Progetto interdipartimentale
"Informazione statistica territoriale e settoriale per le politiche strutturali 2001-2008"

QCS OBIETTIVO 1 (2000-2006)
PON ASSISTENZA TECNICA E AZIONI DI SISTEMA
misura I.3

. . . . . .

Metodologia utilizzata per le stime sull'occupazione residente e le persone in cerca di occupazione nei Sistemi Locali del Lavoro per gli anni 1998-2000

. . . . . .

Roma, ottobre 2002

INDICE[1]

[1] Il presente rapporto è il frutto dell'attività congiunta di un Gruppo di lavoro dell'Istat costituitosi ad hoc per la realizzazione di stime sull'occupazione residente e la disoccupazione coordinato da Sandro Cruciani e composto da: Alessandro Faramondi dell'U.O. OBS/E della Direzione Centrale della Contabilità Nazionale, Stefano Falorsi, Loredana Di Consiglio, Fabrizio Solari e Francesco Paolo Rizzo dell'U.O. MPS/A del Servizio della metodologia di base per la produzione statistica, Antonio Rinaldo Discenza e Silvia Loriga dell'U.O. FOL/A del Servizio formazione e lavoro.

1. PREMESSA

L'Istat, dalla seconda metà del 1999, è attivamente impegnato nella produzione di informazione statistica con dettaglio territoriale adeguato alla programmazione economica. Tale impegno si inquadra in due progetti, dove il secondo di questi è la naturale prosecuzione del primo, a valere sul PON Assistenza tecnica dei due cicli di programmazione dei Fondi Strutturali 1994-99 e 2000-06.

Il progetto afferente al ciclo 2000-06, denominato "Informazione statistica territoriale e settoriale per le politiche strutturali 2000-06", risponde a molteplici esigenze informative espresse dal Ministero dell'Economia e delle Finanze ai fini del corretto impiego delle risorse comunitarie e per poter svolgere un'attiva programmazione nello sviluppo del territorio.

Una delle attività maggiormente innovative previste dal progetto è la realizzazione di stime di parametri socio-economici, con finalità di applicazione alla programmazione economica, ad un dettaglio territoriale più fine delle consuete unità amministrative (province e regioni). La dimensione territoriale scelta è quella del Sistema Locale del Lavoro (SLL), così come definiti dall'ISTAT sulla base dei FLussi di pendolarismo per motivi di lavoro[2].

Un'ultima e importante avvertenza va fatta sulla natura dei dati presentati. Come sarà illustrato in dettaglio nei paragrafi che illustrano la metodologia utilizzata, il lavoro di stima dell'occupazione residente e delle persone in cerca di occupazione è il risultato del miglior compromesso possibile tra precisione delle stime e correttezza delle stesse.
E' utile descrivere le condizioni di fondo che, in molti casi, hanno condizionato le scelte metodologiche operate:

a) la griglia territoriale è composta da 784 unità (i Sistemi locali del lavoro), di dimensione media piuttosto limitata (circa 74.000 residenti di media al 2000). La rilevanza dei SLL di dimensioni piccole o piccolissime è significativa: il 42,7% dei SLL presenta una dimensione inferiore ai 20.000 abitanti. Questa geografia, sfavorevole per la precisione delle stime, è mitigata dal peso percentuale della popolazione: il 6,5% della popolazione risiede in SLL con meno di 20.000 abitanti;

b) la geografia dei SLL presenta pochissime intersezioni con i confini amministrativi regionali: 47 SLL sono costituiti da territori afferenti a due regioni, 1 SLL è composto da zone appartenenti a tre regioni mentre i restanti 736 sono contenuti in una sola regione. Le intersezioni sono maggiori quando si considerano i confini provinciali: solo 614 SLL sono compresi in un'unica provincia (a cui corrisponde il 61,3% della popolazione residente totale), 151 SLL occupano il territorio di due province (31,5% della popolazione residente) ed infine 19 SLL sono a cavallo di tre o quattro province diverse (7,2% della popolazione);

[2] Cfr. ISTAT, I sistemi locali del lavoro 1991, Collana Argomenti n. 10, Roma 1997

c) I sistemi locali del lavoro costituiscono un dominio territoriale di studio *non pianificato* in quanto il disegno di campionamento dell'indagine sulle forze di lavoro prevede una stratificazione dei comuni a livello provinciale ma non a livello di SLL; questo fa sì che in alcuni domini si possa presentare una dimensione campionaria nulla: per il periodo di riferimento considerato (1998-2000) si dispone di circa 1.350 comuni campione provenienti dall'indagine sulle Forze di lavoro il che ha comportato, sempre in media, ad avere circa il 35% dei SLL senza nessuna osservazione diretta. Si tratta però di SLL di dimensioni piccole o piccolissime, tanto che la popolazione totale di riferimento di queste aree non arriva al 9% della popolazione totale italiana;

d) dall'indagine sulle Forze di lavoro si dispone, per le quantità in oggetto, di stime campionarie a livello regionale e provinciale, anche se con diversi livelli di errore;

e) le informazioni disponibili a livello comunale, e pertanto aggregabili per SLL, sono date esclusivamente della popolazione residente per sesso e classi quinquennali di età.

Date le caratteristiche descritte, ed in particolare da quelle riportate nel punto c), ne è scaturita la necessità di affrontare il problema con metodi innovativi, avvalendosi di modelli di stima e non della sola informazione campionaria. Inoltre i metodi utilizzati non sono immediatamente riconducibili a quelli utilizzati per produrre le stime ufficiali sulle forze di lavoro; ne consegue che i risultati qui presentati non trovano un'esatta corrispondenza con le stime a livello provinciale, ma solo con quelle regionali. Va però sottolineato che l'obiettivo primario non è stato quello di riprodurre delle stime già esistenti quanto piuttosto di ottenere le migliori stime possibili a livello di sistema locale del lavoro.

2. LA RILEVAZIONE TRIMESTRALE SULLE FORZE DI LAVORO

2.1 Caratteristiche generali

La Rilevazione Trimestrale sulle Forze di Lavoro (RTFL) è un'indagine campionaria condotta dall'ISTAT a partire dal 1959 che viene svolta con cadenza trimestrale nei mesi di gennaio, aprile, luglio e ottobre. L'obiettivo primario dell'indagine è la stima ufficiale del numero degli occupati e del numero delle persone in cerca di occupazione, delle loro variazioni nette, nonché di altre dettagliate informazioni riguardanti l'atteggiamento della popolazione in età lavorativa nei confronti del mercato del lavoro.

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.)

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). 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.

Il campione della RTFL, operativo dal luglio 1990, è stato progettato per garantire prefissati livelli attesi di precisione delle principali stime regionali. Nell'aprile 1995 il numero dei comuni campione è stato ampliato ed è passato da 1.200 a 1.351 e le famiglie intervistate sono diventate circa 75.000. Tale modifica si è resa necessaria per tenere conto della costituzione di otto nuove province. L'allargamento del campione ha consentito di ottenere stime, in media annuale, anche a livello provinciale del numero di occupati e del numero delle persone in cerca di occupazione.

Le stime vengono prodotte seguendo le innovazioni metodologiche e le nuove definizioni proposte dall'EUROSTAT al fine di rendere maggiormente comparabili le statistiche a livello internazionale. Le stime di occupati e persone in cerca di occupazione si riferiscono ai residenti in famiglia al netto delle persone temporaneamente emigrate all'estero

I principali parametri oggetto di stima sono:

A. Forze di lavoro: comprendono le persone occupate e quelle in cerca di occupazione.

B. Occupati: comprendono le persone di 15 anni e più che alla domanda sulla condizione professionale rispondono:

- di possedere un'occupazione, anche se nella settimana di riferimento non hanno svolto attività lavorativa (occupati dichiarati);

- di essere in una condizione diversa da occupato, ma di aver effettuato ore di lavoro nella settimana di riferimento (altre persone con attività lavorativa).

C. Persone in cerca di occupazione: comprendono le persone di 15 anni e più che dichiarano:

- una condizione professionale diversa da occupato;

- di non aver effettuato ore di lavoro nella settimana di riferimento dell'indagine;

- di essere alla ricerca di un lavoro;

- di aver effettuato almeno un'azione di ricerca di lavoro "attiva" nei trenta giorni che precedono la rilevazione;

- di essere immediatamente disponibili (entro due settimane) ad accettare un lavoro, qualora venga loro offerto.

D. Non forze di lavoro: comprendono le persone che dichiarano di essere in condizione professionale diversa da occupato e di non aver svolto alcuna attività lavorativa, né aver cercato lavoro nella settimana di riferimento; oppure di averlo cercato, ma non con le modalità già definite per le persone in cerca di occupazione. Le non forze di lavoro comprendono inoltre gli inabili e i militari di leva o in servizio civile sostitutivo e la popolazione in età fino a 14 anni.

E. Tasso di occupazione: si ottiene dal rapporto tra gli occupati e la popolazione di 15 anni e più.

F. Tasso di disoccupazione: si ottiene dal rapporto tra le persone in cerca di occupazione e le forze di lavoro.

Si può notare, quindi, che la classificazione degli individui nelle diverse condizioni professionali viene decisa non solo sulla base dell'autopercezione dei soggetti, ma anche attraverso un insieme di altre informazioni raccolte sulle attività effettivamente svolte dagli intervistati nel corso della settimana di riferimento[3].

Le stime prodotte dalla RTFL possono essere utilizzate per analisi sia strutturali sia congiunturali sui principali domini territoriali di studio; è necessario tenere presente che il grado di dettaglio territoriale raggiungibile è subordinato alle esigenze di precisione e

[3] Per approfondimenti si veda la premessa del volume Istat, Forze di Lavoro - Media 2000 ed il comunicato stampa Istat "La revisione delle serie storiche delle forze di lavoro. Ottobre 1992 - Aprile 1999 del 16 luglio 1999"

accuratezza delle stime stesse. Ne consegue che l'evoluzione del mercato del lavoro su base trimestrale può essere studiata a livello nazionale, ripartizionale e regionale. Con riferimento al livello provinciale, invece, l'analisi può essere svolta utilizzando esclusivamente i risultati della media annua, che viene pubblicata dall'Istat come media delle quattro rilevazioni trimestrali.

2.2. La strategia di campionamento dell'indagine RTFL

2.2.1. Il disegno di campionamento

L'indagine RTFL è basata su un disegno campionario di tipo composito. All'interno di ogni provincia i comuni sono divisi in due insiemi: l'insieme dei comuni auto rappresentativi (AR), costituito dai comuni di maggiore dimensione demografica e l'insieme dei comuni non auto rappresentativi (NAR), formato dai restanti comuni. Nell'insieme AR si adotta un disegno di campionamento ad uno stadio stratificato. Ciascun comune costituisce strato a se stante e le unità primarie sono le famiglie selezionate attraverso un campione sistematico. Tutti i membri di ciascuna famiglia sono intervistati.

Nell'insieme NAR il campione è basato su un disegno di campionamento a due stadi con stratificazione dei comuni. Due comuni sono selezionati da ciascuno strato con probabilità proporzionale al numero totale di individui del comune. Le famiglie vengono selezionate attraverso un campione sistematico e per ogni famiglia estratta sono intervistati tutti i componenti.

Il campione di primo stadio è costituito da 1.351 comuni, quello di secondo stadio da 73.000 famiglie, che danno luogo ad un campione di circa 200.000 individui. I comuni campione sono sempre gli stessi alle varie occasioni; per le famiglie, invece, si adotta uno *schema di rotazione* del tipo 2-2-2: ossia, una famiglia è inclusa nel campione per due rilevazioni successive e, dopo una pausa di nove mesi, fa parte del campione per altre due rilevazioni.

Detto schema di rotazione, presentato per due anni, è visualizzato nel prospetto di seguito riportato dal quale si trae che:

- si impiegano quattro gruppi di rotazione per formare il campione di una rilevazione trimestrale (il campione di un anno coinvolge nove differenti gruppi di rotazione);
- ogni trimestre entra nel campione un nuovo gruppo di rotazione;
- a regime, la frazione di sovrapposizione è pari a:
  - 0,50 con la rilevazione a tre mesi di distanza;
  - 0,25 a nove mesi;
  - 0,50 ad un anno di distanza;
  - 0,25 a quindici mesi.

Prospetto 1: Schema di rotazione dell'indagine RTFL

Anno	Rilevazioni	Gruppi di rotazione											
A_1	Gennaio	a	b			e	f						
	Aprile		b	c			f	g					
	Luglio			c	d			g	h				
	Ottobre				d	e			h	l			
A_2	Gennaio					e	f			l	m		
	Aprile						f	g			m	n	
	Luglio							g	h			n	o
	Ottobre								h	l			o

2.2.2. Notazioni e parametro di interesse

Per semplicità, verrà introdotta esclusivamente la notazione riferita al disegno di campionamento a due stadi dei comuni NAR, in quanto è possibile derivare le quantità e le espressioni relative ai comuni AR come casi particolari di quelle relative ai comuni NAR. Con riferimento ad una data generica regione geografica, sia: t (t=1,...,4) l'indice di trimestre; p (p=1,...,L) l'indice di provincia; h (h=1,...,H_p) l'indice di strato; i (i=1,...,N_h) l'indice relativo all'unità primaria (comune); j (j=1,...,M_{hi}) l'indice relativo all'unità secondaria (famiglia); a (a=1,...,A) l'indice relativo alle combinazioni della variabili sesso e classe di età. Inoltre, la famiglia j del comune i dello strato h viene indicata sinteticamente con il simbolo hij, mentre il comune i dello strato h è indicato come hi. Siano, infine: P_h il numero totale di persone in h; P_{hi} il numero totale di persone in hi; P_{hij} il numero totale di persone in hij; P_{ahij} il numero totale di persone in hij appartenenti alla classe a.

Con riferimento alla generica regione geografica è possibile, pertanto, introdurre l'espressione formale del parametro oggetto di stima riferito al trimestre di indagine t:

$$Y_t = \sum_{a=1}^{A} \sum_{p=1}^{L} \sum_{h=1}^{H_p} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{ahij}$$

in cui Y_{ahij} denota il totale della caratteristica di interesse y per gli individui della famiglia hij, intervistata nel trimestre t, e l'espressione formale del parametro oggetto di stima riferito all'anno:

$$Y = \frac{1}{4}\sum_{t=1}^{4} Y_t$$

ottenuto come media dei parametri trimestrali Y_t di un dato anno solare.

2.2.3. Stimatore di Y

Lo stimatore del totale Y si ottiene come media degli stimatori dei totali trimestrali Y_t (t=1,…,4); lo stimatore adottato correntemente per la produzione dei risultati trimestrali dell'indagine RTFL è lo stimatore di *ponderazione vincolata*, noto in letteratura con il nome *calibration estimator* (Deville e Särndal, 1992) espresso da:

$$_{PV}\hat{Y}_t = \sum_{a=1}^{A} {}_{PV}\hat{Y}_{ta} \,, \qquad (1)$$

dove:

$$_{PV}\hat{Y}_{ta} = \sum_{p=1}^{L}\sum_{h=1}^{H_p}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} K_{hij}\,\lambda_{hij}\,Y_{ahij} \,, \qquad (2)$$

in cui:

– K_{hij} rappresenta il *peso base* della generica famiglia hij , dato dalla seguente espressione:

$$K_{hij} = \frac{P_h}{n_h}\,\frac{M_{hi}}{P_{hi}\,m_{hi}} \qquad (3)$$

essendo per i comuni AR:

$$K_{hij} = \frac{M_{hi}}{m_{hi}}$$

in quanto si ha $N_h = n_h = 1$

– λ_{hij} è un *coefficiente di correzione* del peso base K_{hij} ottenuto come soluzione del seguente sistema di minimo vincolato

$$\begin{cases} \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \mathrm{Dist}(\lambda_{hij};1) = \min \\ {}_{PV}\hat{P}_{ta} = P_{ta}, \quad (a=1,\ldots,A) \end{cases} \tag{4}$$

dove la quantità:

$${}_{PV}\hat{P}_{ta} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} P_{ahij}\,\lambda_{hij}K_{hij} \tag{5}$$

rappresenta la stima del totale noto:

$$P_{ta} = \sum_{p=1}^{L}\sum_{h=1}^{H_p}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} P_{ahij}\ . \tag{6}$$

Per la costruzione dello stimatore di ponderazione vincolata sono stati considerati A=28 post-strati così formati: sesso (maschio, femmina) per classe di età (0 –| 14, 15 |– 20, 20 |– 25, 25 |– 30, 30 |– 35, 35 |– 40, 40 |– 45, 45 |– 50, 50 |– 55, 55 |– 60, 60 |– 65, 65 |– 70, 70 |– 75, 75 |–).

3. LA METODOLOGIA ADOTTATA PER LA COSTRUZIONE DELLE STIME RIFERITE AI SISTEMI LOCALI DEL LAVORO

3.1. Parametro di interesse

Con riferimento ad un generico SLL d (d=1,...,D), si denoti con: L_d il numero di provincie[4] che includono uno o più comuni appartenenti al SLL ; H_{dp} il numero degli strati, formati all'interno della generica provincia p, che includono il SLL; N_{dh} il numero dei comuni dello strato h appartenenti al SLL .

Con riferimento al generico sistema locale d, è possibile pertanto introdurre, l'espressione formale del parametro oggetto di stima riferito al generico trimestre di indagine t

$$Y_{td} = \sum_{a=1}^{A} Y_{tda} \ , \qquad (7)$$

dove:

$$Y_{tda} = \sum_{p=1}^{L_d} \sum_{h=1}^{H_{dp}} \sum_{i=1}^{N_{dh}} \sum_{j=1}^{M_{hi}} Y_{ahij}$$

e l'espressione formale del parametro oggetto di stima riferito all'anno

$$Y_d = \frac{1}{4} \sum_{t=1}^{4} Y_{td}$$

ottenuto come media dei parametri trimestrali Y_{td} di un dato anno solare.

Per non appesantire eccessivamente la trattazione, nei successivi paragrafi verranno descritti unicamente gli stimatori del parametro Y_{td} , in quanto una stima del parametro

[4] Le provincie che includono un SLL possono appartenere anche a più di una regione.

Y_d può essere ottenuta facilmente come media delle stime dei parametri trimestrali di un dato anno solare.

3.2 Stimatori del parametro Y_{td}

3.2.1 Stimatori diretti

I sistemi locali del lavoro costituiscono un dominio territoriale di studio *non pianificato* - in quanto il disegno di campionamento dell'indagine sulle forze di lavoro prevede una stratificazione dei comuni a livello provinciale, da ciò deriva che alcuni dei SLL di minore dimensione demografica possono non essere rappresentati nel campione oppure possono essere rappresentati con pochissime unità campionarie. A tale proposito è sufficiente osservare che il numero medio di comuni campione per SLL è inferiore a due in quanto il numero di unità primarie rilevate in ciascun trimestre è di circa 1.350 comuni e la numerosità complessiva dei SLL definiti sul territorio italiano è pari a 784 unità.

Da quanto detto risulta evidente che:

- i SLL di ampiezza limitata hanno una dimensione campionaria attesa molto esigua, cosicché gli errori delle stime dirette sarebbero così elevati da inficiarne l'utilizzo;
- una volta estratto il campione di comuni, gli stimatori *diretti* non possono essere calcolati per tutti i SLL in quanto alcuni possono risultare privi di famiglie campione;

Al fine della costruzione delle stime dirette riferite a ciascun SLL sono stati considerati due tipi di stimatori:

i. il primo è lo stimatore di ponderazione vincolata. Calcolato con riferimento a ciascun SLL:

$$_{PV}\hat{Y}_{td} = \sum_{a=1}^{A} {}_{PV}\hat{Y}_{tda} \qquad (8)$$

dove, è indicata con δ_{dhi} una variabile indicatrice che assume valore 1 se il comune hi appartiene al SLL d e 0 in caso contrario, si ha:

$$_{PV}\hat{Y}_{tda} = \sum_{p=1}^{L_d} \sum_{h=1}^{H_{dp}} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hij} Y_{ahij} \delta_{dhi},$$

essendo:

$$W_{hij} = K_{hij}\,\lambda_{hij}$$

il *peso finale* assegnato alla famiglia hij, ottenuto dal prodotto del peso diretto K_{hij} espresso dalla (3) con il corrispondente correttore λ_{hij} risoluzione del sistema (4);

ii. il secondo stimatore diretto considerato è uno stimatore del *rapporto post-stratificato* che assume la seguente espressione

$$_{R}\hat{Y}_{d} = \sum_{a=1}^{A'} \frac{_{PV}\hat{Y}_{tda}}{_{PV}\hat{P}_{tda}} P_{tda} \qquad (9)$$

dove A' rappresenta il numero di post-strati utilizzati per lo stimatore $_{R}\hat{Y}_{td}$ e

$$_{PV}\hat{Y}_{tda} = \sum_{p=1}^{L_d} \sum_{h=1}^{H_{dp}} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hij}\, Y_{ahij}\, \delta_{dhi},$$

$$_{PV}\hat{P}_{tda} = \sum_{l=1}^{L_d} \sum_{h=1}^{H_{dp}} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hij} P_{ahij}\, \delta_{dhi}$$

sono rispettivamente gli stimatori di ponderazione vincolata dei corrispondenti totali:

$$Y_{tda} = \sum_{p=1}^{L_d} \sum_{h=1}^{H_{dp}} \sum_{i=1}^{N_{dh}} \sum_{j=1}^{M_{hi}} Y_{ahij},$$

$$P_{tda} = \sum_{p=1}^{L_d} \sum_{h=1}^{H_{dp}} \sum_{i=1}^{N_{dh}} \sum_{j=1}^{M_{hi}} P_{ahij}$$

3.2.2. Stimatori indiretti

Per le ragioni descritte nel paragrafo precedente si è reso necessario l'uso di stimatori *indiretti* per la produzione di stime a livello di sistema locale (gli stimatori che vengono riportati sono anche noti come *stimatori per piccole aree*). Tali stimatori si basano:

- sull'utilizzo di informazioni ausiliarie, correlate ai fenomeni oggetto di studio, note a livello di ciascun SLL;

- sull'adozione (implicita o esplicita) di modelli statistici che legano i valori della variabile di interesse e delle variabili ausiliarie a livello di SLL con i valori delle medesime variabili relativi ad un area più grande (detta *macroarea*) contenente la piccola area di interesse e/o relativi ad altre occasioni di indagine oltre quella corrente; tali metodi sono noti rispettivamente come metodi di: *smoothing spaziale*, *smoothing temporale* e *smoothing spaziale e temporale*.

Il ricorso ai metodi di stima indiretti comporta il fatto di accettare un certo livello di distorsione nelle stime compensato però da una minore varianza e conseguentemente da un livello più basso dell'errore quadratico medio. Un problema fondamentale di tali metodi è legato al fatto che si basano su modelli e pertanto le proprietà degli stimatori sono legate alla validità del modello ipotizzato.

Nel par. 3.3**.** sono descritti i metodi di stima presi in considerazione e le analisi empiriche che sono state effettuate per la scelta dello stimatore da utilizzare.

In base alle analisi empiriche effettuate, lo stimatore caratterizzato dalle migliori *performance*, in termini di errore quadratico medio, è risultato lo stimatore composto con una macroarea regionale. Tale stimatore, che è espresso da una combinazione lineare convessa tra uno stimatore diretto ed uno stimatore indiretto di tipo sintetico, rappresenta un valido metodo per bilanciare la naturale instabilità dello stimatore diretto e la potenziale distorsione dello stimatore sintetico. Se da un lato, infatti, lo stimatore diretto è asintoticamente non distorto sotto il disegno è altresì caratterizzato generalmente da una forte variabilità campionaria funzione indiretta delle unità campionarie osservate nel SLL; lo stimatore sintetico, caratterizzato da una variabilità campionaria inferiore rispetto a quella dello stimatore diretto, è, tuttavia, potenzialmente distorto e l'entità della distorsione è tanto più elevata quanto più le ipotesi del modello non sono verificate nella realtà.

Lo stimatore composto utilizzato ha la seguente espressione formale:

$$ {}_{C}\hat{Y}_{t} = \alpha_{d} \; {}_{R}\hat{Y}_{td} + (1-\alpha_{d}) \; {}_{S}\hat{Y}_{td} , \qquad (10) $$

dove α_d è una quantità costante indipendente dal campione estratto $(0 \le \alpha_d \le 1)$ ${}_{R}\hat{Y}_{td}$ è lo stimatore rapporto post-stratificato espresso dalla (9) e ${}_{S}\hat{Y}_{td}$ è lo stimatore sintetico definito dalla seguente espressione:

$$ {}_S\hat{Y}_{td} = \sum_{a=1}^{A} \frac{PV\hat{Y}_{ta}}{PV\hat{P}_{ta}} P_{t\,da} , \qquad (11) $$

che in base alla (4) può essere riscritto come:

$$ {}_S\hat{Y}_{td} = \sum_{a=1}^{A} \frac{PV\hat{Y}_{ta}}{P_{ta}} P_{tda} . $$

Per quanto riguarda i post-strati, definiti sulla base delle variabili sesso ed età, si ha quanto segue:

- per lo stimatore ${}_S\hat{Y}_{td}$, sono stati considerati i medesimi post-strati (A=28) utilizzati per la costruzione dello stimatore di ponderazione vincolata (1);
- per lo stimatore ${}_R\hat{Y}_{td}$, si sono considerati $A' = 4$ post-strati così definiti: sesso (maschio, femmina) per classe di età (0–| 40 , 40 –).

E' importante notare che lo stimatore ${}_R\hat{Y}_{td}$ è basato sulla costruzione di post-strati definiti a livello di SLL, a differenza degli stimatori (8) e (11) in cui i post-strati sono definiti a livello di regione; si è scelto, pertanto, di definire un numero ridotto di post-strati ($A' = 4$) al fine di garantire la presenza di unità campione in ciascuno di essi.

I pesi α_d che compaiono nella (10) sono stati determinati in modo da minimizzare l'errore quadratico medio (EQM) (pari alla varianza dello stimatore ${}_C\hat{Y}_t$ più il quadrato della sua distorsione) dello stimatore ${}_C\hat{Y}_t$ e sono calcolati mediante la seguente espressione:

$$ \alpha_d = \frac{EQM({}_S\hat{Y}_{td})}{EQM({}_S\hat{Y}_{td}) + Var({}_R\hat{Y}_{td})} = \frac{Var({}_S\hat{Y}_{td}) + Bias^2({}_S\hat{Y}_{td})}{Var({}_S\hat{Y}_{td}) + Bias^2({}_S\hat{Y}_{td}) + Var({}_R\hat{Y}_{td})} \qquad (12) $$

Le varianze $Var({}_S\hat{Y}_{td})$ e $Var({}_R\hat{Y}_{td})$ dello stimatore sintetico e rapporto, sono state valutate applicando la trasformata di Woodroof (Woodroof, R.1971, A simple method for approximating the variance of complicated estimate, JASA) ai dati elementari del censimento generale della popolazione del 1991. La distorsione, $Bias({}_S\hat{Y}_{td})$, pari alla differenza tra il valore atteso dello stimatore ${}_S\hat{Y}_{td}$ e il valore vero del parametro oggetto di stima $\hat{Y}_{td}$, è stata valutata per ciascun SLL utilizzando rispettivamente un valore atteso

ed un valore vero del parametro di interesse calcolati in base ai dati del censimento 1991. Il calcolo delle varianze utilizzate nella (11) è risultato particolarmente impegnativo dal punto di vista computazionale, avendo richiesto il trattamento dei 56 milioni di record censuari.

Le stime del totale delle variabili di interesse ottenute attraverso lo stimatore composto (10) sono state riproporzionate in modo che le stime dei totali regionali di tali variabili coincidano con i totali stessi.

3.3. Studi empirici

3.3.1. Analisi delle proprietà empiriche degli stimatori ai fini della scelta dello stimatore migliore

Al fine di valutare le proprietà empiriche, in termini di distorsione ed errore quadratico medio, di alcuni stimatori per piccole aree di rilevante interesse applicativo e teorico, generalmente utilizzati dai principali centri di diffusione statistica, è stato effettuato uno studio simulativo basato sul metodo Monte Carlo. Tale studio consiste nell'estrazione ripetuta da una determinata popolazione di riferimento di un certo numero, R , di campioni definiti in base a determinati criteri. Lo studio può essere sintetizzato come segue:

- per la scelta della popolazione di riferimento da cui selezionare i campioni, piuttosto che generare una popolazione fittizia in base a determinati criteri, si è deciso di utilizzare una popolazione *reale* che fosse il più possibile simile a quella di interesse. Si è scelto pertanto di utilizzare i dati del censimento generale della popolazione del 1991;

- le informazioni relative alle variabili di interesse, alle variabili ausiliarie, ai totali delle variabili di interesse $Y_{td}\ (d=1,\ldots,D)$ e ai totali delle variabili ausiliarie $P_{tda}\ (d=1,...,D;\ a=1,...,A)$ sono tratte dal censimento generale della popolazione del 1991;

- il disegno di campionamento adottato ha struttura e numerosità campionarie (di primo e di secondo stadio) identiche a quelle adottate correntemente per l'indagine RTFL; anche se la stratificazione dei comuni e la suddivisione dei medesimi in autorappresentativi e non autorappresentativi è stata effettuata sui dati del censimento 1991;

- le variabili analizzate nello studio sono: occupati, disoccupati e persone in cerca di prima occupazione;

- le variabili ausiliarie per la post-stratificazione sono sesso e classi di età;

- le piccole aree di interesse sono i 27 SLL della regione Lazio;

- per la simulazione di Monte Carlo sono stati selezionati R=2.000 campioni a due stadi, per ciascuna delle cinque provincie della regione Lazio.

3.3.2. Misure delle performance

Per ciascuno dei 2.000 campioni estratti sono stati valutati gli stimatori in esame; infine sono state utilizzate come misure della bontà degli stimatori la loro distorsione e l'EQM (errore quadratico medio) medie sulle 2.000 replicazioni e su tutte i SLL:

- Valore Medio Assoluto della Distorsione Relativa (DR),

- Valore Medio della Radice Quadrata dell'Errore Quadratico Medio Relativo (REQMR),

espressi rispettivamente dalle formule:

$$DR\left({}_T\hat{Y}\right) = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{1}{R}\sum_{r=1}^{R}\left[\frac{{}_T\hat{Y}_d(r) - Y_d}{Y_d}\right]\right| 100 \quad ;$$

$$R\,\mathrm{EQMR}\left({}_T\hat{Y}\right) = \frac{1}{D}\sum_{d=1}^{D}\sqrt{\frac{1}{R}\left(\sum_{r=1}^{R}\left[\frac{{}_T\hat{Y}_d(r) - Y_d}{Y_d}\right]^2\right)} 100 \quad ;$$

dove ${}_T\hat{Y}_d(r)$ indica il valore del generico stimatore T ottenuto nell'r-sima degli R=2.000 campioni.

Le stesse misure sono, inoltre, state considerate calcolando le medie solo relativamente a opportuni sottoinsiemi dei SLL, con D sostituito dalla numerosità del sottoinsieme.

3.3.3. Descrizione formale degli stimatori indiretti sotto studio

Nell'analisi simulativa sono stati presi in considerazione, oltre agli stimatori diretti descritti al par. 2.2.2, i seguenti stimatori indiretti:

i. lo *stimatore sintetico*, la cui espressione è data dalla (11), il quale si basa

sull'assunzione implicita che, per ciascun post-strato, il rapporto Y_{ta}/P_{ta} (rispettivamente totale della variabile di interesse y e il totale della variabile ausiliaria riferiti alla macroarea di riferimento) sia uguale o abbastanza prossimo a quello di SLL, Y_{tda}/P_{tda} . Pertanto, esso può presentare una certa distorsione quando la suddetta ipotesi viene a cadere;

ii. lo *stimatore composto*, preso in considerazione in differenti forme caratterizzate da diverse scelte degli stimatori diretti presi in esame e da diversi valori dei pesi da attribuire; una delle formulazioni è l'espressione in formula (10) del par. 3.2.2, risultata la migliore dal punto di vista dell'EQM (vedi tab. 2) e per questo utilizzata per le stime prodotte; una differente espressione di stimatore composto si ha considerando un peso costante α per tutti i SLL di una data regione

Analogamente alle quantità α_d, la quantità α viene determinata in modo da minimizzare l'errore quadratico medio (EQM) dello stimatore composto ${}_d\hat{Y}_t$, ed è data dalla seguente espressione:

$$\alpha = \frac{\sum_{d=1}^{D} EQM({}_S\hat{Y}_{td})}{\sum_{d=1}^{D} EQM({}_S\hat{Y}_{td}) + \sum_{d=1}^{D} Var({}_R\hat{Y}_{td})}, \qquad (13)$$

dove le quantità $EQM({}_S\hat{Y}_{td})$ e $Var({}_R\hat{Y}_{td})$ sono state calcolate sulla base dei dati del censimento generale della popolazione del 1991.

Una terza forma di stimatore composto considerata è ottenuta attribuendo un valore del peso dello stimatore diretto unico per sottoinsiemi di SLL di una data regione e di una data dimensione di popolazione. I gruppi presi in considerazione sono: (a) SLL di piccole dimensioni in termini di popolazione residente (*aree piccole*), aventi una percentuale di popolazione rispetto al totale regionale inferiore al 2%; (b) SLL di medie dimensioni in termini di popolazione residente (*aree medie*), aventi una percentuale di popolazione al di sopra del 2%; il SLL di Roma è stato trattato separatamente.

Analoghe forme degli stimatori composti sopra descritti sono state calcolate, sostituendo lo stimatore diretto del rapporto post-stratificato con lo stimatore diretto di ponderazione vincolata (8).

Inoltre al fine di valutare se i pesi utilizzati nelle espressioni del composto; pesi che sono stati ottenuti attraverso i dati censuari del 1991, possano essere impiegati anche per stimare il totale degli occupati o il totale dei disoccupati riferiti a istanti lontani dal 1991, nell'analisi effettuata per la regione Lazio, sono stati calcolati gli stimatori composti in esame per l'anno 1991, assegnando ai coefficienti i valori ottenuti utilizzando per il calcolo delle varianze del sintetico e del diretto e della distorsione del sintetico i dati del censimento 1981; le performance degli stimatori composti risultanti sono state messe a confronto con

quelle degli stimatori con valori "ottimi". Come emerge dalla tabella 2, tale confronto porta a ritenere affidabile l'utilizzo dei valori dei pesi ottenuti attraverso i dati censuari del 1991 anche per gli anni successivi.

Infine, anche la scelta della macroarea dello stimatore sintetico, può essere oggetto di valutazione; in un lavoro precedentemente svolto (Conditional and Unconditional Analysis of Some Small Area Estimators in Complex Sampling, L. Di Consiglio P. D. Falorsi, S. Falorsi, A. Russo. SAE 2001 Conference), era stato effettuato il confronto tra lo stimatore composto in cui lo stimatore sintetico ha macroarea regionale con quello dello stimatore composto in cui lo stimatore sintetico ha macroarea provinciale. In termini di EQM è emerso che il primo metodo risulta preferibile (si veda tabella 2).

Per definizione, tutti gli stimatori composti vengono posti uguali allo stimatore sintetico nel caso in cui la dimensione del campione nel SLL d è uguale a 0;

iii. lo *stimatore sample size dependent*, espresso da

$$ {}_{SD}\hat{Y}_d = w_d \; {}_R\hat{Y}_d + (1-w_d) \; {}_S\hat{Y}_d \qquad (14) $$

dove:

$$ w_d = \begin{cases} 1 & \text{se } {}_{PV}\hat{P}_d \geq \ddot{e} P_d \\ {}_{PV}\hat{P}_d \,/ (\ddot{e} P_d) & \text{altrimenti} \end{cases} $$

$$ {}_{PV}\hat{P}_d = \sum_{a=1}^{A} {}_{PV}\hat{P}_{da} \qquad P_d = \sum_{a=1}^{A} P_{da} $$

Lo stimatore sample size dependent si basa sulla considerazione che il comportamento dello stimatore del rapporto post-stratificato dipende dalla proporzione del campione che cade nel SLL. Se tale proporzione è ragionevolmente ampia, allora lo stimatore (14) coincide con lo stimatore del rapporto post-stratificato, mentre, in caso contrario, diventa uno stimatore composto con un peso $(1-w_d)$ allo stimatore sintetico che cresce al diminuire della dimensione del campione nel SLL.

Per la costruzione dello stimatore sample size dependent sono stati utilizzati differenti valori del parametro (λ=2/3, λ=1,5 e λ=2); poiché il valore λ=2 ha condotto a migliori risultati in termini di EQM, nella tabella 2 vengono riportati unicamente i risultati ottenuti per tale valore.

Come per lo stimatore composto anche lo stimatore sample size dependent può avere formulazioni differenti sostituendo allo stimatore del rapporto post-stratificato lo stimatore di ponderazione vincolata (8).

3.3.4 Altri stimatori possibili

In un precedente lavoro (Conditional and Unconditional Analysis of some Small Area Estimators in Complex Sampling, L. Di Consiglio, P. D. Falorsi, S. Falorsi, A. Russo) è stato valutato uno stimatore della classe degli stimatori basati su modelli a componenti di varianza.

Lo stimatore *best linear unbiased predictor* ha la seguente forma

$$ {}_{P}\hat{Y}_d = \gamma_d \; {}_{PV}\hat{Y}_d + (1-\gamma_d)\, x_d' \, \tilde{\hat{a}} \qquad (15) $$

con:

$$ \tilde{\hat{a}} = \left[\sum_{d=1}^{D} x_d \, x_d' / (\sigma_v^2 + \psi_d) \right]^{-1} \left[\sum_{d=1}^{D} x_d \, {}_{E}\hat{Y}_d / (\sigma_v^2 + \psi_d) \right], \quad \gamma_d = \sigma_v^2 / (\sigma_v^2 + \psi_d) \qquad (16) $$

alla base dello stimatore è il noto modello lineare a livello di area di Fay and Herriot (1979)

$$ {}_{PV}\hat{Y}_d = x_d' \, \hat{a} + \acute{\imath}_d + e_d $$

in cui: β è il vettore dei parametri di regressione; x_d è un vettore di valori assunti da opportune variabili ausiliarie specifiche; ν_d sono variabili casuali incorrelate e identicamente distribuite con media 0 e varianza σ_v^2; e_d sono errori campionari indipendenti di media 0 e varianza ψ_d; $\tilde{\hat{a}}$ è lo stimatore dei minimi quadrati ponderati di β con coefficienti di ponderazione $(\sigma_v^2 + \psi_d)^{-1}$. Dal momento che la varianza σ_v^2 è incognita, nell'espressione (16), essa viene sostituita da una stima $\tilde{\sigma}_v^2$; lo stimatore che ne risulta viene definito in letteratura *empirical best linear unbiased predictor* (EBLUP).

Inoltre nel lavoro citato sono stati considerati due differenti metodi di valutazione delle varianze campionarie incognite ψ_d: nel primo la varianza ψ_d è stata valutata utilizzando i dati censuari; nel secondo, in analogia alle situazioni reali, la varianza ψ_d viene predetta tramite un modello di regressione basato su 12 (corrispondenti a tre anni)

campioni simulati dell'indagine RTFL. Dalle due valutazioni derivano due forme di EBLUP ${}_{EP_1}\hat{Y}_d$ e ${}_{EP_2}\hat{Y}_d$.

3.3.5 La valutazione dell'errore quadratico medio

L'errore quadratico medio dello stimatore composto (10) utilizzato per le stime del totale occupati e persone in cerca di occupazione può essere approssimato, assumendo nulla la covarianza tra lo stimatore sintetico e lo stimatore rapporto post-stratificato, come segue:

$$EQM\left({}_C\hat{Y}_{td}\right) = \alpha_d^2 \; Var\left({}_R\hat{Y}_{td}\right) + (1-\alpha_d)^2 \; EQM\left({}_S\hat{Y}_{td}\right) \qquad (17)$$

Si noti che come per la stima dei totali d'interesse (occupati e persone in cerca di occupazione), anche nel caso delle quantità che compaiono nell'espressione dell'EQM, l'informazione campionaria per la stima diretta nei SLL di piccola dimensione è scarsa, quando non inesistente dove in campione non è presente.

Pertanto la stima dell'EQM è stata realizzata operando le seguenti approssimazioni:

1. la stima delle varianze trimestrali dello stimatore rapporto e dello stimatore sintetico sono state ottenute applicando alle rispettive varianze trimestrali delle stime realizzate secondo un disegno campionario semplice senza ripetizione un effetto della strategia valutata a livello regionale (*deff*);

2. le varianze trimestrali delle stime realizzate secondo un disegno campionario semplice senza ripetizione di cui al punto 1, sono state ottenute utilizzando come stima sintetica della proporzione di occupati o di persone in cerca di occupazione la

3. le varianze annuali sono ottenute come prodotto della media delle quattro varianze trimestrali e un effetto di rotazione (*effr*); tale effetto viene quantificato sulla base del meccanismo di rotazione (prospetto 1) e dei coefficienti di correlazione intraclasse valutati da De Vitiis C, Falorsi S. (1998), in "Indagine forze di lavoro: Analisi e confronto di schemi di rotazione alternativi", *Progetto interarea per la ristrutturazione dell'indagine sulle forze di lavoro*, Doc. n. 1, dicembre 1998;

4. la distorsione relativa dello stimatore sintetico è stata posta pari alla distorsione relativa che le stime presentano al Censimento 1991,

Pertanto per ciascun anno abbiamo:

$$Bias({}_S\hat{Y}_d) = Bias({}_S\hat{Y}_d(1991)) / Y(1991) * {}_S\hat{Y}_d$$

$$Var({}_R\hat{Y}_d) = effr \cdot \frac{1}{4}\sum_{t=1}^{4} \frac{P_t^{\,2}}{n_t} \frac{{}_S\overline{\hat{Y}}_{dt}}{P_t}\left(1 - \frac{{}_S\overline{\hat{Y}}_{dt}}{P_t}\right) deff_t$$

$$\mathrm{Var}({}_S\hat{Y}_d) = \mathrm{effr} \cdot \frac{1}{4} \sum_{t=}^{4} \frac{P_{dt}{}^2}{n_t} \frac{{}_S\overline{\hat{Y}}_t}{P_t} \left(1 - \frac{{}_S\overline{\hat{Y}}_t}{\overline{P}_t} \right) \mathrm{deff}_t \ ,$$

la quantità $deff_t$ è ottenuta tramite l'applicazione della procedura per il calcolo degli errori GSSE su variabili linearizzate.

Tabella 1 - I Sistemi locali del Lazio, Popolazione e numero di Comuni

SLL	Popolazione 1991	Popolazione 1991 (%)	Numero di Comuni
398	6.005	0,12	5
396	7.364	0,14	3
391	8.901	0,17	4
407	11.392	0,22	2
393	12.500	0,24	3
414	12.656	0,25	4
406	13.051	0,25	3
395	16.012	0,31	5
390	19.823	0,39	8
411	23.226	0,45	5
394	30.193	0,59	6
408	45.274	0,88	5
392	51.789	1,01	13
416	59.512	1,16	10
402	71.906	1,40	15
401	72.080	1,40	34
400	72.235	1,41	4
412	78.249	1,52	5
409	88.84	1,73	7
399	97.80	1,90	42
405	114.361	2,23	3
397	133.303	2,60	18
413	146.133	2,85	41
410	170.945	3,33	6
404	198.010	3,86	16
415	259.382	5,05	35
403	3.314.237	64,54	65

Tabella 2 EQM relativo percentuale degli stimatori analizzati per occupati, disoccupati e persone in cerca di occupazione (macroarea regionale)

	Occupati	Disoccupati	Persone in cerca di prima occupazione
Stimatore di ponderazione vincolata	76,52	97,99	84,44
Stimatore rapporto post-stratificato	38,48	58,05	46,13
Stimatore sintetico	4,82	19,38	10,29
Metodi che utilizzano lo stimatore di ponderazione vincolata			
Sample Size Dependent	59,63	80,71	67,29
Composto con alfa ottimo (91)	5,91	17,61	10,00
Composto con alfa unico (91)	10,68	22,71	12,71
Composto con alfa per gruppi(91)	9,03	19,65	11,07
Composto con alfa ottimo (81)	6,73	20,15	10,47
Composto con alfa unico(81)	9,37	21,98	12,34
Metodi che utilizzano lo stimatore rapporto post-stratificato			
Sample Size Dependent	6,35	29,83	16,83
Composto con alfa ottimo (91)	4,36	17,25	9,70
Composto con alfa unico (91)	4,53	18,73	10,06
Composto con alfa per gruppi(91)	4,53	18,57	9,97
Composto con alfa ottimo (81)	4,41	19,38	9,89
Composto con alfa unico(81)	4,55	18,71	10,06

4. ALTRI ASPETTI METODOLOGICI

4.1 Metodo utilizzato per la stima della popolazione per SLL, sesso e classi di età

Per la definizione della popolazione residente per sesso e classi di età a livello di SLL si è fatto ricorso alla rilevazione sulla popolazione residente per sesso, anno di nascita e stato civile nei comuni (POSAS), disponibile con cadenza annuale. Uno dei maggiori problemi di questa rilevazione è la copertura, che per gli anni considerati non è mai stata totale. Nel 1998 (31.12.98) la copertura è stata del 98,65%, che corrisponde a 7.991 comuni rispondenti, raggiungendo il 98,86% nel 2000 (31.12.00) con 8.008 comuni rispondenti.

La metodologia adottata per stimare la popolazione per sesso e classe di età dei sistemi locali del lavoro, in assenza di informazioni complete, riprende in parte la metodologia Istat per la stima della popolazione provinciale per sesso e classe di età. In dettaglio la stima è stata ottenuta nel modo seguente:

- per ogni Provincia i comuni sono stati stratificati in 8 gruppi, in base alla combinazione di tasso di invecchiamento (raggruppamento in quartili), e dimensione demografica (sopra e sotto la mediana);
- all'interno di ciascuno strato è stata calcolata, per sesso, la struttura per età media dei comuni rispondenti;
- la struttura così definita viene *donata* ai comuni non rispondenti appartenenti al medesimo strato;
- applicando la struttura *donata* alla popolazione comunale per sesso (la popolazione residente per sesso nei comuni è disponibile con cadenza annuale per tutti i comuni) si stima per i comuni non rispondenti la popolazione per sesso e classe d'età;
- l'ultima fase riguarda il riproporzionamento delle differenze a livello provinciale, tra i comuni non rispondenti, considerando come vincolo la stima della popolazione provinciale per sesso ed età della rilevazione POSAS (le differenze tra le stime provinciali calcolate dall'Istat periodicamente relativamente alla rilevazione POSAS e le stime provinciali ottenute dall'aggregazione per sottosistemi provinciali si deve al diverso metodo di costruzione degli strati).

Il metodo usato, più noto in letteratura come *imputazione media* (Abbate, 1997) si basa sull'ipotesi che all'interno di ogni strato la variabilità del parametro d'interesse sia molto bassa e quindi che i valori si distribuiscano in un intorno molto piccolo della media.

4.2 Costruzione delle popolazioni trimestralizzate coerenti con le popolazioni trimestrali delle RTFL per sesso e classi di età

La costruzione delle popolazioni trimestralizzate, coerenti con le popolazioni trimestrali delle RTFL, per SLL e per i sottogruppi formati dall'incrocio di sesso e classi di età è stata effettuata in passi successivi. La fonte principale è costituita dalla popolazione per sesso e classi di età per i comuni italiani. Questa popolazione è una stima che si basa principalmente sulla fonte anagrafica. Essa è disponibile solo con riferimento al 1 gennaio di ciascun anno. Per alcuni comuni l'ISTAT non pubblica la popolazione per sesso e classi di età. Ai fini del presente lavoro, per sopperire a questa mancanza si è utilizzata la procedura di ricostruzione della popolazione presentata nel paragrafo 4.1.

Per rendere più omogenea questa popolazione con quella di riferimento della RTFL, sono state sottratte le convivenze per classi di età e sesso, cosi come risultano dal Censimento 91, e ricalcolate per ciascun comune, tenendo conto di tutti i cambiamenti di codifica territoriale intervenute nel corso degli anni (comuni che vengono accorpati o si dividono).

Il totale della popolazione anagrafica residente in famiglia per ciascun anno è ancora diverso da quello risultante dalle stime della media annua RTFL per i seguenti motivi:

- Le stime di media per la RTFL sono ottenute semplicemente come media delle 4 rilevazioni trimestrali;
- Per ciascuna delle rilevazioni, il totale noto e la struttura della popolazione per sesso e classi di età che vengono utilizzati come totali noti, non fanno riferimento alla popolazione residente in famiglia all'inizio dell'anno, bensì ad una sua stima trimestralizzata. Nel momento in cui essa viene utilizzata come totale noto per la produzione delle stime RTFL tale stima risulta ancora provvisoria. Le 4 rilevazioni quindi utilizzano 4 popolazioni trimestralizzate diverse.
- Dalla popolazione di riferimento RTFL sono esclusi anche gli emigrati all'estero (tale aggregato non proviene da dati amministrativi, ma è una stima dell'indagine stessa, per cui soggetta ad errore campionario).

Dovendo comunque garantire la coerenza tra le stime RTFL e quelle SLL a livello Regionale, e dovendo anche garantire la coerenza tra le stime di popolazione per diversi domini territoriali abbiamo ricalcolato la popolazione trimestrale per sistema locale, per sesso e classi di età P_{tda} utilizzando la seguente relazione

$$\frac{P_{tda}}{{}_{PV}\hat{P}_{ta}} = \frac{P_{da}}{P_a}$$

dove:

- t indica il generico trimestre, d indica il generico SLL ed a indica il generico post-strato ottenuto dall'incrocio di sesso e classi di età;

- $_{PV}\hat{P}_{ta}$ è la popolazione risultante dalla RTFL per il trimestre t per il post-strato a;
- P_{da} è la popolazione anagrafica residente in famiglie, appartenenti al SLL d e post-strato a, con riferimento al 1 gennaio dell'anno considerato
- $P_a = \sum_{d \in D} P_{da}$ è la popolazione anagrafica residente in famiglie, appartenente al post-strato a, con riferimento al 1 gennaio dell'anno considerato

Ne segue che la popolazione da ricalcolare è

$$P_{tda} = \frac{P_{da}}{P_a} \; {}_{PV}\hat{P}_{ta}$$

ed è tale che, sommando rispetto a tutti i sistemi locali d, coincide con la popolazione della RTFL:

$$\sum_{d \in D} P_{tda} = \sum_{d \in D} \frac{P_{da}}{P_a} \; {}_{PV}\hat{P}_{ta} = {}_{PV}\hat{P}_{ta}$$

4.3. Metodo utilizzato per ottenere l'additività a livello regionale di occupati residenti e persone in cerca di occupazione

Lo stimatore composto adottato a livello di SLL non consente l'additività a livello regionale, con le stime dell'indagine delle forze di lavoro, e questo dipende sia dall'uso dell'alfa ottimo che del diretto rapporto.

L'importanza di una coerenza a livello regionale con le stime ufficiali dell'indagine forze di lavoro ha portato alla definizione di una procedura di "quadratura", che presenta le seguenti fasi:

1. Aggregazione, a livello regionale, delle stime di occupati e persone in cerca di occupazione per sottosistemi regionali (i sistemi locali del lavoro contenuti in una sola Regione, più le relative quota-parte di territorio di quei sistemi locali del lavoro iscritte in una sola Regione; nel complesso 833 sottosistemi):

   $$\hat{y}_{r,k} = \sum \hat{y}_{d_r,k}$$

   dove:

   r indica la generica Regione,

k è l'indice di occupati (k=1) e persone in cerca di occupazione (k=2),

dr indica il generico sottosistema d della generica Regione r;

2. Differenza tra le stime del punto 1 e le stime dell'indagine forze di lavoro a livello regionale, sia per gli occupati che per le persone in cerca di occupazione:

$$D_{r,k} = {}_{FL}\hat{y}_{r,k} - \hat{y}_{r,k}$$

dove:

${}_{FL}\hat{y}_{r,k}$ è la stima della generica Regione r per occupati/persone in cerca di occupazione dell'indagine Forze di lavoro;

3. Ad ogni sottosistema regionale è poi associata la popolazione ed il tasso di occupazione/disoccupazione provinciale dell'indagine forze di lavoro;

4. Definizione di un peso per ogni sottosistema locale del lavoro regionale: per gli occupati il peso è dato dal prodotto tra popolazione e tasso di occupazione del punto 3; per i disoccupati il peso è dato dal prodotto tra popolazione e tasso di disoccupazione;

$$\mathrm{P}_{\mathrm{d_r,k}} = \mathrm{x}_{\mathrm{d_r}}\, \mathrm{t}_{\mathrm{d_r,k}},$$

dove:

x_{d_r} è la popolazione del generico sottosistema d della regione r,

$t_{d_r,k}$ è il tasso di occupazione/disoccupazione associato al generico sottosistema d della regione r, come indicato al punto 3;

5. calcolo della quota da assegnare al generico sottosistema d della regione r per occupati/in cerca di occupazione:

$$\mathrm{q}_{\mathrm{d_r,k}} = (\mathrm{D}_{\mathrm{r,k}} / \sum_{\mathrm{i_r}} \mathrm{P}_{\mathrm{i_r,k}})\, \mathrm{P}_{\mathrm{d_r,k}};$$

6. stima finale del generico sottosistema d della regione r per occupati/persone in cerca di occupazione:

$$\tilde{y}_{d_r,k} = \hat{y}_{d_r,k} + q_{d_r,k}$$

L'idea alla base del metodo è di introdurre un peso, sulla base di variabili ausiliarie, che consenta di discriminare i sottosistemi all'interno di una regione. Per fare ciò è stata considerata la dimensione demografica e quindi il tasso di occupazione/disoccupazione più "prossimo" ai sottosistemi regionali che è dato da quello calcolato su base provinciale, ciò al fine di poter cogliere in maniera più precisa possibile le diversità intra-regionali.