

Nota informativa sull'utilizzo dei file relativi alla componente longitudinale di EU-SILC (2009-2010-2011-2012)

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.

Dal 2011 l'indagine si avvale della tecnica di rilevazione CAPI (Computer Assisted Personal Interviewing), e di conseguenza il questionario ha subito alcune modifiche e revisioni. Tuttavia, il file standard mantiene la stessa struttura e gli stessi contenuti degli anni precedenti.

Il materiale fornito include:

- File dati:
 - L11D.csv
 - L11H.csv
 - L11P.csv
 - L11R.csv
 - DX090.csv.
- Documentazione:
 - Guidelines 2012: describe l'indagine, le variabili e le raccomandazioni operative date da Eurostat agli Istituti di Statistica.
 - Differences between data collected and UDB: describe le variabili aggiuntive Eurostat, rispetto a quelle descritte nelle Guidelines.
 - UDB descrizione variabili italiane aggiuntive: describe le ulteriori variabili aggiuntive Italiane.
 - Modelli di rilevazione dell'indagine 2009, 2010, 2011 e 2012:
 - Guida_Intervistatore_2009/10
 - Guida_Intervistatore_CAPI_2011/12
 - SILC_REG_2009/10 registro familiare
 - SILC_FAM_2009/10 questionario familiare
 - SILC_IND_2009/10 questionario individuale
 - SILC_CAPI_2011/12 questionario CAPI

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*”.

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) 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 2010 del campione partito nel 2009 è:

$$P_{2010}^{(L)} = P_{2009} - OUT_{2010}$$

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 2009) e sia $RB060_{2009}$ 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.

$$(C4_{2009}, RB060_{2009}) \rightarrow P_{2009}$$

Al tempo $t+1$ (2010) il campione longitudinale è composto dal campione iniziale al netto degli usciti dalla popolazione di riferimento (out_{2010}) e di quanti non partecipano all'indagine pur avendone ancora i requisiti (x_{2010})¹:

$$C4_{2010}^{(L)} = C4_{2009} - (out_{2010} - x_{2010})$$

Assumendo che i fuoriusciti dal campione pesati siano rappresentativi dei fuoriusciti dalla popolazione abbiamo²:

$$((C4_{2010}^{(L)} + x_{2010}), RB060_{2009}) \rightarrow (P_{2009} - OUT_{2009})$$

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

¹ 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.

² 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.

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 ponderazione³.

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.

³ 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.

Tabella 1 Popolazioni longitudinali e coefficienti di riporto all'universo

Campioni longitudinali		ANNI			
		2009	2010	2011	2012
(DB075=1)	C9	w1	w2	w3	w4
(DB075=2)	C10		w1	w2	w3
(DB075=3)	C11			w1	w2

RB060 = Popolazione longitudinale del campione C9

RB060 = Popolazione longitudinale del campione C10

RB060 = Popolazione longitudinale del campione C11

RB062 = Popolazione longitudinale nel panel a due wave

RB063 = Popolazione longitudinale nel panel a tre wave

RB064 = Popolazione longitudinale nel panel a quattro wave

Popolazioni longitudinali

RB060

		2009	2010	2011	2012
(DB075=1)	C9	60,108,862	60,005,475	59,901,665	59,774,075
(DB075=2)	C10		60,191,094	60,086,963	59,958,978
(DB075=3)	C11			60,683,909	60,554,652
Total		60,108,862	120,196,569	180,672,538	180,287,705

Popolazioni longitudinali

PB050

		2009	2010	2011	2012
(DB075=1)	C9	50,830,661	51,435,230	51,921,038	52,022,056
(DB075=2)	C10		51,009,455	51,406,303	51,424,550
(DB075=3)	C11			51,413,475	51,958,455
Total		50,830,661	102,444,684	154,740,817	155,405,061

Popolazioni longitudinali

RB062

		2012
(DB075=1)	C9	19,852,934
(DB075=2)	C10	19,889,654
(DB075=3)	C11	20,110,846
Total		59,853,434

Popolazioni longitudinali

RB063

		2012
(DB075=1)	C9	29,779,402
(DB075=2)	C10	29,834,480
(DB075=3)	C11	
Total		59,613,882

Popolazioni longitudinali

RB064

		2012
(DB075=1)	C9	59,558,803
(DB075=2)	C10	
(DB075=3)	C11	
Total		59,558,803

Maggiori informazioni su EU-SILC sono disponibili sul sito

http://epp.eurostat.ec.europa.eu/portal/page/portal/income_social_inclusion_living_conditions/introduction

Quality report comparativi e nazionali rispettivamente disponibili su:

http://epp.eurostat.ec.europa.eu/portal/page/portal/income_social_inclusion_living_conditions/quality/eu_quality_reports

http://epp.eurostat.ec.europa.eu/portal/page/portal/income_social_inclusion_living_conditions/quality/national_quality_reports

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

http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Living_conditions_introduced

Il sito EUROSTAT:

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

Il sito ISTAT:

<http://www.istat.it>