

## **Nota informativa sull'utilizzo dei file relativi alla componente longitudinale di EU-SILC (2005-2006-2007-2008)**

I file contengono i dati per l'Italia relativi all'indagine campionaria sulle famiglie "Reddito e condizioni di vita", condotta sulla base del Regolamento dell'Unione Europea (n° 1177/2003) che definisce il progetto EU-SILC (European Statistics on Income and Living Conditions).

Questa indagine, i cui risultati confluiscono nei rapporti periodici dell'Unione Europea sulla situazione sociale e sulla diffusione della povertà nei paesi dell'Unione, sostituisce il precedente panel europeo sulle famiglie (ECHP) e ha come obiettivo prioritario quello di fornire, usando definizioni e metodi armonizzati, dati comparabili, sia a livello trasversale che longitudinale, per l'analisi della distribuzione dei redditi, del benessere e della qualità della vita delle famiglie e delle politiche economiche e sociali adottate a livello nazionale e/o europeo.

In particolare, il pacchetto include:

- File dati:
  - L08D.csv
  - L08H.csv
  - L08P.csv
  - L08R.csv
  - DX090.csv.
- Documentazione:
  - Guidelines2008.doc: descrive l'indagine, le variabili e le raccomandazioni operative date da Eurostat agli Istituti di Statistica.
  - SILC L-2008 UDB PROBLEMS AND MODIFICATIONS.xls: descrive modifiche e particolarità rispetto alle indicazioni standard.
  - UDB descrizione variabili italiane aggiuntive.doc
  - Modelli di rilevazione dell'indagine 2005, 2006, 2007 e 2008:
    - Guida\_Intervistatore\_2005/6/7/8.pdf
    - SILC\_REG\_2005/6/7/8.pdf   registro familiare
    - SILC\_FAM\_2005/6/7/8.pdf   questionario familiare
    - SILC\_IND\_2005/6/7/8.pdf   questionario individuale

Nelle pubblicazioni statistiche basate su tale database, si prega di citare come fonte *"ISTAT – Indagine longitudinale sulle condizioni di vita – EU-SILC"*

Già dall'indagine del 2007 il primo campione longitudinale è giunto a conclusione; l'Istat ha arricchito il contenuto informativo dei file aggiungendo altre variabili desunte direttamente dal questionario descritte nel file *"UDB descrizione variabili italiane aggiuntive.doc"*.

La componente longitudinale di EU-SILC rappresenta uno degli elementi più innovativi dal punto di vista metodologico e contenutistico nelle indagini su reddito e condizioni di vita. La struttura longitudinale, se da un lato aumenta le possibilità di analisi con una visione dinamica del fenomeno, dall'altro comporta una maggiore complessità dei metodi e delle procedure per garantire il rispetto della coerenza delle informazioni. La fase del trattamento dei dati longitudinali, intesa come controllo e

correzione per il rispetto della coerenza delle informazioni raccolte, imputazione dei valori mancanti e costruzione dei coefficienti di riporto all'universo, rappresenta uno dei momenti più complessi nella realizzazione di EU-SILC.

Due criteri hanno guidato tutta la fase del trattamento dati in senso longitudinale: il rispetto delle coerenze tra le informazioni riguardanti il reddito e le sue componenti e il minimo cambiamento.

Ogni anno l'aggiunta di una nuova *wave* condiziona la fase del trattamento dei dati. Ad esempio, con riferimento ad uno stesso anno, informazioni riguardanti uno stesso individuo possono cambiare con l'aggiunta di una nuova *wave*. Pertanto, qualora l'utente riscontrasse incoerenze è pregato di segnalarle all'Istituto tramite e-mail ([eusilc@istat.it](mailto:eusilc@istat.it)) inserendo nell'oggetto la parola chiave "LONGITUDINALE".

L'interazione con gli utenti ha l'obiettivo di rendere statisticamente ininfluenti le eventuali incongruenze che di anno in anno possono presentarsi soprattutto in corrispondenza della *wave* finale (la quarta) di ogni campione longitudinale.

### **L'utilizzo dei coefficienti di riporto all'universo**

Per fornire stime tramite un campione longitudinale occorre che queste facciano riferimento alla popolazione dell'anno in cui il campione partecipa per la prima volta all'indagine. Ovviamente, questa popolazione di riferimento deve essere aggiornata al netto delle persone che nel frattempo ne sono uscite (morti, emigrati, trasferiti in istituzione...). La popolazione longitudinale all'anno  $t+1$ , quindi, comprende le persone della popolazione all'anno  $t$  al netto dei fuoriusciti tra l'anno  $t$  e l'anno  $t+1$  ( $OUT_{t+1}$ ).

Ad esempio, la popolazione di riferimento all'anno 2005 del campione partito nel 2004 è:

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

In generale, per un panel all'anno  $n$  che parte nell'anno  $t=1$ , la popolazione longitudinale è pari a:

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

Si noti che la popolazione longitudinale al tempo  $t$  ( $P_t^{(L)}$ ) differisce dalla popolazione al tempo  $t$  ( $P_t$ ) perché non include le persone nate o immigrate nella popolazione di riferimento dal tempo  $t=1$ .

Sia  $C4$  il campione che inizia il suo percorso longitudinale all'anno di partenza dell'indagine (nel nostro caso 2004) e sia  $RB060_{2004}$  il vettore dei coefficienti di riporto all'universo calcolati per il cross-section dell'anno in questione. Per costruzione, questi pesi fanno sì che il campione sia rappresentativo della popolazione di partenza.

$$(C4_{2004}, RB060_{2004}) \rightarrow P_{2004}$$

Al tempo  $t+1$  (2005) il campione longitudinale è composto dal campione iniziale al netto degli usciti dalla popolazione di riferimento ( $out_{2005}$ ) e di quanti non partecipano all'indagine pur avendone ancora i requisiti ( $x_{2005}$ )<sup>i</sup>:

$$C4_{2005}^{(L)} = C4_{2004} - (out_{2005} - x_{2005})$$

Assumendo che i fuoriusciti dal campione pesati siano rappresentativi dei fuoriusciti dalla popolazione abbiamo<sup>ii</sup>:

$$\left( (C4_{2005}^{(L)} + x_{2005}), RB060_{2004} \right) \rightarrow (P_{2004} - OUT_{2005})$$

In generale, per il generico campione  $j$  all'anno  $n$  abbiamo

$$\left( \left( Cj_n^{(L)} + \sum_{t=2}^n x_i \right), RB060_1 \right) \rightarrow \left( P_1 - \sum_{t=2}^n OUT_t \right)$$

Per far sì che il campione rimanente sia rappresentativo della popolazione di partenza al netto degli usciti dalla popolazione di riferimento occorre quindi fare una trasformazione dei pesi di modo che questi tengano anche conto di quanti smettono di partecipare all'indagine pur avendone ancora i requisiti. I pesi dei rimanenti nel campione all'anno  $n$ ,  $RB060_n$ , devono quindi essere tali che:

$$(Cj_n^{(L)}, RB060_n) \rightarrow \left( P_1 - \sum_{t=2}^n OUT_t \right)$$

Ovviamente, la non casualità della mancata partecipazione all'indagine introduce una distorsione nelle stime degli aggregati, e questa distorsione deve essere ridotta per quanto possibile. L'idea di base per la correzione dei pesi anno per anno è stata quella di lavorare sul campione alla prima *wave* per renderlo il più rappresentativo possibile della popolazione di partenza e considerare nel tempo le persone che sono ancora nel campione, inflazionandone i pesi per tenere conto della non casualità dell'attrition. Praticamente, si parte dal peso individuale dell'unità campione e, tramite un processo di aggiornamento che tiene in considerazione i diversi fattori che contribuiscono a spiegare la permanenza degli individui nel panel, si giunge ai nuovi pesi individuali. Si abbandona quindi l'ottica del peso integrato a livello familiare, ma ogni individuo possiede un suo proprio coefficiente di riporto (RB060 e PB050). Si può comunque fare inferenza sulle famiglie di individui panel tramite trasformazione dei pesi dei componenti della famiglia. Per questo motivo si considerano anche gli individui facenti parte della famiglia, ma non appartenenti al campione longitudinale. Ai nuovi nati, si assegna il peso della madre. Per quanti si sono aggiunti alla famiglia, ma facevano già parte della popolazione di riferimento (ad es., individui che appartenevano ad altre famiglie private residenti in Italia) si assegna peso pari a 0, poiché questi erano comunque già considerati nel sistema di ponderazione<sup>iii</sup>.

Si determina in questa maniera quello che viene chiamato *peso base*, che è quindi diverso negli anni anche quando riferito allo stesso individuo, con il quale è possibile fare analisi in ottica longitudinale: non esiste infatti un vero e proprio peso longitudinale, ma questo dipende fondamentalmente dalla natura dell'analisi da svolgere, e si può calcolare a partire dal peso base relativo agli individui rimasti nel campione nei diversi anni del panel.

Volendo fare un'analisi a livello familiare, ad esempio, questo è ovviamente possibile: per questo motivo è riportato nel file il coefficiente di riporto familiare cross-section (DB090). Si consiglia in realtà di utilizzare il "peso condiviso" (DX090), calcolato come media dei pesi base di tutti gli appartenenti alla famiglia (*weight share method*). In questa maniera il peso condiviso familiare della generica famiglia  $h$  è pari a:

$$\omega_h = \frac{\sum_{i=1}^{n_h} RB060_i^h}{n_h}$$

dove  $n_h$  è la numerosità della famiglia (inclusendo i nuovi entrati e i nuovi nati) e  $RB060_i^h$  sono i singoli pesi base individuali degli individui appartenenti alla famiglia (inclusendo i nuovi entrati e i nuovi nati). Il peso condiviso è inserito nel file *DX090.csv* che contiene anche le chiavi per l'aggancio con gli altri file.

Nella tabella 1 si riportano i totali dei vari coefficienti di riporto all'universo presenti nei file rispetto ai diversi campioni longitudinali.

**Tabella 1 Popolazioni longitudinali e coefficienti di riporto all'universo**

Campioni longitudinali		ANNI			
		2005	2006	2007	2008
(DB075=1)	C5	w1	w2	w3	w4
(DB 075=2)	C6		w1	w2	w3
(DB 075=3)	C7			w1	w2

RB060 = Popolazione longitudinale del campione C5

RB060 = Popolazione longitudinale del campione C6

RB060 = Popolazione longitudinale del campione C7

**RB063 = Popolazione longitudinale nel panel a tre wave (RB060/2)**

**RB062 = Popolazione longitudinale nel panel a due wave (RB060/3)**

#### Popolazioni longitudinali

RB060

		2005	2006	2007	2008
(DB075=1)	C5	58,389,937	58,311,695	58,239,388	58,168,336
(DB075=2)	C6		58,892,655	58,819,629	58,747,869
(DB075=3)	C7			59,311,150	59,238,790
Total		58,389,937	117,204,350	176,370,167	176,154,995

#### Popolazioni longitudinali

PB050

		2005	2006	2007	2008
(DB075=1)	C5	49,285,838	49,350,598	49,533,033	49,794,189
(DB075=2)	C6		49,549,003	49,813,247	50,051,451
(DB075=3)	C7			49,897,219	50,147,267
Total		49,285,838	98,899,601	149,243,500	149,992,907

#### Popolazioni longitudinali

RB062

		2008
(DB075=1)	C5	19,389,445
(DB075=2)	C6	19,582,623
(DB075=3)	C7	19,746,263
Total		58,718,332

#### Popolazioni longitudinali s

RB063

		2008
(DB075=1)	C5	29,084,168
(DB075=2)	C6	29,373,934
(DB075=3)	C7	
Total		58,458,103

Maggiori informazioni su EU-SILC:

[http://forum.europa.eu.int/Public/irc/dsis/eusilc/library\\_](http://forum.europa.eu.int/Public/irc/dsis/eusilc/library_).

Quality report comparativi e nazionali rispettivamente disponibili su:

[http://circa.europa.eu/Public/irc/dsis/eusilc/library?l=/quality\\_assessment/comparative\\_quality\\_1&vm=detailed&sb=Title](http://circa.europa.eu/Public/irc/dsis/eusilc/library?l=/quality_assessment/comparative_quality_1&vm=detailed&sb=Title)

[http://circa.europa.eu/Public/irc/dsis/eusilc/library?l=/quality\\_assessment/quality\\_reports&vm=detailed&sb=Title](http://circa.europa.eu/Public/irc/dsis/eusilc/library?l=/quality_assessment/quality_reports&vm=detailed&sb=Title)

Altra documentazione:

[http://circa.europa.eu/Public/irc/dsis/eusilc/library?l=/data\\_dissemination/udb\\_user\\_database&vm=detailed&sb=Title](http://circa.europa.eu/Public/irc/dsis/eusilc/library?l=/data_dissemination/udb_user_database&vm=detailed&sb=Title)

Si può anche consultare il website su “Living conditions and social protection”:

[http://epp.eurostat.ec.europa.eu/portal/page/portal/living\\_conditions\\_and\\_social\\_protection/introduction](http://epp.eurostat.ec.europa.eu/portal/page/portal/living_conditions_and_social_protection/introduction)

Quello EUROSTAT:

<http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home>

Quello ISTAT:

<http://www.istat.it>

---

<sup>i</sup> In realtà, il campione longitudinale è anche al netto di un terzo gruppo di individui: quanti non partecipano all'indagine ma per i quali non si dispone di informazioni per decidere se facciano ancora parte della popolazione di riferimento o ne siano invece usciti. Ogni persona deve quindi essere assegnata all'uno o all'altro gruppo di persone uscite dal campione sulla base di un modello appropriato. Nell'indagine EUSILC italiana si è utilizzato un modello di regressione logistica che stimasse la propensione a rimanere nella popolazione come funzione di un set di variabili esplicative (quella che ovviamente esercita la maggiore influenza è l'età) sul gruppo di individui campione sui quali si dispone delle informazioni necessarie per poi attribuire i parametri del modello agli individui per i quali non si dispone di informazioni determinando così a quale dei due gruppi ricondurli.

<sup>ii</sup> Si noti che non è possibile ricavare queste informazioni da fonti esterne: dal bilancio demografico dell'ISTAT, infatti, sono noti gli ammontari complessivi di morti ed emigrati (tralasciando quanti si trasferiscono in convivenza, numericamente minori), ma non si può stabilire se questi facessero parte della popolazione di partenza; per fare un esempio pratico, se un individuo dovesse morire nel periodo in questione non potremmo sapere se era una persona appena entrata (per immigrazione, nascita o trasferimento da un'istituzione) o se faceva parte della popolazione di riferimento. L'utilizzo di dati demografici comporterebbe comunque l'utilizzo di una stima.

<sup>iii</sup> In realtà ai componenti della famiglia che non facevano parte della popolazione di riferimento (ad es., i nuovi immigrati) si dovrebbe assegnare un peso pari alla media degli individui campione appartenenti alla famiglia. Non disponendo di questa informazione, sono stati trattati come se facessero già parte della popolazione di riferimento, assegnando loro peso pari a 0.