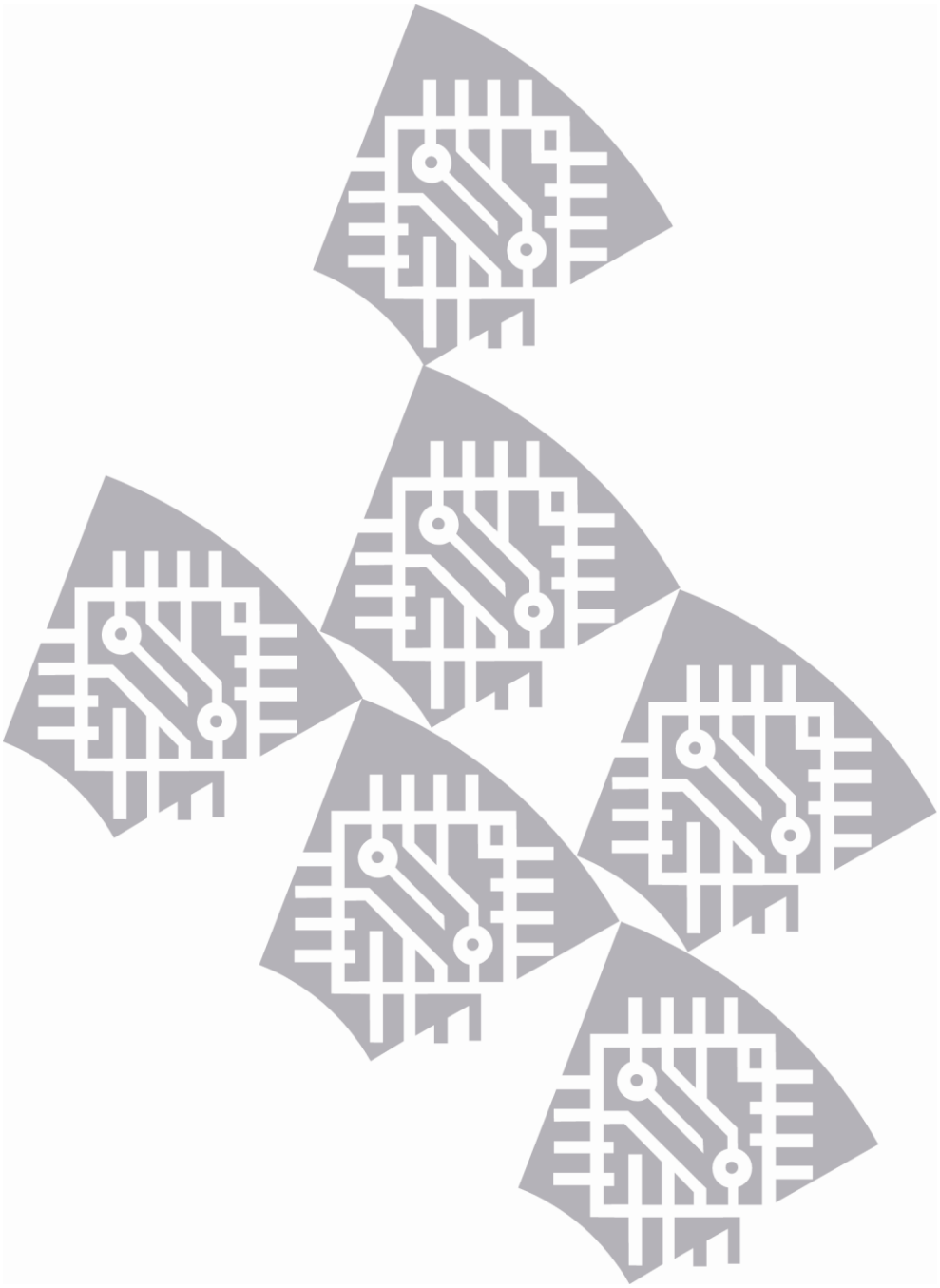
Istat
Istituto Nazionale di Statistica



Rilevazione sulle Forze di Lavoro

Dati longitudinali

Aspetti metodologici dell'indagine

INDICE

1. Introduzione 3
2. Il disegno campionario e la componente longitudinale 4
3. L'abbinamento dei record individuali 5
4. La tecnica di Indagine 6
5. Il trattamento dei dati 6
6. La popolazione di riferimento 7
7. La metodologia di calcolo dei pesi campionari 8
8. Le matrici di transizione 9
9. La diffusione dei dati longitudinali 12
10. Riferimenti bibliografici 13
11. Contatti 13

1. Introduzione

La Rilevazione sulle Forze di Lavoro (d'ora in avanti RFL) costituisce dal 1959 la principale fonte informativa sul mercato del lavoro; essa fornisce, con periodicità trimestrale, da un lato la stima del numero degli occupati e dei non occupati (persone in cerca di occupazione e non forze di lavoro) e dall'altro le variazioni tendenziali e congiunturali dell'occupazione e della disoccupazione. Tali stime si riferiscono a tutti i componenti delle famiglie residenti in Italia al netto dei membri permanenti delle convivenze (ospizi, brefotrofi, istituti religiosi, caserme, ecc.) .

La Rilevazione incorpora una componente longitudinale derivante dal sistema di rotazione delle famiglie nel campione. In particolare, la metà delle famiglie intervistate in un trimestre viene re-intervistata a distanza di 3 e 12 mesi, un quarto a distanza di 15 mesi. Archivi di microdati longitudinali possono essere ottenuti abbinando le informazioni raccolte sugli stessi individui / famiglie in diversi trimestri. Le matrici di transizione, desunte dai file in questione, forniscono una stima del numero di permanenze e di transizioni in entrata e in uscita dalle diverse condizioni occupazionali, e consentono di comprendere con un maggiore dettaglio le dinamiche del mercato del lavoro e le caratteristiche degli individui coinvolti.

I file di microdati longitudinali e le relative matrici di transizione costituiscono un "sottoprodotto" della rilevazione stessa, ed è bene sottolineare che non si tratta di un vero e proprio panel relativo a tutta la popolazione. Difatti, un individuo, intervistato la prima volta in uno dei comuni campione, non viene re-intervistato se nell'arco di tempo tra la prima e la successiva intervista ha cambiato residenza o si è trasferito all'estero. Ne consegue che, in un definito arco temporale, la componente longitudinale non rappresenta tutta la popolazione, ma solo quella residente in uno stesso comune sia all'inizio sia alla fine del periodo considerato, ossia la popolazione longitudinale.

Questa nota riporta gli aspetti più importanti relativi alla costruzione dei file di microdati longitudinali e delle matrici di transizione[1].

[1] Per una trattazione più dettagliata delle scelte metodologiche relative alla realizzazione dei dati longitudinali della RFL si veda Boschetto et al. (2010).

2. Il disegno campionario e la componente longitudinale

A partire dal primo trimestre 2004, la RFL ha subito una profonda trasformazione[2], ma il disegno campionario è rimasto quasi completamente invariato. Si tratta, difatti, di campionamento a due stadi con stratificazione delle unità di primo stadio e rotazione delle unità di secondo stadio; le unità di primo stadio sono i comuni e le unità di secondo stadio sono le famiglie.

Il disegno campionario prevede la sostituzione di una parte delle famiglie campione nelle varie occasioni di indagine. I campioni relativi a trimestri differenti, quindi, risultano parzialmente sovrapposti in base a uno schema di rotazione di tipo 2-2-2 (Figura 1), secondo cui una famiglia viene inclusa nel campione per due rilevazioni successive e, dopo una pausa di due trimestri, viene reinserita nel campione per altre due rilevazioni. A differenza delle famiglie, che ruotano secondo lo schema sotto riportato, i comuni campione rimangono sempre gli stessi nel tempo[3].

Figura 1. Schema di rotazione delle famiglie campione

ANNO		2011				2012				2013				2014			
TRIMESTRE		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
GRUPPOROT	A	A1	A2			A3	A4										
GRUPPOROT	B		B1	B2			B3	B4									
GRUPPOROT	C			C1	C2			C3	C4								
GRUPPOROT	D				D1	D2			D3	D4							
GRUPPOROT	E					E1	E2			E3	E4						
GRUPPOROT	F						F1	F2			F3	F4					
GRUPPOROT	G							G1	G2			G3	G4				
GRUPPOROT	H								H1	H2			H3	H4			

La Figura 1 riporta la ripartizione dei gruppi di rotazione rispetto alle occasioni di indagine. Il campione di famiglie intervistate nel terzo trimestre 2012 è composto da un gruppo di famiglie che entrano per la prima volta nel campione (G1), un gruppo di famiglie intervistate per la seconda volta (F2), un gruppo di famiglie intervistate per la terza volta (C3) ed un gruppo di famiglie intervistate per la quarta ed ultima volta (B4).

[2] Per approfondimenti circa la rilevazione trimestrale che ne costituisce il presupposto, si veda ISTAT (2006).

[3] In realtà alcuni comuni di ridotta dimensione in termini demografici vengono periodicamente sostituiti con altri con le stesse caratteristiche a causa dell'esaurimento delle liste anagrafiche

In particolare, il campione teorico di famiglie relative a ciascuna occasione di indagine (rilevazione trimestrale) è costituito da circa 75mila famiglie[4], suddivise equamente nei quattro gruppi di rotazione costituiti, quindi, ciascuno da circa 19mila famiglie.

La struttura longitudinale così congegnata comporta una sovrapposizione del campione teorico del 50% a tre e a dodici mesi di distanza, del 25% a nove e quindici mesi. Tale sovrapposizione consente di costruire file di microdati longitudinali a 3, a 12 e a 15 mesi di distanza. Ad esempio, il file longitudinale a 12 mesi di distanza, che ha come riferimento il terzo trimestre 2012 e il terzo trimestre 2013 contiene le informazioni delle famiglie presenti contemporaneamente nei gruppi G1 – G3, e di quelle nei gruppi F2 – F4, mentre il file longitudinale a 3 mesi di distanza, che ha come riferimento il terzo trimestre 2012 e il quarto trimestre 2012 fa riferimento alle famiglie presenti nei gruppi F1 – F3, e B3 – B4. Ne consegue che, inevitabilmente, il livello di precisione delle stime longitudinali risulta ridotto rispetto alla rilevazione trimestrale, e quindi eventuali risultati ottenuti per sottogruppi di popolazione e/o domini territoriali molto piccoli potrebbero avere un elevato grado di incertezza.

3. L'abbinamento dei record individuali

La RFL prevede l'uso di codici identificativi familiari e individuali univoci per tutto il periodo in cui la famiglia e i suoi componenti partecipano all'indagine (15 mesi). L'uso della tecnica di indagine CAPI e CATI garantisce che, nella quasi totalità dei casi, questi codici vengono registrati in maniera corretta, permettendo un abbinamento deterministico.

In realtà, mentre il codice identificativo della famiglia risulta sempre corretto, un limitato numero di errori può essere presente sui codici identificativi dei componenti della famiglia. Tra i problemi che possono generare questi errori si possono menzionare i seguenti:

- errori nel funzionamento dei software di acquisizione dei dati;
- errori compiuti dai rilevatori durante la somministrazione del questionario elettronico;
- errori di compilazione e registrazione del questionario cartaceo quando quello elettronico non può essere usato.

Per risolvere completamente il problema dei pochi casi di errati abbinamenti, è stata implementata una procedura di verifica deterministica che, usando variabili di controllo

[4] Si tratta di un campione teorico di circa 300mila famiglie in un anno.

quali sesso, data di nascita e nome, permette l'individuazione sia dei falsi positivi (componenti diversi che hanno lo stesso codice), sia dei falsi negativi (lo stesso componente che in due diverse occasioni ha due diversi codici individuali). Il numero degli errati abbinamenti individuati (corretti con procedure ad hoc) risulta comunque di entità marginale.

4. La tecnica di Indagine

L'uso di una tecnica di indagine assistita da computer, permette di ottenere un dato "grezzo" trasversale qualitativamente elevato rispetto alle indagini eseguite con tecniche non assistite da computer. Grazie all'implementazione diretta sul questionario elettronico di un piano di controllo che agisce al momento dell'intervista (controllo dei domini delle variabili, dei percorsi del questionario e delle incongruenze logiche) il numero di errori dal punto di vista trasversale risulta estremamente ridotto.

In linea generale, la prima intervista (prima wave) di ogni famiglia presente nel campione è condotta mediante tecnica CAPI (intervista faccia a faccia assistita da computer) mentre le tre successive sono da effettuarsi mediante tecnica CATI (intervista telefonica assistita da computer). Alcune eccezioni a queste regole riguardano:

- le famiglie con capofamiglia straniero, che vengono sempre intervistate con tecnica CAPI;
- le famiglie che non dispongono o non vogliono fornire un numero di telefono che devono essere re-intervistate con tecnica CAPI;
- le interviste in prima wave, per le famiglie che dispongono di telefono, nei periodi delle vacanze di Ferragosto e Natale - Capodanno, vengono effettuate con tecnica Cati.

Le tecniche CAPI e CATI consentono di usare, inoltre, per le re-interviste, e solo per alcuni specifici quesiti, la somministrazione di domande con modalità a conferma. Questo metodo consente di ridurre i tempi di somministrazione, la molestia statistica e garantisce una maggiore coerenza delle informazioni rilevate nei diversi trimestri sullo stesso individuo campione.

5. Il trattamento dei dati

Nonostante le regole di controllo implementate nei questionari elettronici, un certo numero di incoerenze longitudinali, relativo alle informazioni rilevate in diversi trimestri, risulta comunque presente nei dati "grezzi". Tali incoerenze possono derivare da una

molteplicità di cause, tra le quali per esempio: gli errori di risposta, che possono intervenire per incomprensione della domanda, oppure per errata collocazione nel tempo di un evento, o se il rispondente proxy non è a conoscenza della reale situazione del familiare per conto del quale fornisce la risposta; errori di registrazione, che si riscontrano per incomprensione della risposta o per errore dell'intervistatore.

Nei file longitudinali rilasciati dall'Istituto, le incoerenze longitudinali vengono comunque risolte mediante un piano di controllo e correzione longitudinale, imputando le informazioni che risultano incoerenti o mancanti. L'aver rilevato una molteplicità di informazioni su più istanti di tempo, consentirà di ricostruire le storie lavorative degli individui campione in maniera più precisa e dettagliata.

La correzione dei dati longitudinali della RFL ha lo scopo principale di garantire la coerenza longitudinale delle informazioni e di preservarne, in secondo luogo, quella trasversale. Le regole generali di imputazione, da usare per la correzione a regime dei dati longitudinali, sono state definite studiando l'intera storia lavorativa, dalla prima occasione di indagine alla quarta occasione (o wave), tenendo in considerazione lo stato di evoluzione delle varie occasioni di intervista alle quali sono arrivate le famiglie presenti nel campione (ad esempio la strategia di correzione su individui alla seconda intervista è diversa rispetto a quelli giunti alla terza ovvero alla quarta intervista).

6. La popolazione di riferimento

Come in ogni altro tipo di indagine, anche per la componente longitudinale della RFL è fondamentale definire la popolazione di riferimento, cioè la popolazione che può essere correttamente rappresentata dal campione longitudinale degli individui abbinati.

Uno dei punti fondamentali da tenere in considerazione è che la popolazione si modifica nell'arco di un determinato periodo a causa di entrate (nascite e iscrizioni anagrafiche dovute a cambi di residenza o immigrazione) e uscite (morti e cancellazioni anagrafiche dovute a cambi di residenza o emigrazione). Inoltre, la componente longitudinale, che è un sottoprodotto della RFL, non può essere considerata come un vero e proprio panel; non fornisce, infatti, informazioni sulla condizione occupazionale, a inizio e fine periodo, di tutta la popolazione di partenza, ma solo relativamente ad una sua parte, seppur considerevole. Questo limite è dovuto al fatto che il disegno campionario della RFL non prevede di seguire sul territorio, per le interviste successive, né gli individui che escono dalla famiglia campione, né le famiglie intere che si trasferiscono verso altri comuni o verso l'estero.

Di conseguenza, in un determinato arco temporale, il campione longitudinale della RFL, che scaturisce dall'abbinamento di due trimestri, non può rappresentare tutta la popolazione, ma è in grado di rappresentare correttamente solo quella residente in uno stesso comune sia all'inizio sia alla fine del periodo considerato[5].

Tale popolazione, che per comodità espositiva chiameremo popolazione longitudinale, è definita come la popolazione residente a inizio periodo (esclusi gli individui che fanno parte di convivenze), al netto delle morti e dei cambi di residenza verso altri comuni e/o verso l'estero verificatisi nel periodo. Grazie alla disponibilità all'interno dell'Istituto di dati demografici - affidabili e tempestivi circa la popolazione residente, le iscrizioni e le cancellazioni - prodotti sulla base delle comunicazioni provenienti dai registri anagrafici, la popolazione longitudinale è stata calcolata a livello di singola provincia con disaggregazione per sesso e classe di età, e a livello di regione con disaggregazione per sesso e cittadinanza (italiani/stranieri). Il metodo di costruzione della popolazione longitudinale garantisce la coerenza tra le popolazioni di riferimento delle stime trimestrali e la popolazione di riferimento delle stime longitudinali.

7. La metodologia di calcolo dei pesi campionari

Lo stimatore scelto per il riporto all'universo dei dati longitudinali della RFL, come per i dati trasversali, appartiene alla classe degli stimatori di ponderazione vincolata[6] . Si tratta di uno stimatore in cui i pesi finali consentono di ottenere, nell'ambito di diversi domini territoriali (ripartizioni, regioni, province), stime di popolazione (per sesso, classi di età e/o cittadinanza) uguali ai corrispondenti totali noti di fonte anagrafica. I pesi sono calcolati a livello familiare e ciò implica che a ciascun individuo appartenente alla stessa famiglia sia attribuito un medesimo peso di riporto all'universo; l'unicità del peso a livello familiare assicura la coerenza delle stime familiari con le stime individuali.

Il peso di riporto all'universo per un campione longitudinale viene determinato mediante diverse fasi, distinte e consecutive, che permettono da un lato il trattamento delle mancate risposte totali, dall'altro la ponderazione delle stime longitudinali sia su totali noti di popolazione di fonte anagrafica, sia sulle stime trasversali della stessa RFL riferite ai trimestri iniziale e finale.

[5] La componente longitudinale non può correttamente rappresentare tutta la popolazione che risiede sul territorio nazionale in quanto il comportamento degli individui che cambiano residenza differisce fortemente da quello degli individui che non la cambiano. L'estensione dei risultati relativi alla popolazione longitudinale alla popolazione complessiva comporterebbe l'introduzione di una distorsione nelle stime di flusso.

[6] Si tratta di uno stimatore usato in ISTAT per tutte le indagini sulle famiglie. Nella letteratura in lingua inglese tale stimatore è noto con il nome di calibration estimator; per ulteriori informazioni si veda Deville & Särndal (1992).

La procedura di riporto all'universo dei file di dati longitudinali a 12 mesi di distanza, prevede le seguenti fasi:

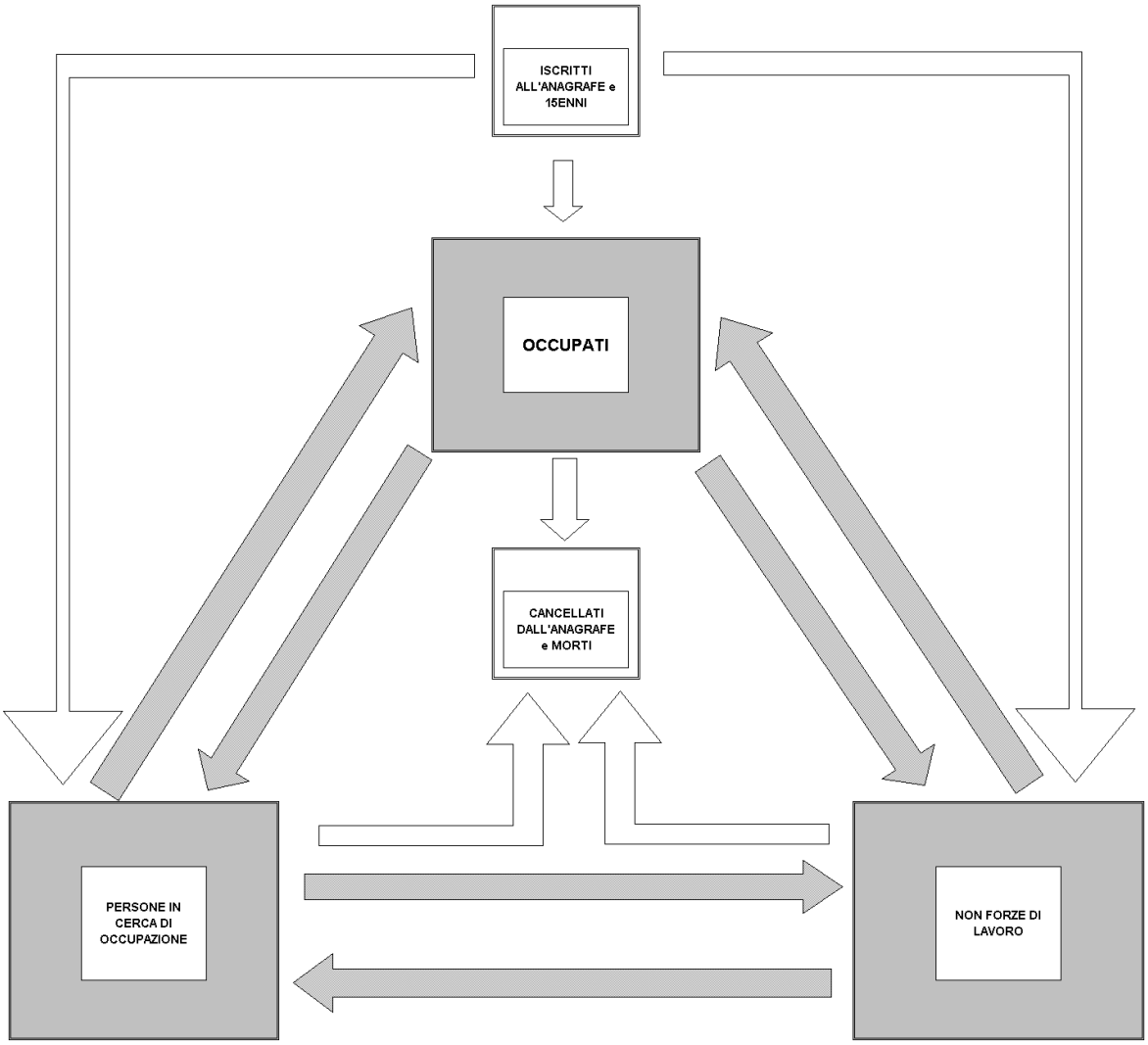
- nel trimestre iniziale (rispetto al periodo di riferimento dei dati longitudinali) tutti gli individui che dovrebbero essere re-intervistati dopo 12 mesi (sottocampione degli abbinabili) vengono selezionati e il loro peso finale trasversale viene utilizzato come peso base longitudinale.
- il peso base longitudinale viene corretto mediante l'uso di una procedura di ponderazione vincolata, in modo che il sottocampione degli abbinabili riproduca esattamente un certo numero di stime trasversali ottenute con il campione trasversale completo (popolazione per sesso, classe di età, territorio e cittadinanza, classificata per condizione professionale e altre sue caratteristiche rilevanti). Al termine di questa fase si ottiene il peso intermedio longitudinale per tutti gli individui abbinabili.
- nella terza fase, vengono selezionati solo gli individui effettivamente abbinati (che formano quindi il dataset longitudinale) e, partendo dai loro pesi intermedi longitudinali, si usa ancora la procedura di ponderazione vincolata per ottenere i pesi finali longitudinali. In questa fase i vincoli della procedura sono costituiti dalla popolazione longitudinale, relativa al periodo in questione, con una suddivisione per cittadinanza, sesso e classi di età, diversa per i diversi domini territoriali.

8. Le matrici di transizione

La costruzione degli archivi longitudinali della RFL, e la conseguente possibilità di fornire matrici di transizione per la stima dei flussi tra condizioni nel mercato del lavoro, è subordinata, e soprattutto limitata, dalla particolare natura del disegno dell'indagine che ha come obiettivo fondamentale quello di fornire stime trimestrali trasversali dei principali indicatori strutturali del mercato del lavoro in ottemperanza alle definizioni internazionali di occupato e di persona in cerca di occupazione (Regolamenti Comunitari n° 1575/2000 e n° 1897/2000).

La sottostante Figura 2 rappresenta i flussi della popolazione complessiva per un dato intervallo di tempo (un trimestre o un anno), che possono essere rappresentati dai campioni trasversali e longitudinali della RFL.

Figura 2. Diagramma dei flussi della popolazione complessiva

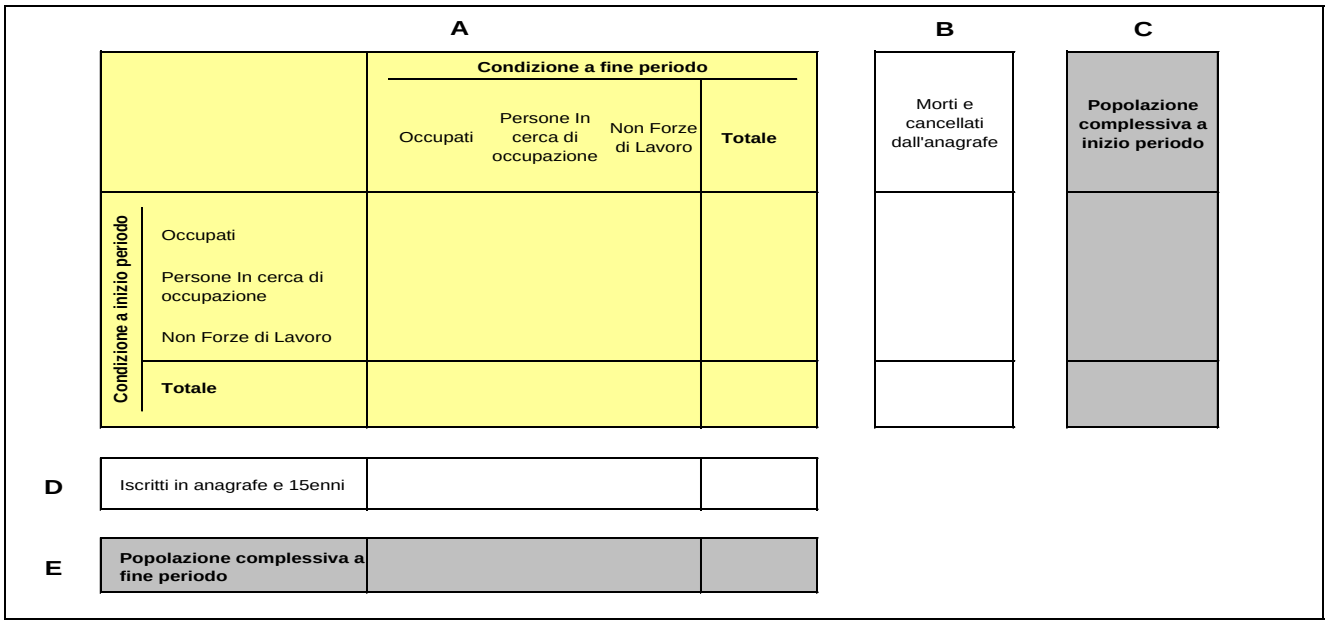


I campioni trasversali della RFL forniscono una stima della distribuzione per condizione professionale, sia della popolazione iniziale, sia di quella finale. Come detto, però, parte della popolazione iniziale può cambiare residenza, emigrare o morire. Ne consegue che di essa se ne conosce solo la condizione a inizio periodo. Allo stesso modo, di quella parte della popolazione che si è iscritta in anagrafe o ha compiuto 15 anni[7] nel periodo sotto osservazione, è nota solo la condizione a fine periodo.

[7] Considerata per convenzione internazionale l'età di ingresso nel mercato del lavoro.

La Figura 3 mostra la matrice completa degli stock e dei flussi della popolazione complessiva che fornisce quindi, con diverso grado di dettaglio, una serie di informazioni riferite agli aggregati di popolazione: iniziale, finale, longitudinale, uscita e entrata. La componente longitudinale, quindi, descrive solo i flussi tra le diverse condizioni (le frecce colorate nel diagramma della Figura 2) intervenuti per la popolazione longitudinale, sintetizzati nella matrice di transizione (indicata con A nella Figura 3).

Figura 3. Schema della matrice completa degli stock e dei flussi della popolazione complessiva



Considerando solo la popolazione in età lavorativa in entrambe le occasioni (con almeno 15 anni a inizio periodo), la matrice completa contiene:

- la matrice di transizione (indicata con la lettera A) con la distribuzione congiunta secondo la condizione a inizio e fine periodo, desunta dalla componente longitudinale della RFL e riferita esclusivamente alla popolazione longitudinale;
- due vettori con la distribuzione per condizione della popolazione complessiva, sia a inizio che a fine periodo, risultante dai relativi campioni trasversali (indicati con C e E);
- due vettori di raccordo tra stime trasversali e stime longitudinali (ottenuti per differenza) che riportano la condizione a inizio periodo per coloro che risultano morti o cancellati dalle anagrafi nel periodo (indicato con B) e la condizione a fine periodo per coloro che compiono 15 anni e si iscrivono in anagrafe nello stesso periodo (indicato con D).

Tenendo conto dello schema sopra riportato, le stime contenute nelle matrici di transizione, prodotte a partire dai dati longitudinali mediante l'uso del peso finale longitudinale descritto in precedenza, risultano coerenti con le stime trasversali ufficiali e relative alla popolazione complessiva della RFL[8].

Le matrici di transizione consentono, quindi, di comprendere con un maggiore dettaglio l'evoluzione degli stock di occupati, disoccupati e inattivi, arricchendo il panorama delle informazioni statistiche sulle dinamiche del mercato del lavoro. Difatti, esse forniscono una stima del numero di permanenze e di transizioni in entrata e in uscita dalle diverse condizioni occupazionali. Consentono, inoltre, di stimare le probabilità di permanenza e/o di passaggio da una condizione di origine a una di arrivo, e di analizzare le caratteristiche degli individui che ne sono coinvolti.

9. La diffusione dei dati longitudinali

Sono previste due versioni di file di microdati: SISTAN (microdati per enti appartenenti al SISTAT) e MFR (microdati per la ricerca). Alcuni dei quesiti presenti nei file microdati trasversali non sono presenti nei microdati longitudinali o per motivi di affidabilità statistica dell'informazione (il file longitudinale si riferisce a meno della metà del campione trasversale) o per motivi di coerenza longitudinale.

In particolare e in merito a questo aspetto occorre precisare che nel file è già presente l'informazione relativa alla durata della ricerca di occupazione, ma essa non può essere usata per la stima della durata della disoccupazione in quanto quest'ultima viene attribuita sulla base del valore tra i mesi di ricerca di lavoro e mesi trascorsi dalla fine dell'occupazione (per coloro che ne avevano una).

[8] Le matrici di transizione prodotte fanno riferimento soltanto alla popolazione longitudinale (che è una parte di quella iniziale e finale), quindi, tutte le analisi possono essere condotte solo su matrici "al netto" dei flussi realizzatisi per la popolazione iscritta e cancellata nel periodo. D'altro canto questo tipo di matrici consentono di raggiungere un elevato dettaglio informativo, con una minima distorsione, soltanto sulla popolazione longitudinale.

10. Riferimenti bibliografici

Boschetto B., Discenza A.R., Fiori F., Lucarelli C., Rosati S., (2010), "Longitudinal data for the analysis of Italian labour market flows", Statistica Applicata, Volume 1/2010, Padova, pp. 129-150.

ISTAT, (2006), "La Rilevazione sulle Forze di Lavoro: contenuti, metodologie, organizzazione.", Collana Metodi e Norme, n. 32, Roma

Deville J. C., Särndal C. E., (1992), Calibration Estimator, in Survey Sampling, Journal of the American Statistical Association, vol. 87, pp.376-382.

11. Contatti

ISTAT - Servizio Istruzione, Formazione e Lavoro – e.mail: infolav@istat.it

Carlo Lucarelli – 0646734565 - e.mail: carlo.lucarelli@istat.it

Barbara Boschetto – 0646734555 – e.mail: barbara.boschetto@istat.it

Filomena De Filippo – 0646734558 – e.mail: defilipp@istat.it

Antonella Iorio - 0646734482 – e.mail: antonella.iorio@istat.it

Marco Iudicone - 0646734772 – e.mail: marco.iudicone@istat.it

Cristiano Marini - 0646734714 – e.mail: cristiano.marini@istat.it