# Le matrici di transizione della Rilevazione trimestrale sulle forze di lavoro

## Nota metodologica

*Roma*
*12 dicembre 2002*

# INDICE

# 1 Aspetti definitori[1]

## 1.1 La Rilevazione trimestrale sulle forze di lavoro

L'Istituto nazionale di statistica realizza ogni trimestre un quadro della situazione del mercato del lavoro in Italia. La principale fonte informativa utilizzata è la Rilevazione trimestrale sulle forze di lavoro (RTFL, nel seguito). La RTFL è un'indagine campionaria che viene condotta continuativamente con cadenza trimestrale a partire dal 1959. Essa consente nell'arco delle quattro rilevazioni trimestrali di acquisire informazioni su oltre 300 mila famiglie per un totale di 800 mila individui, distribuiti in 1351 comuni italiani, l'1,4 per cento della popolazione complessiva nazionale.

Il suo utilizzo per analisi sia di tipo congiunturale sia strutturale, è rivolto allo studio dei principali indicatori del mercato del lavoro. L'evoluzione di tali indicatori può essere analizzata in modo disaggregato a livello territoriale, settoriale e per le principali caratteristiche socio-demografiche della popolazione.

Dalla sua introduzione ad oggi l'indagine è stata più volte ristrutturata per tenere conto, da un lato delle trasformazioni del mercato del lavoro italiano, dall'altro delle crescenti esigenze conoscitive da parte degli utenti sulla realtà sociale ed economica del nostro paese. Nel corso degli ultimi quindici anni un ruolo di primo piano nelle trasformazioni metodologiche dell'indagine è stato svolto dal processo di armonizzazione promosso dall'Eurostat al fine di rendere maggiormente comparabili le statistiche internazionali sul mercato del lavoro. L'ultima ristrutturazione è stata realizzata nell'ottobre 1992 e si è concretata con l'introduzione di un insieme di modifiche rilevanti: nuove definizioni per la popolazione in età lavorativa e per le persone in cerca di lavoro; nuovo modello di rilevazione; nuova classificazione degli occupati per settore di attività economica; nuova procedura di controllo e correzione degli errori; nuove stime della popolazione di riferimento e implementazione degli stimatori calibrati per la determinazione dei coefficienti di riporto all'universo (unico per individui e famiglie).

Nel luglio 1999 l'Istat ha effettuato una revisione delle stime della RTFL relative al periodo ottobre 1992 - aprile 1999. Tale operazione è stata motivata da una pluralità di ragioni, essenzialmente legate al rispetto dei vincoli posti dal nuovo regolamento comunitario in materia di procedure di calcolo dei coefficienti di riporto all'universo e all'adozione dei dati di popolazione prodotti secondo il metodo anagrafico. La revisione ha comportato un cambiamento, talvolta sensibile, delle stime dei principali aggregati del mercato del lavoro.

In tale ambito si inseriscono le matrici di transizione che derivano dal particolare disegno campionario della RTFL. La presente nota metodologica ha l'obiettivo di illustrare i passaggi cruciali che hanno portato alla realizzazione delle matrici di transizione.

A tale proposito è doveroso sottolineare la preziosa e fattiva collaborazione con i ricercatori dell'Università degli studi di Padova che in particolare hanno curato e realizzato la procedura per l'abbinamento dei *record* con discordanze.

[1] Il presente lavoro è stato curato e realizzato da: Claudio Ceccarelli, Antonio Rinaldo Discenza e Simona Rosati del Servizio Formazione e Lavoro dell'Istat e da Adriano Paggiaro e Nicola Torelli dell'Università degli studi di Padova. Tale nota metodologica è da considerarsi una prima stesura di un lavoro più ampio, di prossima pubblicazione, nel quale saranno riportate anche le performance delle metodologie adottate.

### 1.1.1 Universo di riferimento

L'universo di riferimento dell'indagine è costituito da tutti i componenti delle famiglie residenti in Italia, anche se temporaneamente emigrati all'estero. Sono escluse le famiglie residenti in Italia che vivono abitualmente all'estero e i membri permanenti delle convivenze (ospizi, brefotrofi, istituti religiosi, caserme, ecc.)

### 1.1.2 Unità di rilevazione

L'unità di rilevazione è la famiglia di fatto. Questa va intesa come un insieme di persone legate da vincoli di matrimonio, parentela, affinità, adozione, tutela o da vincoli affettivi, coabitanti ed aventi dimora abituale nello stesso comune (anche se non residenti secondo l'anagrafe nello stesso domicilio). Due sono, quindi, le condizioni affinché un insieme di persone diventi una famiglia: coabitazione e presenza di un legame. Ad esempio, un figlio che si sposa, se continua ad abitare con i genitori, costituisce con loro un'unica famiglia.

Una famiglia può essere costituita, naturalmente, anche da una sola persona. Qualora il rilevatore nell'effettuare l'intervista trovi due o più famiglie nella stessa abitazione deve intervistare solo la famiglia estratta e indicata dal comune.

### 1.1.3 Periodicità e riferimento temporale

L'indagine viene svolta trimestralmente a gennaio, aprile, luglio e ottobre di ogni anno al fine di cogliere la stagionalità dei fenomeni rilevati. Le notizie acquisite e di conseguenza i dati pubblicati non fanno riferimento ad una media trimestrale ma ad una situazione puntuale colta di volta in volta nella prima settimana dei mesi indicati. Alla fine dell'anno l'ISTAT pubblica una "Media" dei dati rilevati nelle quattro rilevazioni.
I riferimenti temporali delle notizie raccolte sono:

- giorno di riferimento, che coincide con il venerdì della settimana di riferimento. A questo giorno vanno ricondotte le informazioni sull'età, lo stato civile, il livello di istruzione e la cittadinanza;
- settimana di riferimento: è di norma la prima settimana priva di giorni festivi del mese in cui viene condotta l'inchiesta (dal lunedì alla domenica);
- settimana di rilevazione: è la settimana successiva a quella di riferimento durante la quale gli intervistatori si recano presso le famiglie per le interviste.

### 1.1.4 Il disegno di campionamento

#### *1.1.4.1 Il campione*

Il campione utilizzato è a due stadi con stratificazione delle unità di primo stadio. Le unità di primo stadio sono costituite dai comuni, quelle di secondo stadio dalle famiglie anagrafiche.

La stratificazione delle unità di primo stadio è basata sulla sola popolazione residente nei comuni.

Il disegno tiene conto della condizione di autoponderazione dello strato nell'ambito di ciascuna regione geografica, il che ha comportato l'assegnazione ad ogni provincia di un numero di famiglie campione proporzionale al peso demografico della provincia stessa.

Dall'aprile 1995, con l'istituzione delle nuove province, il numero dei Comuni campione è stato portato a 1.351 unità mentre le famiglie intervistate sono diventate 75.516.

La procedura di selezione dei comuni campione avviene formando strati omogenei per provincia, in modo da ottenere livelli costanti di popolazione complessiva, e determinando una soglia dimensionale per ogni provincia, al di sopra della quale tutti i comuni vengono inclusi nel piano di campionamento (comuni "auto-rappresentativi"), e al di sotto della quale vengono selezionati due comuni per ogni strato elementare (comuni "non auto-rappresentativi"), senza reimmissione e con probabilità proporzionale al peso demografico del comune stesso. I comuni auto-rappresentativi non possono essere sostituiti, mentre i comuni non auto-rappresentativi vengono sostituiti solo quando non sono più in grado di fornire nuove famiglie campione.

L'estrazione delle famiglie-campione dalle liste anagrafiche dei comuni viene effettuata con cadenza annuale, in coincidenza con l'indagine di aprile. Il numero di famiglie da estrarre viene stabilito in modo da assicurare la formazione dei campioni per l'intero ciclo, nonché di un elenco di famiglie di riserva per eventuali sostituzioni.

Il campione è caratterizzato da una struttura longitudinale del tipo 2-2-2 per cui ogni famiglia viene intervistata per due trimestri successivi, esce temporaneamente dal campione per due trimestri e infine rientra nel campione per gli ultimi due trimestri, prima di abbandonarlo definitivamente. Il sistema di rotazione consente di mantenere invariata metà della composizione del campione in due trimestri consecutivi e in trimestri a distanza di anno l'uno dall'altro. La scelta del sistema di rotazione è tale da conciliare in maniera ottimale le esigenze di costruzione di stime di "livello" e stime di "variazione": maggiore è il numero di famiglie che si rinnovano di periodo in periodo, maggiore è la validità degli aggregati stimati dall'indagine (stime di livello); viceversa, la presenza di una quota consistente di famiglie in comune da una rilevazione all'altra garantisce la stabilità delle stime in periodi successivi (stime di variazione).

#### *1.1.4.2 Le stime cross-section*

Per il calcolo dei coefficienti di riporto all'universo si utilizza una procedura generalizzata di stima in tutte le indagini campionarie condotte dall'Istat.

La procedura si basa sull'uso di una famiglia di stimatori, noti in letteratura come *calibration estimator* (stimatori di ponderazione vincolata). La caratteristica fondamentale della metodologia alla base di tali stimatori 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 sia per famiglia. Tale famiglia di stimatori coincide asintoticamente con lo stimatore di regressione generalizzato; pertanto, per campioni sufficientemente grandi, è possibile affermare che tali stimatori abbiano approssimativamente le stesse proprietà (corretti, consistenti e con la stessa varianza campionaria) (Deville e Särndal,1992).

La costruzione dei coefficienti finali di riporto all'universo trimestrali è articolata nelle seguenti fasi:

1. calcolo del coefficiente di riporto base (o peso diretto), ottenuto come reciproco della probabilità di inclusione di ogni famiglia campione;
2. calcolo del fattore di correzione per mancata risposta totale, ottenuto come l'inverso del tasso di risposta per ciascuno strato;
3. calcolo del fattore correttivo che consente di soddisfare la condizione di uguaglianza tra i totali noti della popolazione e le corrispondenti stime campionarie.

Le stime di media annua si ottengono dividendo per 4 i coefficienti finali di riporto all'universo trimestrali.

La procedura vincola le stime ai seguenti totali noti:

- popolazione residente provinciale per sesso;
- popolazione regionale residente per sesso e classi d'età (0-14; 15-19; 20-24; 25-29; 30-34; 35-39; 40-44; 45-49; 50-54; 55-59; 60-64; 65-69; 70-74; 75 e più);

I totali noti sopra indicati si ottengono combinando i dati relativi alle seguenti fonti:

1. ammontare mensile della popolazione regionale (disponibile con cadenza mensile);
2. struttura della popolazione per provincia e sesso (disponibile con cadenza annuale e riferita al 31/12 di ogni anno);
3. struttura della popolazione per regione, sesso ed età (disponibile con cadenza trimestrale);
4. Ammontare delle convivenze (dato censuario).

Per ottenere i totali noti le fonti sono combinate nel modo seguente. Indicando con:

$P_r^{(1)}$ popolazione residente nella regione r desunta dalla fonte n.1;

$P_r^{(2)}$ popolazione residente nella regione r desunta dalla fonte n.2;

$P_r^{(3)}$ popolazione residente nella regione r desunta dalla fonte n.3;

$P_{ps}^{(2)}$ popolazione di sesso s residente nella provincia p desunta dalla fonte n.2;

$P_{ps}^{(4)}$ popolazione che vive in convivenze di sesso s nella provincia p desunta dalla fonte n.4;

si ha:

$$P_{ps}^{*} = \frac{P_r^{(1)}}{P_r^{(2)}} P_{ps}^{(2)} - P_{ps}^{(4)} \qquad (1.1)$$

Ovvero, la popolazione di sesso s residente nella provincia p, al netto delle convivenze, è aggiornata, rispetto al dato di inizio anno (fonte n.2), con la popolazione regionale ultima disponibile (fonte n.1).

$$P_{rse}^{*} = \frac{P_r^{(1)}}{P_r^{(3)}} P_{rse}^{(3)} - P_{rse}^{(4)} \qquad (1.2)$$

Analogamente per la (1.2), la popolazione di sesso s ed età e residente nella regione r, al netto delle convivenze, è aggiornata, rispetto al dato trimestrale (fonte n.3) con la popolazione regionale ultima disponibile (fonte n.1).
La (1.1) e la (1.2) si basano sull'ipotesi che la composizione per sesso ed età della popolazione rimanga costante per periodi di tempo così ristretti.

Definita la metodologia utilizzata per il calcolo dei coefficienti di riporto all'universo risulta sufficientemente agevole definire le stime riferite ad un dato carattere Y della popolazione di riferimento. Pertanto, la stima del numero di occupati in uno specifico dominio *d* può essere rappresentata secondo la seguente formalizzazione.
Sia:

$$Y = \begin{cases} 1 & \text{occupato} \\ 0 & \text{non occupato} \end{cases};$$

$d$ indice di dominio territoriale di riferimento delle stime;
$H_d$ numero di strati nel dominio d;
$M_h$ numero di famiglie residenti nello strato h;
$m_h$ numero di famiglie campione nello strato h;
$P_h$ numero di individui residenti nello strato h;

$p_h$ numero di individui campione nello strato h;
$n_j$ numero di componenti della famiglia j;
$Y_h$ numero di occupati nello strato h;
$y_h$ numero di occupati nel campione dello strato h;
$Y_{hj}$ numero di occupati nella famiglia j dello strato h;
$y_{hj}$ numero di occupati nella famiglia campione j dello strato h.

Il numero di occupati nel dominio *d* è espresso da:

$$ {}_dY = \sum_{h=1}^{H_d}\sum_{i=1}^{P_h} Y_{hi} = \sum_{h=1}^{H_d}\sum_{j=1}^{M_h}\sum_{k=1}^{n_j} Y_{hjk} \qquad (9) $$

Una stima della (9) è data dalla seguente espressione:

$$ {}_d\hat{Y} = \sum_{h=1}^{H_d}\sum_{i=1}^{p_h} y_{hi} w_{hi} = \sum_{h=1}^{H_d}\sum_{j=1}^{m_h}\left(\sum_{k=1}^{n_j} y_{hjk}\right) w_{hj} \qquad (10) $$

in cui $w_{hi} = w_{hj}$, ovvero il *coefficiente finale di riporto all'universo* da attribuire alla famiglia *j* dello strato *h* uguale al attribuito ad ogni suo componente.

## 1.2 La struttura longitudinale della RTFL

Il disegno campionario dell'indagine prevede la sostituzione di una parte delle unità campionarie nelle varie occasioni di indagini. Un tale campione viene denominato *campione ruotato.* In particolare, il campione di famiglie relative a ciascuna occasione di indagine (rilevazione trimestrale) è costituito da quattro *gruppi di rotazione.* Ogni gruppo di rotazione è costituito da un quarto delle famiglie campione, circa 19.000, che corrisponde a circa 49.000 individui campione.

**Schema 1.1 Gruppi di rotazione**

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gennaio 2002 | A4 | B3 | | | E2 | F1 | | | | | | | |
| Aprile 2002 | | B4 | C3 | | | F2 | G1 | | | | | | |
| Luglio 2002 | | | C4 | D3 | | | G2 | H1 | | | | | |
| Ottobre 2002 | | | | D4 | E3 | | | H2 | I1 | | | | |
| Gennaio 2003 | | | | | E4 | F3 | | | I2 | L1 | | | |
| Aprile 2003 | | | | | | F4 | G3 | | | L2 | M1 | | |
| Luglio 2003 | | | | | | | G4 | H3 | | | M2 | N1 | |
| Ottobre 2003 | | | | | | | | H4 | I3 | | | N2 | O1 |
| Gennaio 2004 | | | | | | | | | I4 | L3 | | | O2 | P1 |

Lo Schema 1.1 riporta un esempio di sezioni di rotazione rispetto alle occasioni di indagine. Il campione di famiglie intervistate nell'Aprile 2003 è composto da un gruppo

di famiglie che entrano per la prima volta nel campione (M1), un gruppo di famiglie intervistate per la seconda volta (L2), un gruppo di famiglie intervistate per la terza volta (G3) ed un gruppo di famiglie intervistate per la quarta ed ultima volta (F4). Lo schema di rotazione è detto 2-2-2 in quanto le famiglie campione sono intervistate per due trimestri successivi, escono dal campione per altri due trimestri, rientrano nel campione per altri due trimestri e poi ne escono definitivamente.

Un disegno campionario così strutturato permette di costruire due differenti tipi di archivi longitudinali, quelli riferiti alle sezioni e quelli che hanno come riferimento temporale i trimestri. I file del primo tipo contengono le informazioni riferite a tutte e quattro le occasioni di indagine di una stessa sezione (ad esempio, le informazioni relative alle famiglie della sezione F cioè F1, F2, F3 e F4). Gli archivi che si riferiscono ai trimestri, invece, contengono le informazioni relative alle unità statistiche intervistate nei due trimestri di riferimento, a prescindere dalla sezione di appartenenza.

La struttura longitudinale così congegnata consente di costruire archivi del secondo tipo a 3, a 12 e a 15 mesi di distanza. Ad esempio, l'archivio che ha come riferimento Gennaio-Aprile 2003 riporta le informazioni delle famiglie appartenenti alle sezioni F3, F4, L1 ed L1, nell'archivio a 12 mesi (Aprile 2002-2003) sono presenti le sezioni F2, F4, G1 e G3 mentre nell'archivio Gennaio 2002-Aprile 2003 sono presenti le sezioni F1 ed F4.

Il campione che deriva dall'abbinamento longitudinale, come sopra specificato, ha una dimensione minore del campione trimestrale *cross-section*. Nel caso di file a 3 mesi e nel caso di quelli a 12 mesi, il numero di individui che possono essere teoricamente presenti negli archivi sono circa la metà del campione *cross-section*, mentre nel caso del file a 15 mesi il numero di individui teoricamente presenti si aggira attorno ad un quarto del campione trimestrale. Ciò comporta un riduzione del livello di precisione delle stime e quindi dei domini territoriali di studio rispetto alla rilevazione trimestrale e, a maggior ragione, rispetto alla media annua.

### 1.2.1 L'unità statistica longitudinale

Il disegno di campionamento della RTFL, come di ogni altra indagine su famiglie e individui condotta dall'Istat, prevede l'estrazione di un dato numero di famiglie dall'anagrafe del comune campione. La famiglia campione è individuata tramite le notizie relative all'intestatario della scheda anagrafica. Ogni membro della famiglia così individuata entra automaticamente nel campione.

L'*unità di rilevazione cross-section* è costituita dalla famiglia di fatto (si veda *par*. 1.1.2). Definita l'unità di rilevazione *cross-section* è necessario definire *l'unità statistica longitudinale*. A tal senso si introduce una *regola di continuità* per poter stabilire se una famiglia k al tempo t+1 ($F_{t+1,k}$) può essere considerata come la *continuazione* della famiglia i al tempo t ($F_{t,i}$) oppure se deve essere considerata come una nuova famiglia.
Nel caso della RTFL la regola di continuità utilizzata può essere così definita:

1) nel caso in cui, tra il tempo t ed il tempo t+1, la persona di riferimento esce dalla famiglia (per morte, ecc.) e almeno uno dei membri rimane allo stesso indirizzo, o ad un altro indirizzo dello stesso comune di residenza, allora la seconda famiglia è la continuazione della famiglia originaria;
2) nel caso in cui, tra il tempo t ed il tempo t+1, la persona di riferimento esce dalla famiglia (per morte, ecc.) e nessuno dei membri rimane nello stesso comune di residenza ciò da luogo a due famiglie differenti in senso longitudinale.

La regola di continuità utilizzata per definire una famiglia longitudinale è caratterizzata dalla natura particolare del disegno di campionamento della RTFL.

Come in ogni altro tipo di indagine, anche per la componente longitudinale della RTFL 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 ben presente è che la popolazione si modifica nell'arco di un determinato periodo a causa di entrate (nascite, immigrazione) e uscite (morti, emigrazione) .

E' chiaro che la scelta della popolazione di riferimento condiziona il tipo di matrice di transizione che si può costruire. Alla diffusa esigenza e richiesta di dati di flusso riguardanti l'intera popolazione si contrappone l'esigenza di conservare e garantire un elevato rigore metodologico.

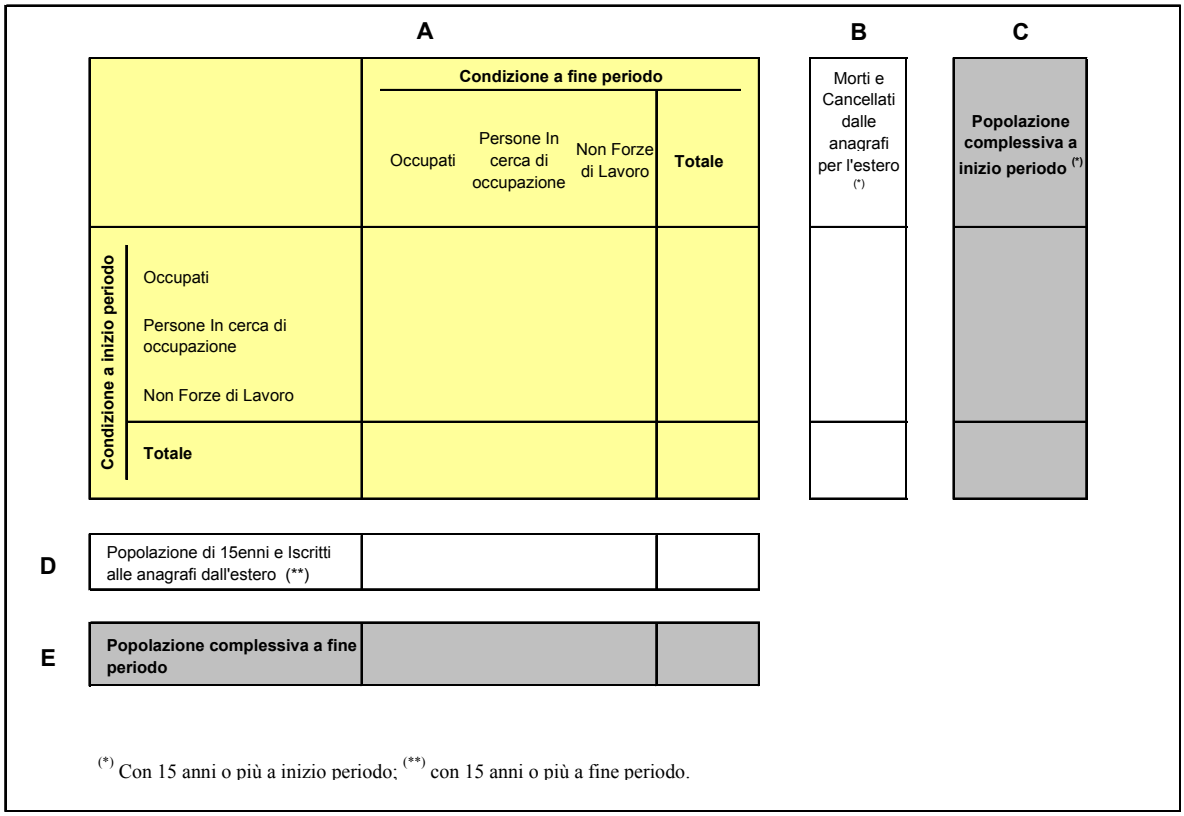
E' d'obbligo precisare che quando si ha a disposizione un vero e proprio panel, in cui gli individui che si spostano nel territorio nazionale vengono comunque intervistati nelle *wave* successive, si può costruire una matrice completa dei flussi tra condizioni cosi come viene riportata nello Schema 1.1. Considerando solo la popolazione in età lavorativa in entrambe le occasioni (con almeno 15anni a inizio periodo), la matrice completa contiene:

- la **matrice di transizione** (indicata con la lettera A nello Schema) con la distribuzione congiunta secondo la condizione a inizio e fine periodo per la popolazione che risiede sempre nel territorio nazionale sia a inizio sia a fine periodo;
- due vettori con la distribuzione per condizione della popolazione complessiva, sia a inizio sia a fine periodo, risultante dai relativi campioni trasversali (rispettivamente indicati con C e E);
- due vettori di raccordo tra dati di trasversali e longitudinali che riportano la condizione a inizio periodo per coloro che risultano morti o emigrati (indicato con B) e la condizione a fine periodo per coloro che compiono 15 anni e gli immigrati (indicato con D).

Come già affermato, la costruzione degli archivi longitudinali della RTFL e la conseguente possibilità di fornire stime sui 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 *cross-section* dei principali indicatori strutturali del mercato del lavoro[2]. La componente longitudinale, che è un sottoprodotto della RTFL, non può essere considerata come un vero e proprio panel, essa infatti non può fornire informazioni sulla condizione occupazionale, a inizio e fine periodo, relativamente a tutta la popolazione di partenza, ma solo per una parte, seppur considerevole, di questa. Questo limite è dovuto al fatto che il disegno campionario della RTFL non prevede di seguire sul territorio, per le interviste successive, ne gli individui che escono dalla famiglia campione, ne le famiglie intere che cambiano residenza verso altri comuni o verso l'estero. Chi cambia comune, anche se all'interno della stessa provincia, ha probabilità pari a zero di entrare nel campione longitudinale.

[2] Essenzialmente, le limitazioni sono dovute: all'obiettivo principale dell'indagine che deve produrre stime *cross-section*, alla regola di continuità della famiglia longitudinale, assenza del codice univoco di identificazione dell'individuo, e dalla ridotta possibilità di controllare in tempo utile i codici di identificazione della famiglia nelle diverse sezioni di rotazione.

**Schema 1.1: Schema della matrice completa degli stock e dei flussi della popolazione complessiva e popolazione residente sul territorio nazionale**




Da tali considerazioni ne consegue che il campione longitudinale RTFL, che scaturisce dall'abbinamento di due trimestri, è in grado di rappresentare correttamente solo la popolazione che **risiede nello stesso comune nei due istanti di tempo considerati**. Tale popolazione, che per comodità espositiva chiameremo *popolazione longitudinale*, è calcolata 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. Ad esempio, la popolazione longitudinale di una regione è composta da tutti gli individui che al trimestre successivo continuano a risiedere nello stesso comune della regione; tale popolazione risulta sicuramente inferiore alla popolazione che continua a risiedere nella regione (Centra, Discenza e Rustichelli, 2001)[3].

Per la scelta della popolazione di riferimento si presentano le due seguenti alternative:

1) la popolazione osservata a inizio periodo al netto dei morti e degli emigrati;
2) la popolazione longitudinale come sopra definita.

[3] Quest'ultima popolazione, con la denominazione di "compresente" è stata utilizzata da M. Centra, A.R. Discenza e E. Rustichelli in "Strumenti per le analisi di flusso nel mercato del lavoro. Una procedura per la ricostruzione della struttura longitudinale della Rilevazione trimestrale Istat sulle forze di lavoro". Monografie sul mercato del lavoro e le politiche per l'impiego. ISFOL, marzo 2001. Il lavoro svolto dal gruppo di ricerca dell'ISFOL, coordinato dalla Dr.ssa Marinella Giovine, è stato prezioso per la determinazione della metodologia sviluppata e adottata dall'Istat per la costruzione delle matrici di transizione.

Per utilizzare la popolazione indicata al punto 1), si deve introdurre l'ipotesi che la distribuzione doppia della condizione professionale a inizio e fine periodo per gli individui che non cambiano residenza tra i due istanti di tempo sia identica a quella degli individui che la cambiano. Ciò equivale a dire che i suddetti gruppi di popolazione abbiano un comportamento simile tra loro rispetto alla condizione professionale, e che quindi il comportamento nel mercato del lavoro di tutta la popolazione iniziale possa essere rappresentato dalla sola componente longitudinale.

Tale ipotesi non trova riscontro nella realtà. A suffragio di tale affermazione, le elaborazioni effettuate sui dati del modulo ad hoc relativo alla condizione professionale ed alla residenza nell'anno precedente[4], hanno evidenziato che tali distribuzioni sono significativamente differenti. L'utilizzo di tale popolazione, quindi, comporterebbe l'introduzione di un'accentuata distorsione (concettuale, oltre che numerica) nelle stime relative stime relative ai flussi tra condizioni.

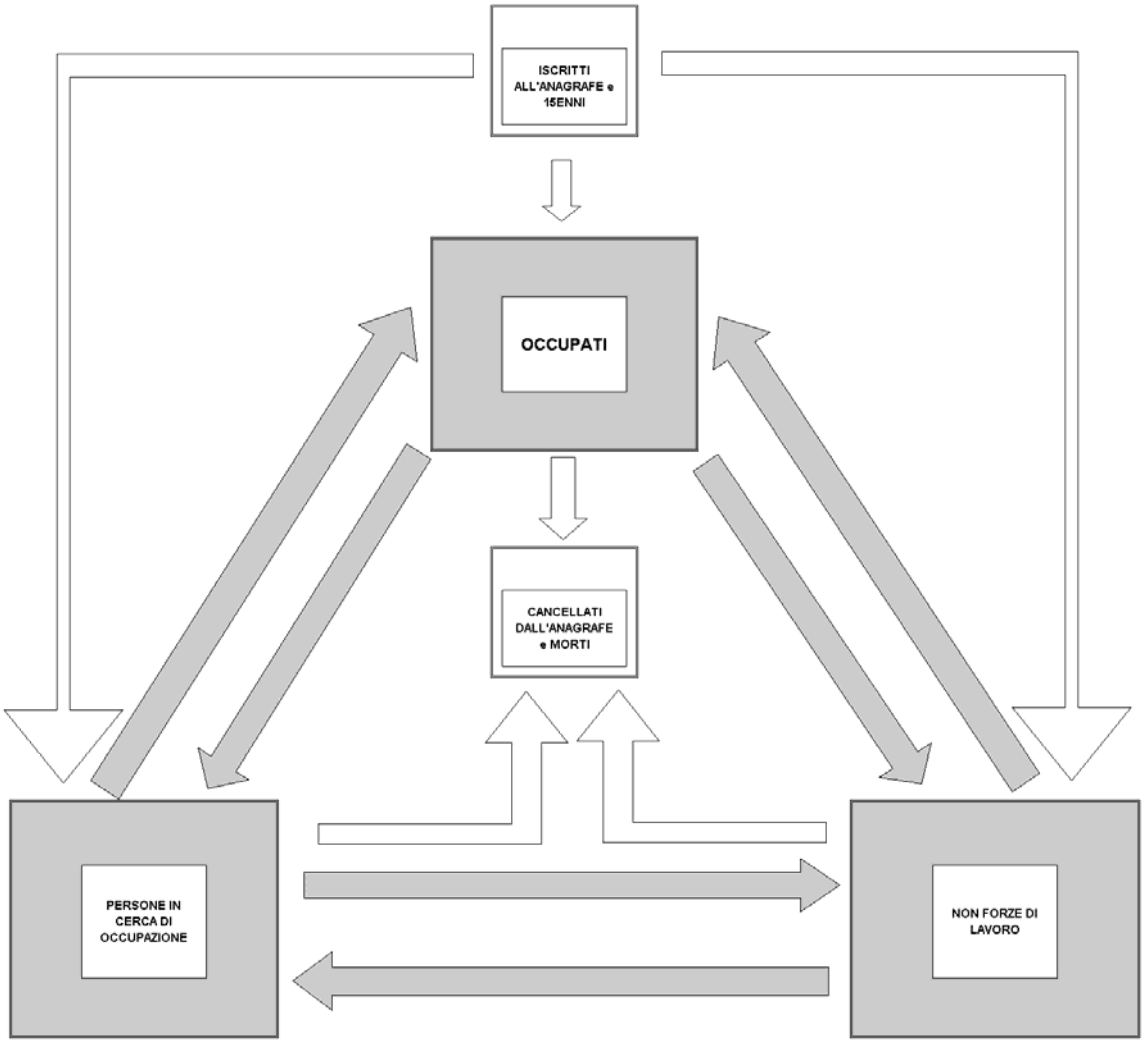
**Schema 1.2: Schema della matrice completa degli stock e dei flussi della popolazione complessiva e popolazione longitudinale**

| | | A | | | B | C |
|---|---|---|---|---|---|---|
| | | Condizione a fine periodo | | | | |
| | | Occupati | Persone In cerca di occupazione | Non Forze di Lavoro | Totale | Morti e Cancellati dalle anagrafi per l'estero o per altro comune (*) | Popolazione complessiva a inizio periodo |
| Condizione a inizio periodo | Occupati | | | | | | |
| | Persone In cerca di occupazione | | | | | | |
| | Non Forze di Lavoro | | | | | | |
| | **Totale** | | | | | | |
| D | Popolazione di 15enni Iscritti alle anagrafi dall'estero o da altro comune (**) | | | | | | |
| E | **Popolazione complessiva a fine periodo** | | | | | | |

(*) Con 15 anni o più a inizio periodo; (**) con 15 anni o più a fine periodo.

[4] Il modulo ad hoc è proposto in ogni rilevazione di aprile. Il numero di cambi di residenza e la quantità delle informazioni raccolte non consentono analisi dettagliate del fenomeno e pertanto non utilizzabili per poter correttamente utilizzare la popolazione iniziale.

**Schema 1.3: Diagramma dei flussi della popolazione complessiva**




L'utilizzo della popolazione longitudinale (indicata al punto 2), risponde ai requisiti di rappresentatività del campione e risolverebbe i problemi brevemente evidenziati. Un elemento di sconvenienza, in sede di analisi, potrebbe derivare dal fatto che le matrici di transizione prodotte fanno riferimento soltanto alla popolazione longitudinale (che è una parte di quella iniziale e finale), e che quindi tutte le analisi possono essere condotte solo matrici "al netto" dei flussi realizzatisi per la popolazione iscritta e cancellata nel periodo. Per contro, un importante elemento a favore, riguarda la possibilità di ottenere un elevato dettaglio informativo, con una minima distorsione, soltanto sulla popolazione longitudinale. Nel caso dell'utilizzo della componente longitudinale della RTFL è possibile ottenere comunque una matrice completa dei flussi relativi alla popolazione complessiva, ma questa volta essa contiene una matrice di transizione solo per la popolazione longitudinale (Schema 1.2). Nella matrice completa sono comunque presenti dei vettori aggiuntivi, ottenibili per differenza con i dati di stock relativi alla popolazione iniziale e finale, che in questo caso contengono le stime della condizione a inizio periodo di coloro che risultano morti o cancellati (per cambio di residenza verso altro comune o per l'estero) e le stime della condizione a

fine periodo dei 15enni e degli iscritti (per cambio di residenza da altro comune o estero).

Le stime possono essere fornite per tutte le persone che nel primo trimestre hanno 15 anni o più. In questo modo è possibile fornire dati comparabili e congruenti con quelli di stock del primo trimestre relativamente alla popolazione di 15 e più. Per quanto riguarda i dati di stock del secondo trimestre, è possibile assicurare la congruenza inserendo una riga aggiuntiva con le stime dei 15enni per condizione professionale fornite proprio dalla indagine stessa nel secondo trimestre.

Nello Schema 1.3 sono rappresentati i possibili flussi per la popolazione complessiva. Le frecce in grigio indicano i flussi tra condizioni occupazionali a inizio e fine periodo per la popolazione longitudinale; quelle in bianco indicano i flussi in entrata e in uscita rispettivamente nella popolazione a fine periodo e dalla popolazione a inizio periodo.

### 1.2.2 Calcolo della popolazione longitudinale

I flussi naturali e migratori intercorsi nel periodo in esame e necessari per il calcolo della popolazione longitudinale (morti, cancellati dall'anagrafe, iscritti in anagrafe, 15enni) sono stimati mediante modelli demografici e permettono di aggiornare trimestralmente l'ammontare e la struttura per sessso e età della popolazione residente. Tale stima, nota come "stima anticipata della popolazione residente per sesso e classi di età", ed è utilizzata per il riporto all'universo delle indagini campionarie su famiglie ed individui dell'Istat (Istat, 1999).

Date le seguenti quantità,

$m_{rse}$ morti che erano residenti nella regione *r*, di sesso *s* ed età *e*;
$c_{rse}$ cancellati per trasferimento di residenza in altro comune o all'estero che erano residenti nella regione *r*, di sesso *s* ed età *e*;
$n_{rse}$ nati nella regione *r*, di sesso *s* ed età *e*;
$i_{rse}$ iscritti per trasferimento di residenza da altro comune o dall'estero che risiedono nella regione *r*, di sesso *s* ed età *e;*

è possibile definire la relazione che lega la popolazione per regione, sesso ed età al tempo 1 e al tempo 2:

$$ {}_1P_{rse} - m_{rse} - c_{rse} + n_{rse} + i_{rse} = {}_2P_{rse} \qquad (1.3) $$

Dalla (1.3) è possibile definire la popolazione longitudinale:

$$ {}_lP_{rse} = {}_1P_{rse} - m_{rse} - c_{rse} = {}_2P_{rse} - n_{rse} - i_{rse} \qquad (1.4) $$

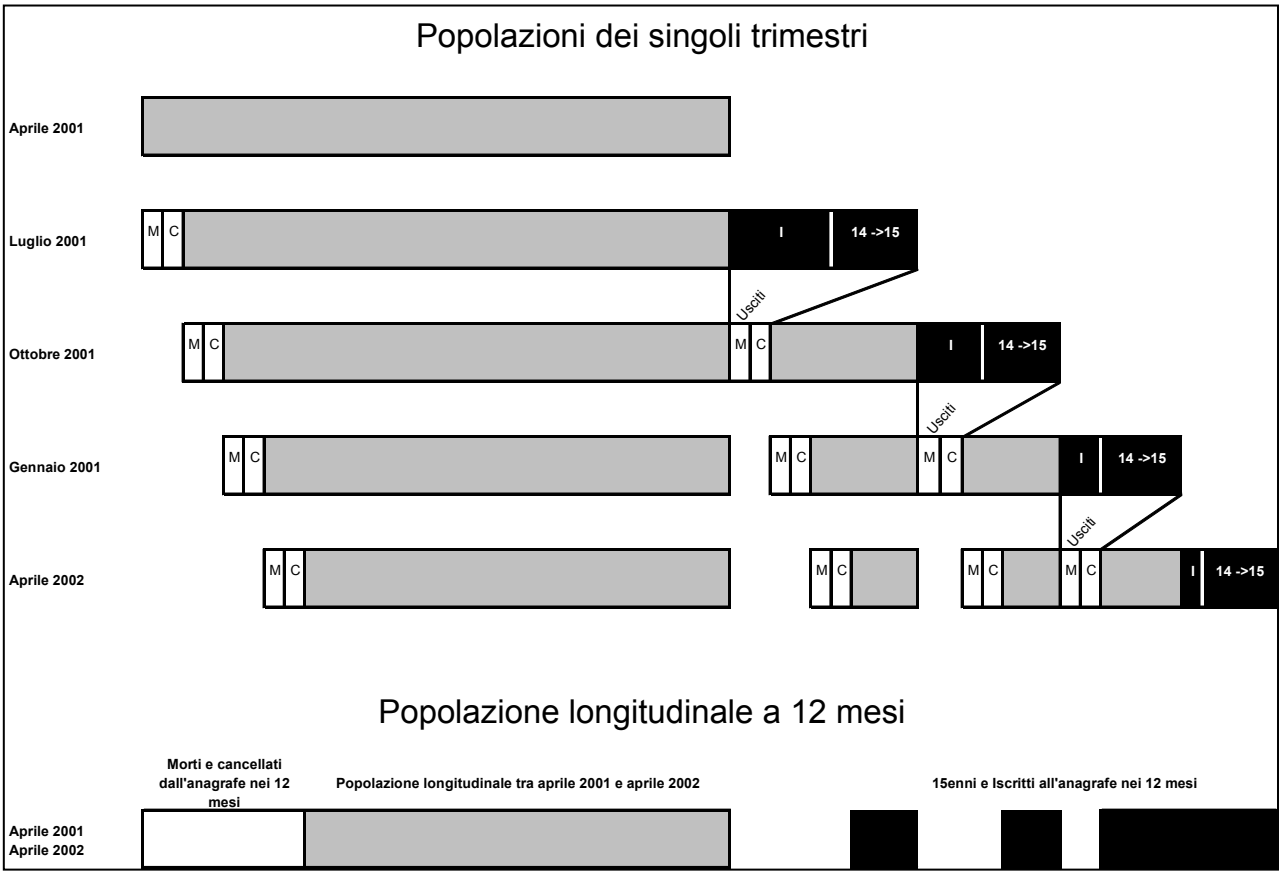
I flussi di popolazione (le quantità indicate nella 1.4) sono disponibili con cadenza trimestrale. Lo Schema 1.4 illustra come si passa, per un dato comune, dalle popolazioni trasversali di 5 trimestri consecutivi alla popolazione longitudinale a distanza di 12 mesi.

In particolare, per ogni trimestre, la popolazione trasversale è data dalla somma della popolazione che nei tre mesi precedenti non ha cambiato residenza (in grigio nello Schema 1.4) e dagli individui che nei tre mesi precedenti si sono iscritti all'anagrafe comunale o sono entrati nell'età da lavoro (in nero). Con riferimento allo

stesso trimestre una quota di popolazione viene cancellata dall'anagrafe per morte o cambio di residenza (in bianco).

La popolazione longitudinale a 12 mesi di distanza è data dalla popolazione del primo trimestre al netto delle cancellazione anagrafiche. Gli iscritti in anagrafe nei 12 mesi sono dati dalla somma degli iscritti in ciascuno dei trimestri, sopravvissuti e residenti nel comune a fine periodo.

**Schema 1.4: Schema di calcolo della popolazione longitudinale tra aprile 2001 e aprile 2002**




### 1.2.3 Le stime longitudinali

Le indagini longitudinali (che misurano un fenomeno sulla stessa unità di rilevazione in diversi istanti di tempo), oltre alle stime riferite ad un dato istante di tempo (stime *cross-section*), consentono la realizzazione di stime riferite ai cambiamenti di stato occorsi tra due istanti di tempo. In tal caso si parla di stime a carattere longitudinale e possono essere espresse nel seguente modo.

La stima del numero di individui che, in uno specifico dominio *d,* dal tempo *t* al tempo *t+k* passano dallo stato $S_t$ allo stato $S_{t+k}$ può essere rappresentata secondo la seguente formalizzazione.

Sia:

$$Y = \begin{cases} 1 & \text{se passa da } S_t \text{ a } S_{t+k} \\ 0 & \text{altrimenti} \end{cases} ;$$

*d* indice di dominio territoriale di riferimento delle stime;

$H_d$ numero di strati nel dominio d;
$_hm_{t,\ t+k}$ numero di famiglie campione rispondenti sia al tempo t sia al tempo t+k, appartenenti allo strato h;
$n_j$ numero di componenti della famiglia j;

La stima del numero di individui che dal tempo *t* al tempo *t+k* passano dallo stato $S_t$ allo stato $S_{t+k}$ è espressa da:

$$ {}_d\hat{Y}_{S_t,S_{t+k}} = \sum_{h=1}^{H_d} {}_h m_{t,t+k} \sum_{j=1}^{n_j} \sum_{k=1} y_{hjk} w_{hjk} \qquad (1.5) $$

in cui $w_{hjk}$ è il *coefficiente finale di riporto all'universo* da attribuire all'individuo k della famiglia *j* dello strato *h* che ha risposto alle occasioni di indagini al tempo *t* e a tempo *t+k*.

Nel proseguo della nota verrà illustrata la metodologia e la strategia adottate per il calcolo dei coefficienti di riporto all'universo.

# 2 L'abbinamento

## 2.1 La procedura di abbinamento con discordanze

Le indagini campionarie che prevedono una struttura longitudinale sono solitamente organizzate in modo che risulti immediato riconoscere quali siano i *record* relativi al medesimo individuo nel corso del tempo, per cui la definizione del campione longitudinale richiede unicamente la disponibilità di algoritmi efficienti di ricerca dei *record* che presentano, in archivi riferiti a tempi diversi, la medesima chiave identificativa. Nel caso della RTFL, tuttavia, un identificatore per ciascun *record* individuale non esiste e la chiave che identifica la medesima famiglia nelle diverse occasioni di indagine è soggetto a errori sia di trascrizione sia di registrazione. Per la creazione di un archivio longitudinale, quanto più esaustivo e privo di duplicazioni ed errori, è quindi necessario ricorrere a metodi per l'abbinamento esatto di *record* (*record linkage*).

L'abbinamento esatto di *record* è una metodologia che consente di combinare informazioni sulle medesime unità contenute in archivi diversi, oppure di identificare duplicazioni all'interno di archivi di grandi dimensioni. Le tecniche di abbinamento esatto consentono, ad esempio, di sfruttare in modo più penetrante le informazioni correntemente raccolte in indagini statistiche (censuarie o campionarie), anche mediante la loro integrazione con i dati presenti in archivi amministrativi. È inoltre frequente il caso in cui si ricorre all'integrazione di uno o più archivi amministrativi per ottenere un unico archivio (corretto) sul quale basare le successive procedure di analisi statistica.

Talvolta, il problema di abbinamento esatto può essere risolto banalmente attraverso l'uso di efficienti procedure informatiche volte a ricercare *record* che nei due archivi presentino lo stesso codice identificativo. E' tuttavia di maggiore interesse (e rilievo) il caso in cui non esiste un codice identificativo univoco e misurato senza errore per ciascun *record*; in tal caso il problema di abbinamento esatto diviene un problema di decisione (si tratta di decidere quando le informazioni identificative parziali disponibili sono sufficienti a concludere che due *record* si riferiscono alla medesima unità) che va risolto con l'utilizzo di adeguate procedure statistiche.

Una prima formalizzazione degli aspetti statistici legati alle procedure di abbinamento esatto si ha nel lavoro di Fellegi e Sunter (1969), che raccolgono le precedenti idee di Newcombe *et a*l. (1959) e Tepping (1968) riportandole ad una struttura coerente legata alla teoria classica della verifica di ipotesi. Rassegne sui problemi metodologici aperti e sugli sviluppi più recenti di tali tecniche si trovano in Winkler (1995) e Torelli (1998).

Tale procedura incorpora alcuni dei risultati più recenti relativi alla teoria dell'abbinamento esatto e quindi costituisce una generalizzazione e un miglioramento delle procedure già proposte al medesimo scopo (Moriani, 1981; Giusti, Marliani e Torelli, 1991).

### 2.1.1 Definizione del problema di abbinamento esatto

Per una trattazione generale del problema dell'abbinamento esatto conviene partire dalla formalizzazione proposta da Fellegi e Sunter (1969). Due archivi A e B, di dimensione $N_A$ e $N_B$, contengono rispettivamente *record* a e b, una parte dei quali sono relativi ai medesimi individui; lo spazio prodotto $A \times B = \{(a,b); a \in A, b \in B\}$, che

include tutte le $N = N_A \times N_B$ possibili coppie di *record* originate dal confronto, è pertanto l'unione di due insiemi disgiunti:

- l'insieme M delle coppie relative allo stesso individuo;
- l'insieme U delle coppie con *record* relativi a due individui differenti.

I procedimenti di abbinamento esatto di *record* sono essenzialmente metodi di decisione per classificare ogni coppia come appartenente ad uno dei due insiemi M ed U. Se esiste un identificatore che permette di individuare con certezza i *record* relativi ad ogni individuo, il procedimento si riduce ad un algoritmo di ricerca di coloro che presentano la medesima chiave di identificazione; nel caso, invece, in cui una chiave identificativa, unica per ogni *record*, non esista oppure sia osservabile con errore, si tratta di impostare il problema di abbinamento come un problema di decisione che auspicabilmente conduca a rendere minimo il numero di errori di classificazione. In questo contesto le decisioni errate possono essere di 2 tipi:

- errati abbinamenti (falsi positivi): si classifica la coppia in M, essendo in realtà *a* e *b* relativi ad individui differenti;
- mancati abbinamenti (falsi negativi): si classifica la coppia in U, essendo in realtà *a* e *b* relativi al medesimo individuo.

Poiché M ed U sono insiemi mutuamente esclusivi, non è possibile minimizzare contemporaneamente il numero di mancati abbinamenti ed il numero di errati abbinamenti: ad un elevato numero di errate decisioni per una delle due classi corrisponde una diminuzione degli errori nell'altra direzione. La scelta di un metodo di abbinamento è legata pertanto alla valutazione della gravità relativa che si attribuisce ai due tipi di errore conseguenti al processo di decisione.

Nella gran parte delle situazioni applicative si ritiene che meriti un maggiore impegno l'obiettivo di evitare gli errati abbinamenti. Ciò è vero specialmente nel caso di abbinamenti di *record* per condurre analisi dinamiche: in tal caso, errati abbinamenti porterebbero ad associare informazioni che sono in realtà relative ad individui differenti, dando luogo a mobilità spuria. I mancati abbinamenti sono invece, almeno in tale situazione applicativa, considerati meno gravi. La più immediata conseguenza dei mancati abbinamenti si traduce, infatti, solo in una non eccessiva riduzione della dimensione campionaria. E' comunque da valutare con attenzione la possibilità che agli errati abbinamenti sia associato un problema di selettività del campione; ciò avviene se gli individui vengono esclusi dall'abbinamento per motivi correlati con quelli di interesse nell'analisi: si pensi ad esempio alla maggior probabilità di commettere errori nella risposta per individui anziani e/o con un basso livello di istruzione.

### 2.1.2 Variabili di confronto e strategie di blocco

Una procedura di abbinamento conduce a stimare, per ognuna delle *N* possibili coppie di *record*, il valore ignoto di una variabile indicatrice *G*, che vale 1 per le coppie in M e 0 per quelle in U. A tal fine è possibile utilizzare i valori assunti, nei *record*, da alcune variabili di confronto; in particolare, Newcombe (1988) osserva che tali variabili devono avere la capacità di discriminare al meglio gli individui presenti nei due archivi, in caso di discordanza, di concordanza o, nella migliore delle ipotesi, in entrambi i casi. La variabile “sesso”, ad esempio, dà poche informazioni se è concordante, mentre fornisce una forte indicazione negativa sull'abbinamento se si osserva una discordanza.

Appare pertanto chiaro che, oltre alla scelta delle variabili di confronto, assume estrema importanza la definizione di concordanza che viene assegnata ad ogni possibile confronto fra le variabili. Al fine di sfruttare al meglio le informazioni provenienti dal confronto, sarebbe necessario tenere in considerazione tutte le possibili combinazioni di modalità che possono ottenersi quando si confronti la stessa variabile presente nei due archivi; è evidente che ciò è tanto più difficile quanto maggiore è il numero di modalità. Sono pertanto necessarie delle modifiche nella definizione di concordanza che permettano di aggregare quei risultati che forniscono informazioni simili sull'abbinamento; la dimensione di tale processo di aggregazione dipende essenzialmente dalla parsimonia richiesta al modello e dalle numerosità campionarie di cui si dispone. Copas e Hilton (1990) mostrano come sia possibile calcolare la perdita di informazione derivante da definizioni più restrittive, e propongono un modello di misura in forma parametrica che permette di sfruttare al meglio i risultati del confronto pur mantenendo una scelta parsimoniosa.

Fra i metodi più utilizzati, il confronto può dare semplicemente un risultato dicotomico, con valori 1 in caso di concordanza e 0 con discordanza fra le variabili; per una specificazione più dettagliata, possono essere previsti diversi livelli di concordanza (ad esempio per l'età, tenendo conto della differenza in anni), o diverse capacità discriminanti per specifici valori delle variabili (ad esempio nel caso di cognomi più o meno comuni, per cui la concordanza di cognomi diffusi fornisce minori informazioni).

Una volta scelte le variabili di confronto e le definizioni di concordanza, l'informazione ottenibile per la j-esima coppia può essere riassunta in un vettore, che ha come singolo elemento il risultato del confronto fra le i-esime variabili:

$$\gamma_j = \left[\gamma_j^1, \gamma_j^2, ..., \gamma_j^i, ..., \gamma_j^I\right], \qquad j = 1...N.$$

Il confronto effettuato su tutte le coppie di *record* appartenenti ai due archivi può comunque portare ad un carico computazionale molto elevato per archivi di grandi dimensioni. Se è possibile osservare variabili di confronto con elevata affidabilità ed alto potere discriminante, una buona strategia consiste nel ridurre lo spazio dei confronti all'interno di un blocco di *record* che presentano concordanza perfetta su tali variabili.

Il vantaggio di tale strategia è di ridurre, spesso drasticamente, il numero di confronti ammissibili, ottenendo contemporaneamente una notevole riduzione del carico computazionale e una forte protezione contro i falsi positivi; le dimensioni di tali effetti dipendono ovviamente dalla capacità discriminante delle variabili di blocco. Di contro, tale procedura può condurre ad un aumento di falsi negativi se non è elevata l'affidabilità delle variabili di blocco prescelte. La scelta dipende pertanto essenzialmente dal peso che si vuole dare ai due tipi di errori, oltre alla disponibilità di tempo e mezzi dal punto di vista computazionale; Kelley (1985) suggerisce alcuni metodi per definire una strategia di blocco che permetta di minimizzare i costi complessivi, sia computazionali che in termini di errore negli abbinamenti.

### 2.1.3 I pesi di abbinamento: metodi deterministici e probabilistici

Definiti i vettori di confronto $\gamma$, rimane da stabilire come questi possano essere utilizzati per la decisione sulla classificazione delle coppie in M o U. Una possibilità è l'assegnazione ad ogni vettore di un peso *w*, sul valore del quale si basa il seguente processo decisionale per la j-esima coppia:

- $w_j \geq K_u \Rightarrow (a_j, b_j) \in M$ la coppia viene abbinata;
- $K_l \leq w_j < K_u$ la decisione viene rinviata; (2.1)
- $w_j < K_l \Rightarrow (a_j, b_j) \in U$ la coppia non viene abbinata.

La stima dei pesi *w* e la scelta delle soglie K sono ovviamente cruciali nella definizione del procedimento. Nel caso più semplice si può utilizzare un criterio deterministico, dove implicitamente i valori dei pesi e delle soglie sono fissati a priori in funzione degli specifici obiettivi dell'abbinamento; in tal caso, la scelta deve essere definita in base alla predisposizione verso gli errori di abbinamento, con soglie elevate che proteggono dai falsi positivi ma sono spesso associate ad un numero elevato di falsi negativi. L'intervallo tra le due soglie non deve essere inoltre troppo ampio, in quanto la scelta di rinviare la decisione, associata spesso ad un controllo manuale delle coppie, presenta solitamente costi elevati.

Un semplice esempio di criterio deterministico consiste nell'associare i pesi *w* al numero di concordanze osservate; la scelta di abbinare può avvenire ad esempio per tutte le coppie con al massimo una discordanza, con una soglia unica implicita pari ad *l*-1 nel processo decisionale (2.1). In alternativa, si possono utilizzare pesi differenti per le singole variabili, ammettendo ad esempio due errori su quelle ritenute meno discriminanti.

Nella formulazione di Fellegi e Sunter i pesi vengono invece stimati in modo probabilistico attraverso il rapporto fra le due verosimiglianze del vettore di confronti, rispettivamente nel caso di coppie relative allo stesso individuo (M) e coppie abbinate casualmente (U):

$$w_j = \ln \frac{P(\gamma_j | M)}{P(\gamma_j | U)} = \ln \frac{m_j}{u_j}. \qquad (2.2)$$

Essendo *w* un rapporto di verosimiglianza, statistica sufficiente per il problema di decisione, Fellegi e Sunter dimostrano che utilizzando la (2.2) la regola di decisione (2.1) è ottimale per ogni coppia di soglie $(K_l, K_u)$; l'ottimalità assume qui il significato di minimizzazione della regione di indecisione, ed ha come conseguenza, ad esempio, la possibilità di fissare a priori i livelli di errore desiderati, sia per quanto riguarda i falsi positivi che i falsi negativi, rendendo minimo il numero di coppie da abbinare manualmente.

Kirkendall (1985), oltre a proporre alcuni esempi pratici per il calcolo dei pesi in (2.2) con differenti variabili di confronto, ne propone un'ulteriore possibile interpretazione in termini di teoria dell'informazione: nel caso i logaritmi siano espressi in base 2, i pesi sono esprimibili come *odds ratios* che permettono di aggiornare l'informazione a priori attraverso i risultati del confronto.

### 2.1.4 Stima dei pesi con l'algoritmo EM

Il problema principale della procedura proposta da Fellegi e Sunter è la stima delle probabilità *m* ed *u* definite in (2.2), la cui accuratezza condiziona fortemente la proprietà di ottimalità. Come osserva tra gli altri Winkler (1995), risulta infatti spesso irragionevole l'assunzione che esistano campioni per i quali sia certa l'appartenenza delle coppie ad U e, soprattutto, a M. Inoltre, anche se tali campioni fossero disponibili, le stime risultanti per un'applicazione potrebbero non adattarsi ai veri, ma ignoti, valori relativi al campione che si vuole effettivamente abbinare. A tal fine, sarebbe invece

necessario conoscere esattamente il valore della variabile *G* per tutte le coppie da abbinare, il che non è ovviamente possibile.

Tepping (1968) propone di effettuare una partizione preliminare delle coppie negli insiemi M ed U, stimando all'interno di questi campioni le probabilità necessarie; in questo modo si potrebbe tra l'altro evitare di ricorrere, nella stima di *m* ed *u*, alle ipotesi spesso poco realistiche di indipendenza fra gli errori nelle singole variabili di confronto, necessarie per i metodi di stima proposti da Fellegi e Sunter. Seguendo Jaro (1989), è possibile effettuare una partizione simile a quella proposta da Tepping in modo iterativo, imputando ad ogni passo il valore di *G* per tutte le coppie, e ristimando le probabilità seguendo la logica dell'algoritmo EM (Dempster *et al*., 1977).

Per poter applicare l'algoritmo, è necessario definire la funzione di verosimiglianza dei parametri *m* ed *u*, congiuntamente a *p*, la frazione di coppie da abbinare:

$$L(m,u;p)=\prod_{j=1}^{N}\left[P(M)P(\gamma_j|M)\right]^{g_j}\left[P(U)P(\gamma_j|U)\right]^{1-g_j}=\prod_{j=1}^{N}\left[pm_j\right]^{g_j}\left[(1-p)u_j\right]^{1-g_j}.$$

Se si fissano i valori dei parametri *m*, *u* e *p*, al passo *E* dell'algoritmo EM è possibile stimare il valore atteso della variabile indicatrice *G*:

$$\hat{g}_j=E(g_j|m_j,u_j,p)=\frac{pm_j}{pm_j+(1-p)u_j}=\frac{m_j/u_j}{m_j/u_j+(1-p)/p}=\frac{e^{w_j}}{e^{w_j}+(1-p)/p}. \qquad (2.3)$$

Si noti come il valore atteso di *G* abbia un legame diretto (*logit*) con i pesi *w* di Fellegi e Sunter, che vengono così riportati in una scala 0-1 e resi più interpretabili rispetto ai valori originali.

Il passo *M* consiste nel massimizzare la verosimiglianza per i parametri *m* e *p*, condizionatamente al valore assunto da *G*. Seguendo Jaro (1989), si ritiene invece migliore una stima degli *u* effettuata al di fuori dell'algoritmo, su un campione di coppie abbinate casualmente senza tenere conto del blocco; in questo modo è inoltre possibile allentare l'ipotesi di indipendenza fra gli errori, in modo da tener conto delle eventuali correlazioni fra le diverse variabili (si pensi ad esempio alla stretta relazione fra “nome proprio” e “sesso”). Per la stima di *m*, l'ipotesi di indipendenza fra gli errori nelle singole variabili appare invece più realistica (Thibaudeau, 1993) e permette notevoli semplificazioni computazionali, pur non essendo una scelta obbligata per il metodo proposto.

Con l'assunzione di indipendenza, i valori di *m* per le singole variabili possono essere stimati su un campione “virtuale” di coppie appartenenti a M, pesando ogni singola coppia con il valore atteso di *G* calcolato in (2.3); la stima avviene attraverso le frequenze con cui i singoli risultati del confronto si presentano nel campione pesato:

$$\hat{m}^i=\frac{\sum_{j=1}^{N}\gamma_j^i\hat{g}_j}{\sum_{j=1}^{N}\hat{g}_j},\quad i=1..I. \qquad (2.4)$$

La stima di *m* per le singole coppie, sempre per l'ipotesi di indipendenza, avviene in modo moltiplicativo, utilizzando le stime provenienti dalla (2.4) a seconda dei risultati del confronto presenti nei vettori $\gamma$:

$$\hat{m}_j = \prod_{i=1}^{I} \left(\hat{m}^i\right)^{\gamma_j^i} \left(1 - \hat{m}^i\right)^{1-\gamma_j^i} . \qquad (2.5)$$

Infine, la stima di *p* è data semplicemente dalla numerosità relativa del campione "virtuale", ottenuta attraverso la media dei valori assunti dalla variabile indicatrice *G*:

$$\hat{p} = \frac{\sum_{j=1}^{N} \hat{g}_j}{N} . \qquad (2.6)$$

Poiché il metodo dipende esclusivamente dai risultati del confronto, è possibile ottenere una rappresentazione più compatta dei vettori attraverso la distribuzione delle frequenze di tutti i possibili risultati ammissibili, compatibilmente con la codifica delle concordanze e la procedura di blocco. Se, ad esempio, si definiscono i vettori $\gamma$ in modo dicotomico, si ottiene la seguente distribuzione:

$$\gamma_{(k)} = [1,0,1...,1,0] \quad \text{con frequenza } f_{(k)}, k = 1...K, K \leq 2^I .$$

Con questa nuova caratterizzazione, il metodo viene reso notevolmente più veloce, poiché è sufficiente un'unica stima della (2.5) per tutti i vettori $\gamma$ che si rivelano identici. Inoltre, anche l'utilizzo di (2.4) e (2.6) viene semplificato, con l'immediata estensione al caso in cui le medie calcolate vengono ponderate attraverso le frequenze con cui ogni singolo tipo di vettore viene osservato.

### 2.1.5 Scelta della soglia e stima degli errori

Al fine di valutare l'efficienza di un qualunque metodo di abbinamento di *record* è cruciale disporre di stime del numero di errati abbinamenti e mancati abbinamenti che conseguono alla sua applicazione; tali stime, fra l'altro, consentono di affrontare razionalmente il problema della scelta della soglia che consente di decidere quali coppie abbinare a quali no. Belin e Rubin (1995) osservano, attraverso alcune verifiche empiriche sui risultati dell'abbinamento, che l'effetto maggiore sugli errori di abbinamento è legato ad una cattiva scelta della soglia, mentre sembrano di minor rilievo la fase di definizione e stima dei pesi. La soglia deve pertanto essere determinata in un'ottica di minimizzazione degli errori, che devono essere stimati con precisione. Belin e Rubin mostrano che invece gran parte dei metodi usualmente utilizzati in precedenza si rivelano eccessivamente ottimisti, con una notevole sottostima degli errori; ciò vale in particolare per i metodi che prevedono l'ipotesi di indipendenza fra gli errori nelle diverse variabili di confronto.
Un primo semplice metodo di verifica possibile è un controllo manuale effettuato su un campione di *record*; in particolare, se il sistema di confronti permette di discriminare con buona precisione le coppie, risulta spesso sufficiente prevedere un controllo limitato alle coppie con pesi "vicini" alla soglia, la cui assegnazione è più dubbia.
Un metodo spesso utilizzato per una prima approssimazione della soglia migliore, o per definire quali siano le coppie da verificare manualmente, consiste in un'analisi grafica della distribuzione dei pesi sull'intero campione. Questa è la mistura di due distribuzioni che, se la strategia utilizzata è sufficientemente discriminante, sono concentrate in punti distanti fra loro; la distribuzione osservata dovrebbe pertanto presentare un'accentuata bimodalità, con l'altezza relativa delle due mode che dipende essenzialmente dal numero totale di confronti effettuati e dalle strategie di blocco. La

maggiore incertezza rimane nella zona in cui le code delle due distribuzioni condizionate si intersecano, e gli errori dipendono dal punto esatto in cui si posiziona la soglia, con una conferma della relazione inversa fra le proporzioni di falsi positivi e falsi negativi.

Una recente alternativa al controllo manuale è fornita da metodi legati alla modellazione statistica degli errori di abbinamento. Belin e Rubin presentano metodi che si basano, seppure con interpretazioni differenti, su una stima che tenga conto della presenza della variabile latente *G*, relativa alla classificazione delle singole coppie in M o U. Si distinguono in particolare due classi di modelli: (a) un approccio diretto che prevede una regressione logistica della variabile *G* sui pesi *w*; (b) un approccio indiretto basato sulla stima di un modello di mistura che tenga conto delle due distinte distribuzione condizionate dalle quali è formata la distribuzione osservata dei pesi.

Un punto vincolante della formulazione di Belin e Rubin è che in entrambi i metodi è richiesta la disponibilità di un campione di abbinamenti certi, a partire dal quale poter stimare alcuni parametri da utilizzare nella stima degli errori. Tale assunzione, oltre a non essere spesso attuabile nella pratica per la mancanza di tale campione, pone delle restrizioni forti sulla somiglianza delle differenti situazioni di abbinamento. In particolare, per la regressione logistica l'assunzione è che i parametri che legano *G* ai pesi *w* siano gli stessi per tutte le procedure di abbinamento, e si utilizzano le stime sul campione iniziale per abbinare il campione di interesse. Per il modello di mistura, l'assunzione riguarda invece il rapporto fra le varianze delle due distribuzioni condizionate dei pesi, oltre ai parametri delle trasformazioni Box-Cox necessarie per renderle normali; in particolare la prima ipotesi appare restrittiva, alla luce delle forme differenti che assumono le distribuzioni dei pesi al variare, ad esempio, delle strategie di blocco o della proporzione di individui potenzialmente abbinabili nei due archivi.

Winkler (1995) sostiene che il metodo di Belin e Rubin, oltre a dipendere fortemente dal campione utilizzato per le stime iniziali, fornisce risultati soddisfacenti solo nel caso particolare in cui l'ipotesi di indipendenza fra gli errori non sia troppo restrittiva e le due distribuzioni condizionate dei pesi risultino ben distinte. Per una stima più precisa delle probabilità di errore nei frequenti casi in cui tali assunzioni non sono verificate, fra le altre proposte Winkler suggerisce alcune restrizioni di convessità nello spazio parametrico, che consentono di limitare le possibili soluzioni a quelle ritenute più realistiche e velocizzare la massimizzazione della verosimiglianza.

Torelli e Paggiaro (1999) propongono invece un'alternativa che permette di stimare i parametri direttamente attraverso l'algoritmo EM proposto nel sottoparagrafo 2.1.4. In particolare, il legame logistico fra probabilità di abbinamento e pesi *w* riscontrato nella (2.3) fa propendere per un metodo diretto di stima, che a differenza di quello analizzato da Belin e Rubin consente di allentare l'ipotesi di indipendenza. I risultati principali, ottenuti anche a partire da alcuni esperimenti di simulazione, confermano che la sottostima degli errori osservata da Belin e Rubin dipende essenzialmente dalla qualità della stima dei pesi. In particolare, si osserva una buona stima della quota di errati abbinamenti se si considera la dipendenza fra le variabili nella stima di *u*, mentre vi è un'evidente sottostima con l'assunzione di indipendenza.

## 2.2 La procedura di abbinamento applicata ai dati della RTFL

### 2.2.1 Gli individui abbinabili

Nel paragrafo 1.2 è stato illustrato lo schema di rotazione della RTFL ed evidenziato che i file longitudinali a tre e dodici mesi possono avvalersi al massimo della metà del campione trimestrale.

Se in teoria, il 50% del campione potrebbe essere abbinabile, in pratica tale limite non si raggiunge perché gli individui abbinabili sono solo quelli residenti in comuni che partecipano a entrambe le occasioni di indagine. Anche escludendo i comuni che non hanno partecipato ad entrambe le indagini, non tutti gli individui abbinabili possono essere abbinati sia a causa delle effettive uscite dal campione sia per gli errori sui codici identificativi o per il rifiuto dell'intervista.

**Schema 2.2 Individui di 15 anni e più abbinabili, abbinati e non abbinati**

<table>
<tr><th colspan="2">Individui campione del trimestre $t_1$</th><th>Abbinabili secondo il risultato dell'abbinamento e il motivo del mancato abbinamento</th><th>Informazioni relative al trimestre $t_1$</th><th>Informazioni relative al trimestre $t_2$</th><th>Composi-zione percentuale media</th><th>Popolazione rappre-sentata</th></tr>
<tr><td rowspan="5">Individui abbinabili (90%)</td><td>ABBINATI</td><td>ABBINATI CON E SENZA DISCORDANZE</td><td>PRESENTI</td><td>PRESENTI</td><td>91,0%</td><td rowspan="3">Individui residenti nello stesso comune ad inizio e fine periodo – Individui eleggibili per rappresentare la popolazione longitudinale</td></tr>
<tr><td rowspan="4">NON ABBINATI</td><td>ERRORI SULLE CHIAVI</td><td>PRESENTI</td><td>NON PRESENTI</td><td rowspan="3">5,6%<br>(di cui il 2,5% per sostituzioni)</td></tr>
<tr><td>RIFIUTI</td><td>PRESENTI</td><td>NON PRESENTI</td></tr>
<tr><td>IRREPERIBILITÀ</td><td>PRESENTI</td><td>NON PRESENTI</td></tr>
<tr><td>MORTI</td><td>PRESENTI</td><td>NON PRESENTI</td><td rowspan="2">3,4%</td><td rowspan="2">Individui usciti dalla popolazione – Individui non eleggibili per rappresentare la popolazione longitudinale</td></tr>
<tr><td>Cambiamenti di residenza per altro comune o estero</td><td>PRESENTI</td><td>NON PRESENTI</td></tr>
<tr><td colspan="2">NON ABBINABILI (10%)</td><td>Comuni che partecipano solo all'indagine del trimestre $t_1$</td><td colspan="4"></td></tr>
</table>

Essenzialmente, il campione degli individui abbinabili del primo trimestre possono essere classificati in due gruppi principali:

1. gli individui che hanno partecipato all'indagine sia nel primo che nel secondo trimestre e che sono stati abbinati (senza e con discordanze). Su questi si hanno le informazioni per entrambi i trimestri;
2. gruppo residuale degli individui che hanno partecipato all'indagine nel trimestre $t_1$ ma che, per vari motivi, non sono stati abbinati. Su di questi abbiamo l'informazione solo nel primo trimestre.

L'attenzione va posta su questo secondo gruppo che a sua volta può essere distinto in:

a. *individui eleggibili*, tutti coloro che sono potenzialmente intervistabili al trimestre $t_2$ perché risultano ancora residenti nello stesso comune e fanno ancora parte, "di fatto" della stessa famiglia campione (cfr. par. 1.1.2);
b. *individui non eleggibili*, tutti coloro che nel periodo considerato sono usciti dalla famiglia di origine, sono morti, hanno cambiato residenza o sono emigrati all'estero.

Lo Schema 2.2 riassume la situazione degli individui abbinabili tra due generici trimestri classificando gli individui non abbinabili e non abbinati secondo il motivo del mancato abbinamento. Tale classificazione, purtroppo, può essere fatta solo a livello concettuale, poiché non materialmente possibile identificare sul file i record dei non abbinati per i diversi motivi. Come si vedrà nel successivo paragrafo, questo problema pone alcuni limiti sia nella definizione della popolazione di riferimento per la componente longitudinale sia nella possibilità di utilizzare modelli per la "mancata risposta" individuale e familiare.

### 2.2.2 I risultati dell'abbinamento

Nell'attuale procedura della RTFL non esiste una chiave identificativa unica degli individui, ma solo della famiglia. L'individuo viene quindi identificato dalla combinazione di delle seguente set di variabili ad esso relative:

- *chiave identificativa della famiglia*, regione, provincia, comune, codice della famiglia (come da modello P48[5]), sezione di rotazione;
- *notizie relative all'individuo*, sesso, data di nascita, stato civile, titolo di studio e relazione di parentela con il capofamiglia.

Date le variabili la procedura per la realizzazione degli archivi longitudinali può essere riassumere nei seguenti punti essenziali:

1. *abbinamento esatto*: gli individui sono abbinati quando le variabili identificative nei diversi trimestri risultano identiche. In media si abbina l'80% degli individui potenzialmente abbinabili a 3 mesi e il 75% dei record a 12 mesi

2. *abbinamento con discordanze: i record che verosimilmente si riferiscono allo stesso* individuo sono abbinati ammettendo la possibilità di un numero limitato di errori sulle variabili identificative nei diversi trimestri. Tale procedura ecessaria in quanto l'evidenza empirica mostra che tali errori non sono casuali. Considerare solo i record che si abbinano senza discordanze introduce, quindi, una distorsione. In media, nel complesso, si abbina circa il 90% degli individui potenzialmente abbinabili.

3. *procedura di recupero*: tale procedura effettua il controllo deterministico e l'eventuale abbinamento di individui appartenenti a famiglie con individui abbinati (una quota variabile tra lo 0,5 e l,1,5 per cento).

[5] Il P48 è un modello ausiliario della RTFL dove sono riportate le notizie relative alle famiglie da intervistare per quel comune.

### 2.2.3 Le possibili distorsioni

In teoria l'abbinamento non porrebbe particolari problemi, ma in pratica ci sono una serie di problemi che possono provocare distorsioni di diversa natura nei risultati. Le distorsioni che si possono verificare sono di tre tipi:

*Distorsioni dovute alla procedura di abbinamento*

1. Maggiore probabilità ottenere falsi positivi: sono gli errori più gravi in quanto derivanti dall'erroneo abbinamento di individui che in realtà sono diversi. Essi generano dati incoerenti nel tempo per le variabili di interesse e quindi dei flussi spuri che portano ad una sovrastima della mobilità effettiva.
2. Abbinamenti Mancati (falsi negativi): consistono nel mancato abbinamento delle interviste relative al medesimo individuo. Comportano una riduzione delle osservazioni campionarie, con conseguente minore efficienza nella stima dei parametri della popolazione e soprattutto dei flussi lordi tra due periodi. Possono essere molto gravi e possono produrre stime della mobilità potenzialmente distorte nel caso che la distribuzione degli abbinati rispetto ad alcuni caratteri strutturali risulti significativamente diversa dalla corrispondente distribuzione dei non abbinati

*Mancate risposte comunali, familiari e individuali*

Tale problema, noto come "attrition" o autoselezione dei rispondenti, genera una distorsione simile a quella dei falsi negativi. L'attrito riduce il numero dei casi disponibili con un conseguente aumento della varianza, inoltre genera una distorsione tanto più elevata quanto più il comportamento degli individui non rispondenti, o di quelli all'interno dei comuni che non partecipano ad entrambe le indagini, risulta significativamente diverso, secondo alcuni caratteri strutturali, rispetto a quello dei rispondenti.

*Errori di risposta o registrazione*

Possono avvenire per incomprensione della domanda, o della risposta, per errore di registrazione, o se l'intervistato mente consapevolmente. Anche in questo caso producono una distorsione. Nel campione longitudinale, infatti, in numero delle transizioni è molto più basso di quello delle permanenze. Un errore di risposta è molto più probabile che generi quindi una transizione non vera al posto di una permanenza. Di conseguenza tali errori molto probabilmente fanno aumentare i flussi. Una possibile fonte di tali errori è sicuramente la risposta proxi.

# 3 Il controllo e la correzione dei dati longitudinali della RTFL

Il controllo e la correzione dei dati longitudinali della RTFL avviene mediante un *piano di compatibilità* che agisce a livello di singola unità (*record*) per identificare ed imputare le mancate risposte parziali. Per mancate risposte parziali intendiamo sia risposte mancanti ad uno o più quesiti sia valori incoerenti tra una o più variabili rilevate[6].

L'identificazione degli errori avviene mediante un insieme di regole di incompatibilità, cioè un insieme di asserzioni sulla non ammissibilità di codici (modalità) per la singola variabile o di combinazioni di codici relativi a più variabili. Le regole vengono formulate dagli esperti dell'indagine e hanno lo scopo di individuare i valori fuori dominio, le mancate risposte ad uno o più quesiti e le incoerenze logiche tra variabili. Le regole sono di tipo *formale* o *sostanziale*. Sono regole del primo tipo quelle derivanti dalle norme di compilazione e dalla struttura del questionario (per esempio, se ha risposto NO al quesito 1, passare al quesito 3 altrimenti passare al quesito 2) e dal piano di registrazione su supporto informatico; appartengono al secondo tipo quelle derivanti da informazioni "a priori" sulla realtà indagata (per esempio, se l'età è minore di 18 non è possibile che la "condizione" sia militare di leva). Tuttavia, le regole di incompatibilità sono in grado di individuare solo i valori delle variabili che verificano le incoerenze logiche stabilite.

La correzione consiste nel modificare i valori risultati errati assegnandone di altri. Tale assegnazione può avvenire secondo due logiche diverse, l'una detta *deterministica*, l'altra *probabilistica*.

## 3.1 L'approccio deterministico

Un piano di compatibilità di tipo deterministico è generalmente costituito da un insieme di regole del tipo

SE [condizione di errore] ALLORA [azione di correzione].

La condizione di errore della regola esprime una relazione di incoerenza tra le variabili coinvolte; l'azione di correzione consiste nell'imputare un solo valore predeterminato, oppure un valore casualmente scelto da una distribuzione predeterminata e potrà riguardare o meno le variabili incluse nella parte "SE"; può dipendere, inoltre, dai valori assunti da altra o altre variabili. In quest'ultimo caso il metodo si basa su un albero decisionale che organizza le variabili rilevate secondo gerarchie prestabilite, facendo dipendere il valore delle variabili di livello più basso dalle combinazioni di valori di quelle di livello più alto.

Un approccio di questo tipo può portare a gravi inconvenienti. In primo luogo, il programma che traduce un piano di compatibilità deterministico è di tipo sequenziale: vengono prima controllate le condizioni che attiverebbero la prima regola e, in caso positivo, vengono eseguite le relative azioni; poi si procede alla verifica delle condizioni della seconda regola e così via. Poiché le regole agiscono in maniera sequenziale, il procedimento implica un ordinamento gerarchico tra esse. La scelta della gerarchia influenza i risultati dell'algoritmo; la compatibilità e la correzione della *i*-esima variabile sono, infatti, funzione dei valori assunti o modificati delle precedenti. Ne risulta che alla

[6] Le incoerenze logiche tra valori singolarmente ammissibili di variabili differenti, i valori mancanti e i valori fuori dominio possono essere assimilati alle mancate risposte parziali in quanto la logica di correzione adottata dipende dalle relazioni esistenti tra le variabili, in termini di regole di compatibilità, mentre non dipende dai valori che assumono le variabili stesse.

fine del controllo e correzione nulla assicura la correttezza di ogni singolo *record* (Barcaroli e Di Pace, 1991; Masselli e Barcaroli, 1994). Il motivo è che la correzione avviene senza considerare tutte le altre possibili situazioni di errore che tale correzione potrebbe attivare. In nessun caso, poi, un piano deterministico può garantire il minimo cambiamento, cioè che il numero di variabili modificate per riportare un *record* errato ad una situazione di correttezza sia il minimo possibile (Fellegi e Holt, 1976).

Non potendo applicare il principio del minimo cambiamento la scelta della gerarchia deve garantire dalla possibilità di errori indotti dalla procedura. Un criterio per la scelta della sequenza da adottare consiste nel minimizzare la probabilità di modificare un valore vero ordinando la sequenza in funzione delle probabilità di errore, o di ripristino del valore vero delle variabili coinvolte nelle diverse regole; in mancanza di tali informazioni le probabilità di errore delle variabili possono essere interpretate come pesi assegnati alle diverse variabili in funzione della loro importanza ai fini degli obiettivi dell'indagine; in caso di quesiti di salto si può scegliere quella sequenza di variabili per la quale sia massimo il rapporto tra la probabilità di avere osservato una data combinazione di codici, sotto l'ipotesi che la relativa sequenza sia vera, e la probabilità dello stesso evento sotto l'ipotesi che la sequenza sia falsa (Masselli, 1989 e 1990).

Dal punto di vista della progettazione e della realizzazione informatica, i programmi deterministici sono molto meno complessi di quelli probabilistici; essi, generalmente, sono costruiti ad hoc per la singola indagine, mentre i secondi, per la cui produzione sono necessarie risorse notevolmente maggiori che per i primi, sono programmi generalizzati, validi per più indagini differenti.

Molto schematicamente, possiamo ascrivere ai vantaggi del metodo deterministico (Barcaroli *et al*. 1993, Masselli e Barcaroli, 1994):

(i) la completa applicabilità: un piano deterministico è sempre applicabile ai dati una volta tradotte le regole di imputazione in istruzioni di programma;
(ii) l'efficienza elaborativa: il tempo necessario per eseguire il programma che traduce il piano deterministico è lineare rispetto al numero di regole di imputazione e al numero di *record*;
(iii) l'orientabilità degli effetti: lo statistico può orientare i risultati dell'applicazione del piano deterministico definendo opportunamente la parte imputazione di ogni regola e la sequenza di queste nel piano.

Quest'ultimo elemento è di una certa importanza: ad esempio, sulla base della fiducia che lo statistico nutre rispetto alla correttezza delle variabili, egli può implicitamente stabilire una gerarchia tra queste, orientando la modifica verso quelle che egli ritiene meno affidabili.

La correzione deterministica, tuttavia, è più adatta a trattare errori di tipo sistematico, al contrario dell'altra, che risulta più efficiente per quanto riguarda gli errori provenienti da un modello di generazione casuale .

## 3.2 L'approccio probabilistico

Al contrario di quello deterministico, un piano probabilistico non prevede la possibilità (o la necessità) di definire a priori, per ogni situazione di errore, l'elenco delle azioni da intraprendere per eliminare gli errori: l'esperto statistico deve limitarsi a definire le situazioni di errore (o di incompatibilità), demandando ad un prefissato algoritmo il compito di riportare il *record* ad una situazione di correttezza (o compatibilità). Le regole stabilite devono costituire un insieme coerente, cioè tale da garantire:

- la non ridondanza, ovvero la non ripetizione di regole già poste in altra forma o derivabili da altre;

- la non contraddittorietà tra regole.

Le regole ridondanti e quelle contraddittorie, infatti, inficiano le procedure di correzione basate sul principio del minimo cambiamento.

Una volta individuato l'errore, la correzione viene effettuata mediante metodi probabilistici di imputazione attribuendo alla variabile con valore errato o mancante un valore desunto dalla distribuzione dei valori esatti di tale variabile nelle altre unità statistiche. Ciò può essere fatto mediante criteri diversi[7]:

(1) mediante il metodo del "donatore", utilizzato principalmente per l'imputazione di variabili qualitative. Con tale metodo viene attribuito il valore che la variabile mancante o errata assume nell'unità statistica più "vicina" secondo una predefinita funzione di distanza;
(2) mediante la scelta di un modello (di solito, la scelta ricade sul modello di regressione lineare). Il valore della variabile mancante o da correggere viene attribuito utilizzando la funzione di regressione stimata su un insieme di unità statistiche esatte. Le variabili indipendenti di natura dicotomica possono essere introdotte nel modello come variabili *dummy*, mentre se la variabile dipendente è dicotomica, si può ricorrere ai modelli *logit* e *probit*.

Per la costruzione e l'esecuzione di piani probabilistici si fa comunemente riferimento alla metodologia e ai formalismi definiti da Fellegi e Holt (1976) dell'Istituto di Statistica Canadese. Pur prevedendo anche il trattamento delle variabili quantitative il metodo è stato sviluppato essenzialmente per quelle qualitative (codificate e non soggette ad alcuna metrica).

I vantaggi dell'approccio probabilistico sono speculari ai limiti di quello deterministico: la garanzia di correttezza finale dei *record*, la minimalità del cambiamento, il rispetto delle distribuzioni delle variabili. D'altra parte, la predisposizione di un piano di compatibilità probabilistico per indagini di media grande dimensione, quale, ad esempio, la RTFL, oltre a richiedere notevoli sforzi operativi e computazionali, costringe gli esperti a trovare soluzioni di compromesso per quegli algoritmi, che pur essendo validi da un punto di vista formale, si rivelano intrattabili dal punto di vista computazionale (Barcaroli e Di Pace, 1991).

### 3.2.1 La metodologia di Fellegi e Holt

La metodologia di Fellegi e Holt è una metodologia generalizzata, cioè valida per qualsiasi tipo d'indagine, che prevede, in sostanza, la definizione di due fasi. Nella prima fase occorre analizzare l'insieme delle regole esplicite, definite dall'esperto, al fine di produrre un insieme di regole completo, privo di regole ridondanti o contraddittorie, che contenga anche le regole implicite logicamente derivate dall'insieme delle regole esplicite. La costruzione dell'insieme delle regole implicite si dimostra indispensabile al fine di garantire la correttezza finale di un *record*. Infatti, pur essendo sufficienti ad individuare la presenza di errori all'interno dei *record*, le regole esplicite non sono sufficienti a determinare, senza ricorrere a cicli, quali variabili correggere per riportare il *record* ad una situazione globale di correttezza, e tanto meno il minor numero (*insieme minimale*) di variabili da modificare. Infatti, se le variabili da modificare venissero selezionate considerando solo le regole esplicite attivate, può avvenire che, una volta modificate tali variabili, queste siano disattivate e diventino invece attive altre regole, mantenendo quindi una situazione di errore nel *record*. Una volta ottenuto l'insieme completo delle regole si procede all'effettiva

[7] La letteratura esistente propone diversi e più ampi schemi classificatori dei metodi di imputazione (Kalton e Kasprzyk, 1982), ma in realtà corrispondono a casi particolari, o misture, dei due criteri di classificazione adottati.

correzione delle variabili errate mediante l'imputazione delle stesse. Quest'ultima fase è preceduta dall'identificazione delle variabili e modalità da correggere. Seguendo i criteri della metodologia di Fellegi e Holt questo passo è compiuto scegliendo l'*insieme minimale*, cioè il minor numero di variabili da modificare per riportare il *record* errato ad una situazione di correttezza. Nel caso in cui venissero individuati più insiemi minimali, si può scegliere l'insieme costituito da quelle variabili che, a priori, si ritiene siano maggiormente affette da errori, cioè meno attendibili. Ciò equivale ad assegnare, a priori, a ciascuna variabile un peso che ne rappresenti l'attendibilità e scegliere l'insieme minimale a cui corrisponde il minor prodotto dei pesi.

I metodi di imputazione proposti da Fellegi e Holt sono di tipo *hot-deck*: le modalità errate delle variabili appartenenti all'insieme minimale vengono sostituite con quelle di una o più unità "donatrici" (*record* donatori) scelte tra le unità del campione nelle quali non è stato riscontrato alcun errore (*record* esatti).

Le strategie di imputazione proposte sono:

1. imputazione *sequenziale*, che consiste nell'imputare separatamente ogni singola variabile dell'insieme minimale;
2. imputazione *congiunta* in cui le variabili dell'insieme minimale vengono imputate congiuntamente.

Il metodo sequenziale presenta degli inconvenienti. Il più grave è dovuto proprio all'imputazione sequenziale delle variabili, che assicura il solo mantenimento delle distribuzioni marginali nell'insieme dei dati esatti (eccetto il caso di indipendenza, in cui verrebbero ovviamente preservate anche le distribuzioni congiunte). Inoltre, possono essere necessari tanti donatori quante sono le variabili dell'insieme minimale. Il metodo di imputazione congiunta, invece, imputando congiuntamente tutte le variabili dell'insieme minimale preserva nell'insieme dei dati esatti sia le distribuzioni marginali sia le distribuzioni congiunte. Tale metodo, che denominiamo più propriamente imputazione *congiunta allargata*, può essere modificato restringendo la ricerca del donatore a quei *record* che nei campi corrispondenti alle variabili *matching* (variabili che non necessitano di essere imputate) abbiano esattamente gli stessi valori che tali variabili assumono nel *record* da imputare. Parleremo, in tal caso, di imputazione *congiunta ristretta*.

Per entrambi i metodi, qualora non fosse possibile individuare un donatore, si può ricorrere all'imputazione forzata di ciascuna delle variabili dell'insieme minimale. Tale metodo consiste nell'imputare, alla variabile in esame, un qualunque valore estratto casualmente dalla distribuzione marginale semplice della variabile stessa nell'insieme dei dati esatti.

Entrambi i metodi di imputazione, sequenziale e congiunta, possono essere migliorati considerando potenziali donatori quei *record* il cui valore in uno o più campi, scelti a priori, è identico (o simile) a quello/i del *record* in esame, facendo uso, in tal modo, anche dell'informazione proveniente dal *record* da imputare.

## 3.3 L'imputazione di dati longitudinali

L'imputazione di un *record* longitudinale, in cui figurano i valori delle variabili rilevate in differenti occasioni d'indagine su una stessa unità, può essere effettuata secondo due diverse strategie: imputare tutte le variabili che lo richiedono oppure limitare l'imputazione soltanto alle variabili di una data occasione, mantenendo inalterate quelle delle occasioni precedenti. Quest'ultima, sebbene ponga ulteriori restrizioni alla metodologia di imputazione (vincoli di "fissità" sulle variabili non imputabili), si rivela più praticabile laddove sia necessario produrre, per ogni occasione d'indagine, risultati definitivi. Seguendo l'altra strategia, infatti, i dati rilevati sarebbero considerati sempre provvisori sino all'ultima occasione d'indagine.

Più precisamente, considerando due sole occasioni d'indagine, l'una al tempo $t_1$ e l'altra al tempo $t_2$, la modifica di una o più variabili del *record* longitudinale, necessaria per riportare il *record* stesso ad una situazione di correttezza, può riguardare sia le variabili osservate al tempo $t_1$, $X^i_{t1}$, sia quelle osservate al tempo $t_2$, $X^j_{t2}$. Per un'indagine come la RTFL, progettata per ottenere, a cadenza trimestrale, risultati sulla misura e sulle caratteristiche dell'occupazione e della disoccupazione, è ragionevole assumere che i dati osservati al tempo $t_1$ non subiscano variazioni. Ciò significa che la correzione longitudinale tra le rilevazioni di gennaio 2001 ed aprile 2001, per esempio, non modifica affatto le variabili di gennaio 2001. Analogamente, la correzione tra aprile 2001 e luglio 2001 è tale da lasciare immutate le variabili di aprile 2001. Tale ipotesi, che riflette l'esigenza di avere risultati definitivi ad ogni occasione d'indagine, è stata accolta nella definizione della procedura di correzione longitudinale.

## 3.4 La procedura di controllo e correzione longitudinale

La procedura di correzione dei dati longitudinali della RTFL è caratterizzata, sostanzialmente, da un modulo probabilistico ed uno deterministico.

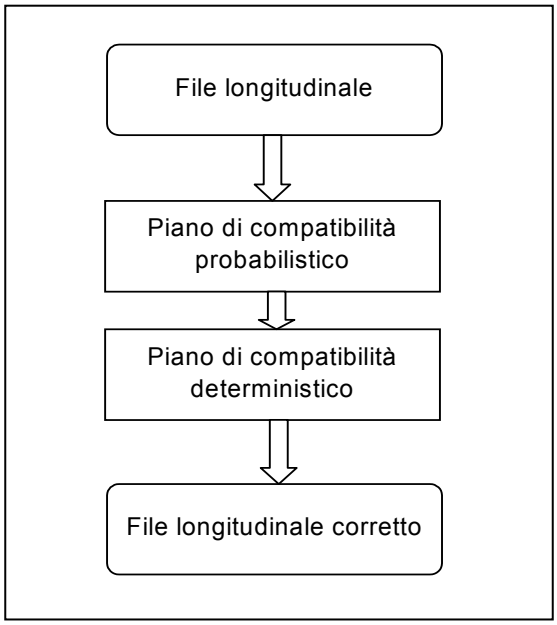
La correzione probabilistica si basa interamente sulla metodologia di Fellegi e Holt ed è implementata in SCIA (**S**istema di **C**ontrollo e **I**mputazione **A**utomatica), un sistema di correzione dei dati realizzato dal Dipartimento d'Informatica dell'ISTAT (Riccini Margarucci *et al*., 2000). In particolare, è stata utilizzata la versione di SCIA contenuta in CONCORD (**CON**trollo e **COR**rezioni **D**ati), un sistema progettato per ambienti operativi Windows (Riccini Margarucci e Floris, 2000). La correzione deterministica, invece, consiste in una procedura SAS in cui sono definite le regole deterministiche relative soprattutto a variabili quantitative (per esempio, durata della ricerca di lavoro, durata dell'attuale attività lavorativa, ecc.). Si tratta di informazioni, rilevate mediante quesiti retrospettivi, che risentono maggiormente degli errori connessi al processo di memoria e che si prestano, quindi, ad essere corrette con il metodo deterministico.

Le regole di incompatibilità longitudinali sono regole che indicano la presenza di incoerenze tra le modalità di due o più variabili rilevate in due diverse occasioni d'indagine su uno stesso individuo. Per l'esatta costruzione di tali regole il confronto tra variabili deve svolgersi in entrambe le direzioni: $t_1 \rightarrow t_2$ e $t_1 \leftarrow t_2$; poiché, in molti casi, ambedue i controlli sono necessari e non ridondanti. Consideriamo, ad esempio, le seguenti regole longitudinali, che rappresentano un errore di incompatibilità logica rispetto al modello di rilevazione della RTFL:

$$SESSO_{t1}(j) \in \{\text{femmina}\} \cap (CONDIZIONE_{t2}(j) \in \{\text{militare di leva}\})$$
$$SESSO_{t2}(j) \in \{\text{femmina}\} \cap (CONDIZIONE_{t1}(j) \in \{\text{ militare di leva }\})$$

pur sembrando a prima vista identiche, in realtà esprimono due incoerenze logicamente distinte, da una parte l'incoerenza tra le variabili "sesso" al tempo $t_1$ e "condizione" al tempo $t_2$, dall'altra l'incoerenza tra "sesso" al tempo $t_2$ e "condizione" al tempo $t_1$. Si tratta, quindi, di due regole in cui sono coinvolte quattro variabili aventi a due a due lo stesso dominio di definizione. Perciò in corrispondenza dell'individuo j può verificarsi al più soltanto una tra le due regole sopra esplicitate.

**Figura 3.1** - Fasi principali della procedura di controllo e correzione dei dati longitudinali della RTFL




La Figura 3.1 riassume le fasi principali della procedura di controllo e correzione dei dati longitudinali della RTFL. Una volta effettuato l'abbinamento tra il *file* del tempo $t_1$ e quello del tempo $t_2$, entrambi corretti trasversalmente, il *file* longitudinale viene corretto prima dal piano di compatibilità probabilistico e successivamente da quello deterministico[8].

La procedura di correzione ha lo scopo principale di garantire la correttezza longitudinale delle informazioni e di conservarne, in secondo luogo, quella trasversale. A seguito della correzione longitudinale, infatti, alcune informazioni potrebbero risultare "errate" dal punto di vista trasversale. Pertanto, per non introdurre ulteriori distorsioni nei dati, sono state inserite alcune regole di incompatibilità trasversale nel piano probabilistico, mentre nel piano deterministico le imputazioni sono state effettuate in modo tale da tener conto anche della coerenza trasversale. D'altra parte, sarebbe stato inopportuno in questa sede progettare un piano di compatibilità che potesse agire contemporaneamente sia sugli "errori" trasversali sia su quelli longitudinali del *file* longitudinale originale (cioè non corretto in alcun modo), in quanto l'obiettivo di produrre dati longitudinali è di gran lunga posteriore alla serie di dati trasversali già pubblicati. In altri termini, era opportuno che i dati prodotti fossero congruenti con quelli trasversali già noti.

Il problema del trattamento congiunto degli "errori" trasversali e longitudinali può essere risolto aggiungendo regole di incompatibilità longitudinali all'insieme di regole trasversali della RTFL. Tale metodo, che si rivela più adatto per il funzionamento a regime della procedura di correzione della RTFL, dimostra di essere uno strumento flessibile nonché ampiamente diffuso che può essere utilizzato per correggere le mancate risposte parziali, trasversali e longitudinali, delle indagini su larga scala (Rosati, 2000).

[8] A titolo esemplificativo riportiamo alcuni indicatori delle prestazioni della procedura di correzione. Il numero di *record* errati nel piano probabilistico è pari al 5% del numero totale dei *record*. Al termine dell'intera procedura di correzione, risulta che l'82% dei *record* ha subito al più una correzione, mentre il 96% ne ha subite, al massimo, due. Il numero di variabili coinvolte nella procedura sono 62 per la parte probabilistica e 18 di quella deterministica (alcune di queste fanno parte anche del piano probabilistico).

# 4 I coefficienti di riporto alla popolazione longitudinale

La metodologia di riporto all'universo dei dati longitudinali è simile a quella correntemente utilizzata dall'ISTAT per il riporto dei dati trasversali della RTFL. Tale procedura si basa sull'uso degli stimatori di ponderazione vincolata.

Il campione longitudinale può produrre stime di flusso tra due periodi ma anche stime si stock sulla popolazione longitudinale. Tali stime (sulla popolazione longitudinale) devono essere in qualche modo congruenti con le stime trasversali riferite a tutta la popolazione iniziale e finale (Schema 1.2) .

La procedura di riporto all'universo per i dati longitudinali RTFL implementata all'Istat considera che il campione longitudinale sia rappresentativo della sola popolazione longitudinale. La metodologia di calcolo dei pesi assicura (per la maggior parte deli aggregati di interesse a livello ripartizionale e regionale) la coerenza dei dati delle matrici di transizione con le stime trasversali.

## 4.1 La procedura per la costruzione dei coefficienti di riporto all'universo.

La metodologia di riporto all'universo dei dati longitudinali è simile a quella correntemente utilizzata dall'ISTAT per il riporto dei dati trasversali della RTFL. Tale procedura si basa sull'uso degli stimatori di ponderazione vincolata. I pesi sul campione longitudinale sono calcolati a livello individuale[9] in quanto la ridotta numerosità campionaria unita all'elevato numero di vincoli necessari a far rispettare le congruenze tra stime longitudinali e *cross-section* trimestrali, aumenterebbe considerevolmente la variabilità dei pesi.

La procedura di riporto all'universo per i dati longitudinali RTFL implementata all'Istat considera che il campione longitudinale sia rappresentativo della sola popolazione longitudinale. Il campione longitudinale, comunque, può produrre oltre alle stime di flusso tra due periodi, anche stime si stock sulla popolazione longitudinale. Tali stime devono essere in qualche modo congruenti con le (e comunque non superiori delle) stime trasversali riferite a tutta la popolazione iniziale e finale (Schema 1.2). La metodologia di calcolo dei pesi assicura (per la maggior parte deli aggregati di interesse a livello di ripartizione e regione) la coerenza dei dati delle matrici di transizione con le stime trasversali.

Come già evidenziato nel capitolo 1, le due sezioni abbinabili ad inizio periodo possono essere considerate rappresentative di tutta la popolazione trasversale di quel trimestre in quanto si tratta, teoricamente, di un campione casuale del campione completo. A causa della caduta di parte dei comuni campione, gli individui abbinabili potrebbero risultare un campione distorto rispetto al campione completo.

La procedura di riporto si sviluppa in diversi passi che, mediante tecniche di post-stratificazione successive, tendono a ridurre le possibili distorsioni. Il peso finale si ottiene a partire da un peso base per gli individui abbinabili a cui si applica un fattore correttivo della distorsione dovuta alla "mancata risposta comunale". Si ottiene un peso iniziale che, solo per gli individui abbinati, viene successivamente corretto per tenere parzialmente conto della "mancata risposta familiare/individuale" dovuta ai non abbinati.

[9] I pesi sul campione trasversale sono calcolati a livello familiare per l'ottenimento del peso unico familiare/individuale.

### 4.1.1 Calcolo del peso base sugli individui abbinabili

Il peso base è lo stesso che è stato calcolato per le stime di stock trasversali ed è ottenuto come reciproco della probabilità di inclusione della famiglia nel campione. Esso è calcolato nell'usuale modo,

$$K_{kij} = \frac{P_h M_{hi}}{P_{hi} m_{hi} n_h} \tag{4.1}$$

in cui :

$P_h$ rappresenta la popolazione dello strato h;

$M_{hi}$ il numero delle famiglie del comune i dello strato h;

$P_{hi}$ la popolazione del comune i dello strato h;

$m_{hi}$ la numerosità campionaria delle famiglie in i di h;

$n_h$ quella dei comuni in h.

Considerando solo gli individui abbinabili, si calcola un fattore correttivo per tenere conto della mancata risposta comunale ottenendo il peso base longitudinale. E' noto che ad ogni rilevazione alcuni comuni, pur facendo parte del campione, non partecipano all'indagine per diverse ragioni. E' noto anche che in occasione dell'indagine di aprile di ciascun anno vengono sostituiti alcuni comuni che hanno esaurito le famiglie campione e/o che non vogliono/possono partecipare più all'indagine.

Nel campione longitudinale quindi può verificarsi che manchino alcuni comuni o interi strati. Si è scelto quindi di calcolare un fattore correttivo a livello regionale che riattribuisce il peso della popolazione dei comuni mancanti su quella dei comuni presenti. In una prima fase si era sperimentato il metodo del "collassamento" degli strati mancanti a strati presenti "vicini". Questo metodo, che portava però ad un aumento eccessivo (in alcuni casi notevole) della variabilità dei pesi base a livello regionale, è stato abbandonato.

### 4.1.2 Calcolo del peso iniziale sugli individui abbinabili

Dal peso base longitudinale, quindi, abbiamo calcolato un peso iniziale sugli abbinabili imponendo di ottenere esattamente le stesse stime trasversali fornite dal campione completo del primo trimestre e già pubblicate dall'Istat per alcuni grandi aggregati (popolazione regionale per sesso e classe di età, condizione professionale, settore di attività, posizione nella professione, ricerca di lavoro, durata della ricerca). Tale correzione si è resa necessaria per correggere l'ulteriore distorsione, rispetto al campione completo trasversale.

Lo Schema 4.1 riassume i vincoli utilizzati per il calcolo del peso iniziale sugli individui abbinabili.

**Schema 4.1 Vincoli utilizzati per il calcolo del peso iniziale sugli individui abbinabili**

| Vincoli a livello di ripartizione (Nord est, Nord ovest, Centro e Mezzogiorno) | | |
|---|---|---|
| Popolazione per classi di età quinquennali<br>15-19, 20-24, …, 65-69, 70-74, 75 e oltre | * | Maschi e Femmine |
| occupati autonomi a tempo pieno;<br>occupati autonomi a tempo parziale;<br>occupati dipendenti - indeterminato - tempo pieno;<br>occupati dipendenti - indeterminato - tempo parziale;<br>occupati dipendenti - a termine - tempo pieno;<br>occupati dipendenti - a termine - tempo parziale<br>persone in cerca di occupazione da meno di 6 mesi;<br>persone in cerca di occupazione da 6 a 11 mesi;<br>persone in cerca di occupazione da 12 mesi o più;<br>nfl potenziali;<br>nfl non disponibile e/o non in cerca;<br>nfl 65 anni e più;<br><br>occupati in agricoltura<br>occupati nell'industria<br>occupati nell'industria in senso stretto<br>occupati nel commercio, alberghi e ristoranti<br>occupati negli altri servizi<br><br>popolazione con diploma di laurea breve, laurea, dottorato<br>popolazione con qualifica professionale,diploma di scuola superiore<br>popolazione con scuola media inferiore | * | Maschi e Femmine |

**Schema 4.1 (segue) Vincoli utilizzati per il calcolo del peso iniziale sugli individui abbinabili**

| Vincoli a livello di singola regione | | |
|---|---|---|
| Popolazione per le seguenti classi di età<br>15-24, 25-44, 45-64, 65 e più | * | Maschi e Femmine |
| Occupati<br>Disoccupati<br>Non forze di lavoro | * | Maschi e Femmine |
| occupati in agricoltura<br>occupati nell'industria<br>occupati nel commercio, alberghi e ristoranti<br>occupati negli altri servizi<br><br>popolazione con diploma di laurea breve, laurea, dottorato<br>popolazione con qualifica professionale,diploma di scuola superiore<br>popolazione con scuola media inferiore<br><br>occupati autonomi<br>occupati dipendenti a tempo indeterminato<br>occupati dipendenti a termine | * | Totale |

### 4.1.3 Calcolo del peso finale sugli abbinati

A partire dal peso iniziale si calcola il peso finale solo sugli individui "abbinati" imponendo come totale noto la popolazione longitudinale per regione, sesso e 6 classi di età (15-24, 25-34, …, 65e +). Questa procedura riesce ad attenuare gli effetti distorsivi dovuti al mancato abbinamento degli eleggibili (individui usciti dalla famiglia campione, non rispondenti, non rintracciabili).Una proprietà importante di questo ultimo peso è che, all'interno del campione longitudinale, tiene conto (e corregge) di quella parte della mancata risposta familiare ed individuale correlata alle variabili di post-stratificazione (ipotesi di indipendenza della mancata risposta condizionata alla combinazione di regione, sesso e età).

Lo Schema 4.2 riassume i vincoli utilizzati per il calcolo del peso finale sugli individui abbinati.

**Schema 4.2 Vincoli utilizzati per il calcolo del peso finale sugli abbinati**

| **Vincoli sulla popolazione longitudinale** | | |
|---|---|---|
| Popolazione per<br>ripartizione geografica (nord est, nord ovest, centro e mezzogiorno)<br>e classi di età decennali<br>15-24, 25-34 , 35-44, 45-54 55-64, 65 e più | * | Maschi e<br>Femmine |
| Popolazione regionale per le seguenti classi di età<br>15-24, 25-44, 45-64, 65 e più | * | Maschi e<br>Femmine |

Il vantaggio di questa architettura è che otteniamo sicuramente stime affidabili per la popolazione longitudinale e, per differenza con le stime di stock correntemente pubblicate trimestralmente dalla RTFL, riusciamo ad ottenere stime sulla condizione ad inizio periodo degli usciti (morti + cancellati) dalla popolazione e sulla condizione a fine periodo degli entrati nella popolazione (15enni e iscritti). Queste ultime, si possono scomporre ulteriormente distinguendo le stime relative agli individui usciti per morte da quelle relative agli individui cancellati dall'anagrafe, e distinguendo quelle relative alla popolazione di 15 anni a fine periodo da quelle relative agli iscritti. In particolare, le stime della condizione a inizio periodo degli usciti per morte si ottengono applicando i relativi tassi di mortalità (per sesso, regione e classi di età quinquennali) al campione trasversale del primo trimestre. Le stime sulla condizione a fine periodo dei 15enni prendendo si ottengono direttamente dal campione completo del secondo trimestre.

**ALLEGATO**

# CODICI DEI TITOLI DI STUDIO

## CORSI DI LAUREA

**001 GRUPPO SCIENTIFICO**
Matematica
Fisica
Astronomia
Scienza dei materiali
Discipline nautiche
Informatica
Scienze dell'informazione

**002 GRUPPO CHIMICO-FARMACEUTICO**
Chimica
Chimica industriale
Farmacia
Chimica e tecnologia farmaceutica
Biotecnologie farmaceutiche

**003 GRUPPO GEO-BIOLOGICO**
Scienze geologiche
Scienze naturali
Scienze biologiche
Scienze ambientali
Biotecnologie (vari indirizzi)

**004 GRUPPO MEDICO**
Medicina e Chirurgia
Odontoiatria e protesi dentaria

**005 GRUPPO INGEGNERIA**
Biennio Propedeutico
Ingegneria Mineraria
Ingegneria Meccanica
Ingegneria Elettrotecnica
Ingegneria Elettronica
Ingegneria Nucleare
Ingegneria Chimica
Ingegneria Navale e Meccanica
Ingegneria Aerospaziale (e aeronautica)
Ingegneria Civile
Ingegneria e Tecnologie Industriali
Ingegneria Civile Difesa suolo e pianificazione
Ingegneria Forestale
Ingegneria dei Materiali
Ingegneria Informatica
Ingegneria Elettrica
Ingegneria delle Telecomunicazioni
Ingegneria Gestionale
Ingegneria per Ambiente e Territorio
Ingegneria Edile
Ingegneria Navale
Ingegneria Biomedica

**006 GRUPPO ARCHITETTURA**
Architettura
Urbanistica
Disegno industriale
Pianificazione territoriale e urbanistica
Storia e conservazione dei beni architettonici e ambientali

**012 GRUPPO LINGUISTICO**
Lingue e letterature straniere moderne
Lingue e letterature straniere
Lingue e civiltà orientali
Lingue e letterature orientali
Traduzione e interpretazione
Interprete
Traduttore

**007 GRUPPO AGRARIO**
Scienze Agrarie
Scienze Forestali
Scienze Forestali e Ambientali
Medicina Veterinaria
Scienze della produzione animale
Scienze delle preparazioni alimentari
Scienze Agrarie tropicali e sub-tropicali
Scienze e Tecnologie Alimentari
Scienze e Tecnologie Agrarie
Biotecnologie agro-industriali

**008 GRUPPO ECONOMICO-STATISTICO**
Economia (immatricolati comuni a più corsi)
Economia e commercio
Scienze economiche
Scienze economiche e bancarie
Scienze statistiche e demografiche
Scienze statistiche demografiche e sociali
Scienze statistiche e attuariali
Scienze statistiche ed economiche
Economia aziendale
Economia bancaria
Economia politica
Economia delle istituzioni e mercati finanziari
Economia amministrazioni pubbliche e istituzioni internazionali
Economia e legislazione per l'impresa
Economia del turismo
Statistica e informatica per l'azienda
Scienze economiche e sociali
Discipline economiche e sociali
Commercio internazionale e mercati valutari
Economia marittima e dei trasporti
Economia bancaria, finanziaria e assicurativa
Economia ambientale
Economia assicurativa e previdenziale

**009 GRUPPO POLITICO-SOCIALE**
Scienze politiche
Sociologia
Scienze internazionali e diplomatiche
Scienze della comunicazione
Relazioni pubbliche

**010 GRUPPO GIURIDICO**
Giurisprudenza
Scienze dell'amministrazione

**011 GRUPPO LETTERARIO**
Lettere
Materie letterarie
Filosofia
Geografia
Discipline arti, musica e spettacolo
Storia
Conservazione beni culturali
Studi islamici
Filologia e storia dell'Europa orientale
Musicologia
Teologia

**013 GRUPPO INSEGNAMENTO**
Pedagogia
Scienze dell'educazione

**014 GRUPPO PSICOLOGICO**
Psicologia

## CORSI DI DIPLOMA UNIVERSITARIO E SCUOLE DIRETTE AI FINI SPECIALI

**101 GRUPPO SCIENTIFICO**
Matematica
Metodologie fisiche
Scienza dei materiali
Informatica

**102 GRUPPO CHIMICO-FARMACEUTICO**
Chimica
Scienze e tecniche cartarie
Tecniche erboristiche
Controllo qualità nel settore industriale farmaceutico
Informazione scientifica sul farmaco
Tecnologie farmaceutiche

**103 GRUPPO GEO-BIOLOGICO**

**104 GRUPPO MEDICO**
Dietologia e dietetica applicata
Fisioterapista e terapista nella riabilitazione
Igienista dentale
Logopedia
Ortottista e assistente in oftalmologia
Ostetricia
Podologo
Riabilitazione psichiatrica e psico-sociale
Scienze infermieristiche
Tecnico in tecnologie mediche
Tecnico di neurofisiopatologia
Tecnico di audiometria e audioprotesi
Tecnico di laboratorio biomedico
Tecnico di radiologia medica
Terapia riabilitazione neuro e psicomotricità età evolutiva
Terapista della riabilitazione

**105 GRUPPO INGEGNERIA**
Edilizia
Ingegneria aerospaziale
Ingegneria biomedica
Ingegneria chimica
Ingegneria dell'ambiente e delle risorse
Ingegneria dell'automazione
Ingegneria delle infrastrutture
Ingegneria delle telecomunicazioni
Ingegneria delle telecomunicazioni (a distanza)
Ingegneria elettrica
Ingegneria elettrica (a distanza)
Ingegneria elettronica
Ingegneria elettronica (a distanza)
Ingegneria energetica
Ingegneria informatica
Ingegneria informatica e automatica
Ingegneria informatica e automatica (a distanza)
Ingegneria logistica e della produzione
Ingegneria meccanica
Ingegneria meccanica (a distanza)
Sistemi informativi territoriali

**106 GRUPPO ARCHITETTURA**
Disegno industriale
Tecniche e arti della stampa

**107 GRUPPO AGRARIO**
Gestione tecnica e amministrativa in agricoltura
Igiene e sanità animale
Produzione agrarie tropicali e subtropicali
Produzioni animali
Produzioni vegetali
Tecniche forestali
Tecniche forestali e tecnologie del legno
Tecnologie alimentari
Biotecnologie agro-industriali

**108 GRUPPO ECONOMICO-STATISTICO**
Amministrazione aziendale
Commercio estero
Economia delle amministrazioni pubbliche
Economia imprese cooperative e organizzazione no profit
Economia e amministrazione delle imprese
Economia e amministrazione delle imprese alimentari
Economia e gestione dei servizi turistici
Gestione delle amministrazioni pubbliche
Gestione delle imprese alimentari
Marketing e comunicazione d'azienda
Statistica
Statistica e informatica per la gestione delle imprese
Statistica e informatica per le amministrazioni pubbliche

**109 GRUPPO POLITICO-SOCIALE**
Altri corsi (gruppo politico-sociale)
Giornalismo
Servizio sociale
Tecnica pubblicitaria

**110 GRUPPO GIURIDICO**
Consulente del lavoro
Operatore giudiziario
Operatore giuridico d'impresa
Operatore della pubblica amministrazione
Relazioni industriali

**111 GRUPPO LETTERARIO**
Archivisti paleografi
Bibliotecari
Conservatori di manoscritti
Operatori/conservatori dei beni culturali
Operatore di costume e moda
Paleografia e filologia musicale
Storia e didattica della musica

**112 GRUPPO LINGUISTICO**
Traduttore e interprete
Traduttore, interprete e corrispondente in lingue estere

**113 GRUPPO INSEGNAMENTO**
Insegnamento della lingua italiana a stranieri
Abilitazione alla vigilanza nelle scuole elementari

**114 GRUPPO PSICOLOGICO**

**115 GRUPPO EDUCAZIONE FISICA**
Educazione fisica

## DIPLOMA POST-SECONDARIO NON UNIVERSITARIO

199 Accademia delle Belle Arti

199 Altro diploma post-secondario

### DIPLOMA DI MATURITÀ (CORSO DI SCUOLA SECONDARIA SUPERIORE DI 4-5 ANNI )

**MATURITÀ DI ISTITUTO PROFESSIONALE**

211 Agrario
212 Industria e Artigianato
213 Marinaro
214 Per i servizi commerciali, turistici e pubblicità
215 Per i servizi alberghieri e ristorazione
216 Per i servizi sociali
217 Altro istituto professionale

**MATURITÀ DI ISTITUTO TECNICO**

220 Agrario
221 Industria e Artigianato
222 Nautico
223 Aeronautico
224 Commerciale
225 Per geometri
226 Per il turismo
227 Per periti aziendali
228 Femminile

**MATURITÀ DI LICEO**

241 Scientifico
242 Classico
243 Linguistico
252 Artistico

**ALTRO TIPO DI ISTITUTO**

231 Scuola magistrale (ciclo lungo, se ciclo breve cod. 331)
232 Istituto magistrale
251 Istituto d'arte (II ciclo, se ciclo I cod. 321)

### DIPLOMA DI QUALIFICA PROFESSIONALE (CORSO DI SCUOLA SECONDARIA SUPERIORE DI 2-3 ANNI CH NON PERMETTE L'ACCESSO ALL'UNIVERSITÀ)

**QUALIFICA DI ISTITUTO PROFESSIONALE**

311 Agrario
312 Industria e Artigianato
313 Marinaro
314 Per i servizi commerciali, turistici e pubblicità
315 Per i servizi alberghieri e ristorazione
316 Per i servizi sociali
319 Altro istituto professionale

**ALTRI DIPLOMI**

341 Licenza di conservatorio
341 Istituto di musica
341 Accademia di danza

321 **QUALIFICA DI ISTITUTO D'ARTE**
(I ciclo, se ciclo II cod. 251)

331 **LICENZA DI SCUOLA MAGISTRALE**
(ciclo breve, se ciclo lungo cod. 231)

399 **ALTRA QUALIFICA O LICENZA**

## Riferimenti bibliografici


Barcaroli G. e Di Pace L. (1991), Un sistema intelligente di supporto alla correzione di dati statistici. Note di Informatica, Ricerca Tecnologia Applicazione, n.26, IBM.

Belin T.R., Rubin D.B. (1995). *A method for calibrating false-match rate in record linkage*. Journal of the American Statistical Association, 90, 694-707.

Centra M., Discenza A.R., Rustichelli E. (2001), Strumenti per le analisi di flusso nel mercato del lavoro. Una procedura per la ricostruzione della struttura longitudinale della Rilevazione trimestrale Istat sulle forze di lavoro. Monografie sul mercato del lavoro e le politiche per l'impiego. ISFOL.

Copas J.B., Hilton F.J. (1990). *Record linkage: statistical models for matching computer records*. Journal of the Royal Statistical Society A, 153, 3, 287-320.

Dempster A.P., Laird N.H., Rubin D.B. (1977). *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society B, 39, 1-38.

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

Falorsi S., Rinaldelli C. (1998), *Uso di un software generalizzato per il calcolo delle stime e degli errori di campionamento,* Statistica Applicata, Vol. 10, n.2.

Fellegi I.P. e Holt D. (1976), A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, 71, 17-35.

Fellegi I.P., Sunter A.B. (1969). *A Theory for record linkage*. Journal of the American Statistical Association, 64, 1183-1210.

Giusti A., Marliani G., Torelli N. (1991). *Procedure per l'abbinamento dei dati individuali delle forze di lavoro*. In Trivellato, U. (a cura di), *Forze di Lavoro: Disegno dell'Indagine e Analisi Strutturali*. ISTAT, Annali di Statistica, 9, 11.

Istat (2000), Popolazione per sesso, età e stato civile nelle province e nei grandi comuni. Stime regionali al 1.1.2000 - Anno 1999, Collana Informazioni n 55.

Jaro M.A. (1989). *Advances in record linkage methodology as applied to matching the 1985 census of Tampa, Florida*. Journal of the American Statistical Association, 89, 414-420.

Kelley R.P. (1985). *Advances in record linkage methodology: a method for determining the best blocking strategy*. In Kills B. e Alvey W. (eds.), *Record Linkage Techniques-1985, Proceedings of the Workshop on Exact Matching Methodologies*, Statistics of Income Division, U.S. Internal Revenue Service, 1299, 2-86.

Kirkendall N.J. (1985). *Weights in computer matching: applications and an information theoretic point of view*. In Kills B. e Alvey W. (eds.), *Record Linkage Techniques-1985, Proceedings of the Workshop on Exact Matching Methodologies*, Statistics of Income Division, U.S. Internal Revenue Service, 1299, 2-86.

Masselli M. (1989). *Manuale di tecniche di indagine – il sistema di controllo della qualità dei dati*. Vol. 6, Note e Relazioni, Istat.

Masselli M. (1990), *Un Modello per l'Individuazione della Sequenza di Regole e Variabili in un Piano di Compatibilità di Tipo Deterministico*, ISTAT (documento interno).

Masselli M. e Barcaroli G. (1994), *La Revisione dei Dati nelle Indagini Statistiche*. In: Colombo B., Cortese A., Fabbris L. (a cura di), La Produzione di Statistiche Ufficiali, C.L.E.U.P., Padova

Moriani C. (1981). *Forze di lavoro e flussi di popolazione*. Supplemento al Bollettino Mensile di Statistica, Istat, 15, 5-15.

Newcombe H.B., Kennedy J.M., Axford S.J., James A.P. (1959). *Automatic linkage of vital records*. Science, 130, 954-959.

Paggiaro A., Torelli N. (1999). *Una procedura per l'abbinamento di record nella rilevazione trimestrale delle forze di lavoro*, working paper n. 15, ottobre 1999, progetto di ricerca MURST "Lavoro e disoccupazione: questioni di misura e di analisi", CLEUP, Padova.

Riccini Margarucci E. e Floris P. (2000), *Controllo e correzione dati* - Manuale utente. ISTAT – Dipartimento d'Informatica Gruppo Software Generalizzato.

Riccini Margarucci E., Silvestri F., e Floris, P. (2000), *S.C.I.A. - Sistema di Controllo e Imputazione Automatica* - Manuale utente. ISTAT – Dipartimento d'Informatica Gruppo Software Generalizzato.

Rosati S. (2000), *La correzione di dati longitudinali nell'indagine Forze di lavoro*. Rivista di Statistica Ufficiale – Quaderni di Ricerca, n. 3, ISTAT.

Tepping B.J. (1968). *A model for optimum linkage of records*. Journal of the American Statistical Association, 63, 1321-1332.

Thibaudeau Y. (1993). *The discrimination power of dependency structures in record linkage*. Survey Methodology, 19, 31-38.

Torelli N. (1998). *Integrazione di dati mediante tecniche di abbinamento esatto: sviluppi metodologici e aspetti applicativi*. Atti della XXXIX riunione scientifica SIS.

Verma V., (1995), *Weighting for Wave 1.* Documento EUROSTAT, doc. PAN 36/95.

Winkler W.E. (1995). *Matching and record linkage*. In Cox, B.G. et al. (eds.), Business Survey Methods, New York, J. Wiley.