a certain distance function) to the initial weights $DB080^{(N)}$. These new weights are calibrated on the totals X_j of the J auxiliary variables; in other words they verify the calibration equations:

$$\forall\, j = 1 \ldots J \qquad \sum_{k \in s} DB090_k \cdot x_{jk} \;=\; X_j \,.$$

Where $DB090_k = g_k \times DB080^{(N)}$

The SAS macro CALMAR, developed in the French Statistical Office (INSEE), can calculate calibrated weights.

When using CALMAR, it is recommended to use a bounded method and to impose lower and upper bounds LO and UP on the weight adjustment factors g_k, usually referred to as g-weights. In practice, one has to keep in mind that the choice of the bounds is not free and directly depends on the calibration variables which are chosen: the limits must be adjusted taking into account the differences between the estimates based on the "old" initial weights and the benchmark totals that the new weights are to reproduce, so CALMAR can find a solution within the constraints applied to the problem. In practice, those limits are determined by some "guess and check": we start with a small interval [LO,UP] and we enlarge it until CALMAR finds a solution. Putting calibration bounds prevents from negative and extreme weights. Extreme

weights can lead to unexpected values especially for domain estimates. Negative weights are not acceptable from a practical point of view.

At this stage, Eurostat recommends an “integrative” calibration. The idea is to use both household and individual external information in a single-shot calibration at household level. The individual variables are aggregated at household level by calculating household totals such as the number of male/female in the household, the number of persons aged of 16 and over…..The calibration is done then at household level using household variables and the individual variables in their aggregated form. This technique ensures "consistency" between household and individual estimates by making the household and the individual weights equal.

In the framework of calibration, it is critical that the external control variables are strictly comparable to the corresponding survey variables, the distribution of which is being adjusted. For instance, EU-SILC micro data must not be calibrated on the basis of ILO LFS counts if ILO status is not measured properly in EU-SILC.

When a sample of persons is involved, the final household weights DB090 determined above can be used to compute the corresponding weights PB060 for the selected respondents:

$$\mathrm{PB060} = \mathrm{PB070} \cdot \left[\frac{\mathrm{DB090}}{\mathrm{DB080}} \right]$$

4. Personal weights (SILC target variables RB050 and PB040)

The calibrated household weight is assigned to each of its members $\mathrm{RB050}_{j \in h} = \mathrm{DB090}_h$. The weights PB040 are derived simply by filtering RB050 to the persons who have received an individual questionnaire (PB040 = RB050). This is based on the fact that for responding households all the individual questionnaires are completed. If individual non-response is restricted, Eurostat recommends indeed imputing individual records at least for individual income components. In this case, personal weights should not be adjusted for individual non-response and consistency between total income and income components is preserved.

III. <u>Computation of base weights</u> (SILC target variables RB060, PB050 and PB080)

The base weights are the back spine for the computation of both cross-sectional weights and longitudinal weights. They are computed and updated for a single panel and, as such, they will rarely be used for estimating population parameters. The cross-sectional and longitudinal weights are obtained by combining the base weights in an appropriate way, which will be described later.

In the following we consider a panel (sub-sample, rotation group) selected fresh at time t=1 from population P_1, and then enumerated for a total of 4 waves, t=1 to 4. Let s_1 be the sample of all persons in the households enumerated at t=1.

For each person in this set, we define the personal base weight at wave t=1 as:

$$\omega_1^{(RB)} = RB060 = RB050$$

Similarly, we define (when applicable) the personal base weight at wave t=1 for selected respondent as:

$$\omega_1^{(SB)} = PB080 = PB060$$

At each subsequent wave, persons have left the population between years t and t+1, due to death, migration out of the country, movement out of the private household sector to an institution or collective household, or have become excluded from the target population for any other reason. We also have to deal with total non-response (attrition) when, for a person who still is in the target population, the measured variables are missing. Some possible causes of missingness are:

- No contact
- Refuse to participate to the survey
- Information not available
- Unable to trace a unit that has moved
- Questionnaires lost

"Base weights" at subsequent waves are obtained by adjusting for attrition base weights from wave 1. In general, in order to determine base weight $\omega_t^{(RB)}$ (t = 2, 3 or 4) from known $\omega_{t-1}^{(RB)}$, we can use the following procedure. Consider the set of persons, denoted s_t, enumerated at (t-1) who are still in-scope at t. For each person j in this set, we define the following binary variable:

- r_j=1 if the person is in s_t, i.e. 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. For any person j in s_t, the required weight is:

$$RB060 = \omega_{t,j}^{(RB)} = \frac{\omega_{t-1,j}^{(RB)}}{p_j} \text{ for wave t>2.}$$

The application of the above procedure requires that for each person enumerated at (t-1), the person's status at t is precisely known. This means that each person in the panel at (t-1) can be classified into one of the following categories uniquely:

1. enumerated at t
2. remains in the population, but not enumerated at t
3. moves out of the population

In practice, for a proportion (of non-enumerated persons) it cannot be determined whether they belong to (2) or (3). Each such person has to be assigned to one or the other of these two groups on the basis of some appropriate exogenous information or model. This may be done, for instance, on the basis of a logit regression model determining the person's propensity to remain in the population as a function of a set of auxiliary variables.

In so far as most non-response occurs at the household level, a majority of the relevant auxiliary variables will be geographical and household level variables (region, household size and type, tenure); also constructed variables (household income, household work status, …). Some personal variables are also likely to be useful (gender, age, employment status,…) – the sort of variables correlated with persons moving to new address, setting up a new household, remaining traceable, etc. 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.

There are certain (small) categories of households and individuals which, according to EU-SILC rules, are not followed-up. Examples are households not enumerated at wave 1 or for two consecutive waves thereafter, or even not enumerated at a single wave for some specified reasons. Also, persons below a certain age (under 14, or under 16 in some countries) are not followed up if they move "alone", i.e. without being accompanied by an adult sample person. For the present purpose, all these categories are treated as non-respondents – even if these have not been recorded as such in the survey because of particular EU-SILC tracing rules. The above applies to all household members covered in R-file, including persons aged 16+, and irrespective of whether an initial sample of households or persons has been used.

For the personal interview sample (P-file), the above provide the starting weights but a further adjustment is required depending on the type of the sample. For a sample of households, the adjustment arises from within-household non-response (which in most countries is very small). First, base weights $\omega^{(RB)}$ are applied to the completed (and accepted) personal interview sample. Then the results are calibrated on gender and age (in single years) according to the distribution of the R-file sample aged 16+ weighted by the same base weights. The resulting weights $\omega^{(PB)}$ = PB050 for the completed individual interview sample are these post-calibration weights. The result is that the P-file sample gives the same age-sex distribution, as the R-file sample for persons aged 16+.

For a sample of persons, there is no “within hh non-response”. For income and other data compiled for all hh members aged 16+, the already computed base weights are used unchanged, which gives PB050=RB060.

For non-income variables collected through the personal interview with selected respondents (one per household), the personal interview data weighted by:

$$\text{PB080} = \omega_t^{(SB)} = \omega_{t-1}^{(SB)} \cdot \left\{ \frac{\omega_t^{(RB)}}{\omega_{t-1}^{(RB)}} \right\}, \; t \geq 2 \,,$$

are calibrated on gender and age (in single years) according to the distribution of the total sample aged 16+ weighted differently, namely by $\omega^{(RB)}$. The resulting weights for the completed individual interview sample are these post-calibration weights.

There remain some additional categories of persons to be considered:

- Children born to sample women. They receive the weight of the mother.
- Persons moving into sample households from outside the survey population. They receive the average of base weights of existing household members.
- Persons moving into sample households from other non-sample households in the population – these are “co-residents” and are given zero base weight.

These are also persons who had a base weight to begin with, then were not enumerated for one wave, but subsequently returned as the sample persons not enumerated in wave 2 but returning in wave 3; and those not enumerated in wave 3 but returning in wave 4. Since during their absence, base weights of other persons in the sample are adjusted to take into account that absence; on return these persons cannot be re-assigned a positive weight without adjusting the weights of other persons. Hence it is convenient at this stage (during their absence) to retain a zero base weight for such “returnees”.

IV. <u>Cross-sectional weights, year 2 onwards</u>

In the next figure the rotation groups in a rotational design are represented and particularly the structure of the cross-sectional sample at each year (in bold):

panel introduced in Year							
	X						
	X	X					
	X	X	X				
Y-3	X	X	X	**X**			
Y-2		X	X	**X**	X		
Y-1			X	**X**	X	X	
Y				**X**	X	X	X
SURVEY YEAR				Y			

The cross-sectional sample at Y is clearly representative of the target cross-sectional population at Y, through the selection of a sub-sample at Y in this population. On the other hand, the three "old" sub-samples do not represent some "immigrants" entering the target population. The figure below summarizes the situation for each panel.

Panel introduced in year	Sample and weight	Population
Y	$(s_{1,}\omega_1^{(RB)})$	P_Y
Y-1	$(s_{2,}\omega_2^{(RB)})$	$P_Y - IN_Y^{(new)}$
Y-2	$(s_{3,}\omega'^{(RB)}_3)$	$P_Y - (IN_Y^{(new)} + IN_{Y-1}^{(new)})$
Y-3	$(s_{4,}\omega''^{(RB)}_4)$	$P_Y - (IN_Y^{(new)} + IN_{Y-1}^{(new)} + IN_{Y-2}^{(new)})$

- P_Y is the target cross-sectional population at Y.
- $IN_Y^{(new)}$ is the population entering the target population and forming separate household (no initial population member) during the year preceding Y.
- s_k is the sample enumerated in k-th year of a specified panel, for example in year 3 of panel Y-2 in the third row above.
- $\omega_k^{(RB)}$ is the corresponding base weight at k-th year of the specified panel. (RB) indicates that the reference is to base weights of the total population (R-file).
- $\omega'^{(RB)}_3$ are the base weights $\omega_3^{(RB)}$ at t = 3, modified to incorporate re-entries into the sample. This refers to sample persons who were – of course – present at t = 1, not present in the sample at t = 2, but are re-enumerated again at t = 3. $\omega'^{(RB)}_3$ is an extension of $\omega_3^{(RB)}$: it gives a non-zero weight to returnees at t = 3. This requires that the $\omega_3^{(RB)}$ weights of all other sample persons have to be adjusted (see annex).
- $\omega''^{(RB)}_4$ These are a modification of base weights $\omega_4^{(RB)}$ at t = 4, as will be defined below.

We use a simple procedure to estimate first adjustment to the base weights $\omega_4^{(RB)}$:

$$\omega_4'^{(RB)} = \left(\frac{\omega_3'^{(RB)}}{\omega_3^{(RB)}}\right) \cdot \omega_4^{(RB)}.$$

Then, in the same way as going from $\omega_3^{(RB)}$ to $\omega'^{(RB)}_3$ to accommodate returnees at t = 3, we go from $\omega'^{(RB)}_4$ to $\omega''^{(RB)}_4$, adjusting for returnees at t = 4.

In order to put the four cross-sections together we first multiply the weights of units according to their origin (initial population or immigrants at previous wave) in order to take into account the number of times the subpopulation they refer to is represented in the different panels. We have:

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

- $IN_{Y-2}^{(new)}$ by 4/3

- $IN_{Y-1}^{(new)}$ by 4/2=2

- $IN_Y^{(new)}$ by 4/1=4

Let ω_j be the weight of unit j after the above mentioned modification. Within a household, each member j has been assigned a weight ω_j, except for "co-residents", i.e. for every household members who are not eligible for inclusion in the panel, for whom ω_j=0. Average of these weights over all household members (including co-residents) is assigned to each member (including co-residents). We recommend applying this averaging process to all households, including households not containing a co-resident. Finally, the four panels are combined and the weights are scaled by a factor of 4:

$$\omega'_j = \omega_j /4$$

The last step will be to calibrate these weights against external standards; using the approach already described in II. Integrative calibration will ensure that members in the same household all receive the same weight. Household cross-sectional weight is the same as the average personal cross-sectional weight: DB090=RB050

By replacing $\omega^{(RB)}$ by $\omega^{(PB)}$ throughout, we obtain personal cross-sectional weight PB040. Similarly, by replacing it by $\omega^{(SB)}$, we obtain the selected respondent cross-sectional weight PB060.

Remark: trimming

Trimming refers to recoding of extreme weights to more acceptable values. The objective of trimming is to avoid excessive increase in variance due to weighting, and possibly give rise to influential data, even though the process introduces some bias. The aim is to seek a trimming procedure which reduces the mean squared error. Basically, at each step of the weighting procedure, the distribution of the resulting weight adjustments should be checked.

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:

- $\omega_i^{(1)}$ = weight before adjustment (non-response, calibration…)

- $\omega_i^{(2)}$ = weight determined after adjustment

- $\overline{\omega}^{(1)}, \overline{\omega}^{(2)}$ their respective mean values,

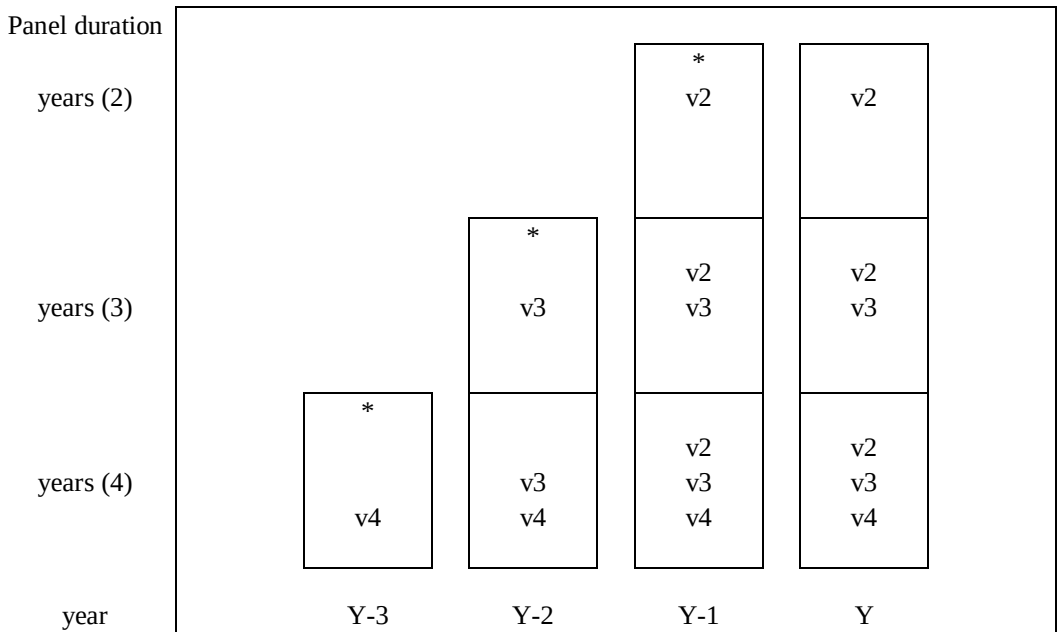
Any computed adjusted weights outside the following limits should be recoded to the boundary of these limits:

$$1/C \leq \frac{\omega_i^{(2)} / \overline{\omega}^{(2)}}{\omega_i^{(1)} / \overline{\omega}^{(1)}} \leq C$$

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.

V. <u>Longitudinal weights</u> (SILC variables RB062 and RB063)

We 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.



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

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 diagram, 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.

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

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

* 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

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

Then the required target variables can be computed as follows: $RB062_j = \omega_j / 3$ where ω_j is the weight for any unit j as defined above.

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

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

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

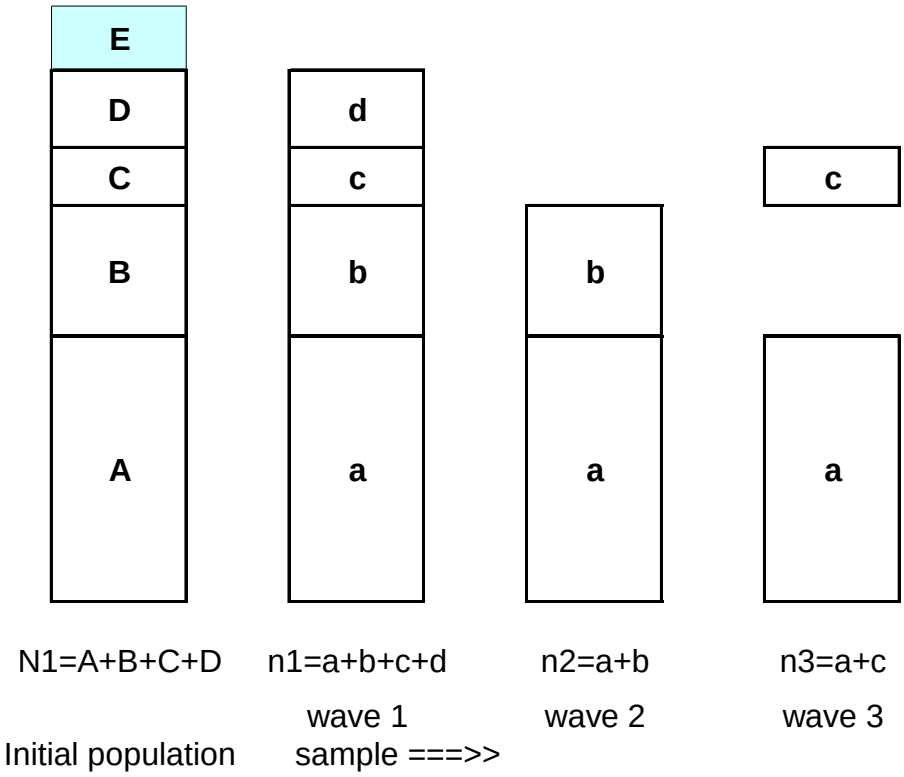
$$RB063_j = \omega_j / 2$$

For the four year panel, (Y-3) to Y: the modified base weights $\omega_4^{'(RB)}$ directly give the required target weights.

Longitudinal set of four years duration, for (Y-3) to Y

Only panel 4 with data from years (Y-3) to Y contributes to this set and the longitudinal weight is given by $\omega_4^{'(RB)}$ = RB060

ANNEX: Weighting of re-entries



---

1. In order to explain the basic procedure, we consider the following situation.

A sample is selected from the initial population at wave 1. We conceptualise the population as being divided into 5 parts, A to E, according to its potential response status.

A = part of the population which potentially responds at all three waves, W=1 to 3.

B = potential respondents at W=1 and W=2, but not at W=3

C = potential respondents at W=1 and W=3, but not at W=2 (re-entries)

D = potential respondents at W=1, but not at any subsequent wave

E = potential non-respondents at W=1.

The last mentioned group (E) are not followed up in EU-SILC. They only affect the sample weights at W=1, but not thereafter. Hence they are of no further interest in this note.

Let us assume for simplicity that all quantities above refer to the 'longitudinal' population, i.e. to all units at wave 1 that remain in-scope at waves 2 and 3.

2. Suppose that w1, the (person-level) weights at wave W=1, have been appropriately determined taking into account the design weights, non-response at wave 1, and any calibration adjustments.

Then, wave 2 weights, w2, can be determined from the wave 1 weights, w1, by taking into account non-response between waves 1 and 2. For instance, conditional on appropriately selected auxiliary variables, we determine response propensities

$$r_{12} = P(n_2|n_1),$$

where the right-hand side is an abbreviation indicating the propensity, for the n1 sample units, to be present in sample n2[5]. With the previous wave weights w1 modified as

(1) $$w_2 = w_1 / P(n_2|n_1),$$

the achieved sample (n2=a+b) at wave 2 with weights w2 represents the population N1, just as does n1 with the original weights w1. Similarly, with weights at wave 2 modified as

(2) $$w_3 = \frac{w_2}{P(a|n_2)} = \frac{w_1}{P(a|n_2).P(n_2|n_1)},$$

[5] As noted above, strictly this refers to units from sample n1 which are *still in scope* of the target population at n2.

the 'longitudinal' sample a with weights w3 represents the population N1, just as does n2 with weights w2, and n1 with the original weights w1.

3. The objective is to determine weights w_3' such that n3=(a+c) represents the population N1. Clearly, in terms of the propensity to be in a conditional on being in (a+c), weights w_3 and w_3' relate as:

$$w_3 = w_3' / P(a|n_3), \text{ giving}$$

(3) $$w_3' = \frac{w_2}{P(a|n_2)} = \frac{w_1 . P(a|n_3)}{P(a|n_2) . P(n_2|n_1)}.$$

As noted, the response propensities are determined conditional on appropriately selected auxiliary variables. Generally, we use auxiliary variables referring to wave 1 for estimating $P(n_2|n_1)$, to wave 2 for estimating $P(a|n_2)$, and to wave 3 for estimating $P(a|n_3)$. This means that the three propensities in (3) are defined only for units common to all the three waves, namely for set a. Therefore, (3) can be used to define the required weights only for units in set a.

For the remaining units (c) in sample n3, we may use the following slightly approximate solution.

4. An alternative to (3) is

(4) $$w_3'' = w_1 / P(n_3|n_1).$$

The objective of (4) is the same as that of (3): to provide weights such that n3=(a+c) represents the population N1. However, (4) is less precise than (3), as it goes directly from wave 1 to wave 3 and disregards information specific to wave 2. However, it is determinable for all units n3=(a+c).

We can use (3) for units in the larger set a, and (4) for the remaining (re-entries) c.

5. The introduction of wave 4 causes no further complication in EU-SILC. For units which are enumerated in both waves 3 and 4, we have the required wave 4 weights analogous to (1):

$$w_4 = w_3' / P(n_4|n_3).$$

For units enumerated in waves 2 and 4, but not in wave 3, we have the required wave 4 weights analogous to (4):

$$w_4 = w_2 / P(n_4 | n_2).$$

Note that in accordance with EU-SILC follow-up rules for non-response, no cases are retained in wave 4 which were not enumerated in both waves 2 and 3. Hence no further complications are involved.

Riferimenti Bibliografici

European Community (2003), Regulation (EC) No 1177/2003 of the European Parliament and of the Council. Official *Journal of the European Union*, L165, pp. 1-9.

Verma, V. e Betti, G. (2006), EU Statistics on Income and Living Conditions (EU-SILC): Choosing the survey structure and sample design, *Statistics in Transition*, 7(5), pp. 935-970.

Verma, V., Betti, G. e Ghellini, G. (2007), Cross-sectional and longitudinal weighting in a rotational household panel: applications to Eu-Silc, *Statistics in Transition*, 8(1), pp. 5-50.

Verma, V. e Clemenceau, A. (1996), Methodology of the European Community Household Panel, *Statistics in Transition*, 2(7), pp. 1023-1062.