File Standard

Istat
Istituto Nazionale di Statistica

Unione Europea
Fondo sociale europeo

Ministero del Lavoro, della Salute e delle Politiche Sociali
Direzione Generale del Mercato del Lavoro

INDAGINE MULTISCOPO SULLE FAMIGLIE
"Criticità dei percorsi lavorativi in un'ottica di genere"
Anno 2007

Manuale utente e tracciato record

"Criticità dei percorsi lavorativi in un'ottica di genere"
Anno 2007
Documentazione tecnica e descrizione del file

Premessa

Il Decreto Legislativo n. 322 del 6 settembre 1989 regola la diffusione delle informazioni statistiche prodotte nell'ambito del Sistema Statistico Nazionale al fine di garantire la riservatezza dei rispondenti. In particolare, per la diffusione di dati elementari, l'articolo 10, comma 2, dispone quanto segue: "Sono distribuite altresì ove disponibili, su richiesta motivata e previa autorizzazione del Presidente dell'Istat, collezioni campionarie di dati elementari, resi anonimi e privi di ogni riferimento che ne permetta il collegamento con singole persone fisiche e giuridiche".

Nell'osservanza di tale Decreto Legislativo e della Legge n. 675 del 31 dicembre 1996 l'Istat ha adottato misure e tecniche che rendono impossibile, o altamente improbabile, il collegamento dei dati rilasciati con l'unità statistica cui si riferiscono. Per tale motivo sono state apportate alcune modifiche sui files originali delle indagini, nell'intento di garantire la massima protezione ai dati, contenendo al minimo l'eventuale perdita di informazioni.

Le metodologie applicate si concretizzano nell'accorpamento e/o riclassificazione di modalità di variabili e nell'oscuramento di variabili. In quest'ultimo caso nei campi del tracciato record è riportata la dicitura "RISERVATO ISTAT".

Va considerato, inoltre, che la stessa dicitura è stata utilizzata anche per quelle variabili non attendibili dal punto di vista campionario e quindi non analizzabili statisticamente.

Finalità e caratteristiche dell'indagine

A partire dal dicembre 1993 l'Istat ha avviato il nuovo corso delle Indagini Multiscopo sulle Famiglie. Ogni anno, accanto all'indagine "Aspetti della vita quotidiana", si affiancano un'indagine a cadenza quinquennale, che approfondisce tematiche particolari, e un'indagine trimestrale su "Viaggi e vacanze".

Nel 2007 è stata condotta la rilevazione "Criticità dei percorsi lavorativi in un'ottica di genere", ovvero un'indagine di ritorno su un sottocampione di 10.000 individui già intervistati nell'indagine quinquennale "Famiglia e soggetti sociali - 2003". A differenza di quest'ultima, l'indagine "Criticità dei percorsi lavorativi in un'ottica di genere" ha come unità di analisi gli individui che nel 2003 avevano tra 18 e 64 anni e non le famiglie di fatto[1].

L'obiettivo principale è quello di esaminare le caratteristiche e il tessuto relazionale degli individui e ricostruire, a distanza di tre anni, le storie di vita individuali e familiari. Vengono analizzate, ad esempio, le transizioni dalla fase di formazione all'ingresso nel mercato del lavoro, il passaggio dallo stato di occupato a quello di ritirato dal lavoro, i flussi di entrata e di uscita nelle diverse fasi del ciclo di vita familiare (uscita dalla famiglia di origine, figli avuti, formazione e scioglimento delle unioni) e come queste si intrecciano con quelle a livello lavorativo, ecc. Grazie ai dati del 2007, quindi, è possibile valutare se e come i cambiamenti attesi, sia in ambito lavorativo che familiare, si siano verificati, quale sia stata la loro interferenza con fattori esterni e quali le modalità di adattamento e di risposta ad eventuali eventi critici verificatisi in tre anni.

La tecnica adottata per le re-interviste è di tipo CATI (Computer Assisted Telephone Interview). Le informazioni sono state raccolte per intervista diretta. Nei casi in cui l'individuo, per qualsiasi motivo, non sia stato disponibile all'intervista, le informazioni sono state fornite da un altro componente della famiglia (interviste proxy).

[1] Per le diverse definizioni, classificazioni e note alle tavole vedi Appendice A.

Avvertenze per l'utilizzazione del file

Per gli utenti esterni all'ISTAT viene messo a disposizione un file con le seguenti caratteristiche:

	Anno 2007
lunghezza record:	5.582
numero records individuali:	9.997
numero famiglie:	7.588

Ogni record individuale contiene una prima parte di informazioni generali e sulle caratteristiche anagrafiche di ciascun componente della famiglia di appartenenza, una seconda parte sulle caratteristiche individuali, rilevate nell'indagine di ritorno e una terza parte che riporta le informazioni già acquisite mediante l'indagine Famiglia e soggetti sociali del 2003.

Ogni componente è individuato dal numero progressivo univoco e il numero totale di appartenenti al campione è pari al numero di records: 9.997. Per selezionare i componenti della stessa famiglia, nel caso di doppia intervista[2], si considerano tutti i records individuali che hanno lo stesso numero generale progressivo della famiglia e numero d'ordine progressivo individuale all'interno della famiglia.

Ciascun record permette anche di studiare la composizione familiare, le relative tipologie familiari e di nucleo nonché il ruolo di ciascun individuo anche in relazione al 2003. Infatti, è possibile ricondurre ciascun membro della famiglia del 2007 allo stesso individuo della famiglia del 2003, qualora già presente.

Costruzione delle stime ed errori di campionamento

Le informazioni riportate nel file sono di carattere campionario. Per ottenere stime relative all'intera popolazione oggetto d'indagine è necessario moltiplicare ciascuna informazione per il coefficiente di riporto all'universo.

Tali coefficienti sono stati determinati in modo da poter essere utilizzati per costruire stime relative alle persone.

L'indagine ha la finalità di fornire stime riferite a:

1. l'intero territorio nazionale;
2. le cinque ripartizioni geografiche (Nord-Ovest, Nord-Est, Centro, Sud e Isole);
3. tre aree basate sulla tipologia socio-demografica dei comuni.

Nel diffondere i risultati di un'indagine campionaria occorre fornire agli utilizzatori le informazioni necessarie per valutare l'attendibilità delle stime ottenibili. Ad ogni stima corrisponde un errore campionario relativo; ciò significa che per consentire un uso corretto delle stime sarebbe necessario fornire per ogni stima il corrispondente errore campionario relativo. Questo, tuttavia, comporterebbe notevoli difficoltà per l'utilizzatore, dovute al fatto che la tutela della riservatezza impedisce di fornire i codici identificativi territoriali sui quali è basato il disegno dell'indagine. Per questo si ricorre ad una presentazione sintetica degli errori tramite il metodo dei modelli regressivi. Questo metodo si basa sulla determinazione di una funzione matematica che mette in relazione ciascuna stima con il proprio errore relativo.

[2] I 9.997 records dell'indagine "Criticità dei percorsi lavorativi in un'ottica di genere" riguardano: 5.179 interviste su un solo individuo per famiglia (in 5.179 famiglie); 4.818 interviste di due componenti della stessa famiglia (in 2.409 famiglie).

Si riportano in allegato le informazioni relative al campionamento e al calcolo degli errori di stima da cui è possibile individuare gli esempi di calcolo degli errori campionari. In seguito sono accluse le tavole per il calcolo degli errori relativi ai dati contenuti nel file standard, per stime sugli individui.

Appendice A

Definizioni, classificazioni e note alle tavole

I dati generali individuali fanno riferimento alle caratteristiche delle persone all'epoca dell'intervista del 2007.

In particolare:

- l'**età** è espressa in anni compiuti;

- il **titolo di studio** è quello più elevato conseguito;

- le **forze di lavoro** comprendono le persone occupate e le persone in cerca di occupazione;

- le **non forze di lavoro (inattivi)** comprendono le persone che non fanno parte delle forze di lavoro, ovvero quelle non classificate come occupate o in cerca di occupazione;

- gli **occupati** sono persone di 21 anni e più che nella settimana precedente l'intervista:
  - hanno svolto almeno un'ora di lavoro da cui hanno ricavato o ricaveranno un guadagno;
  - hanno svolto almeno un'ora di lavoro non retribuito nella ditta di un familiare (esclusi i lavoretti svolti per pagare le vacanze e i divertimenti);
  - sono assenti dal lavoro (ad esempio, per ferie o malattia, cassa integrazione guadagni, ecc.);

- le **persone in cerca di occupazione (disoccupati)** sono persone non occupate di 21 anni e più che:
  - hanno effettuato almeno un'azione attiva di ricerca di lavoro nelle 4 settimane precedenti l'intervista (hanno cercato un lavoro, anche part-time o occasionale, o hanno cercato di avviare un'attività economica autonoma) e sono disponibili a lavorare entro le due settimane successive all'intervista;
  - inizieranno un lavoro entro 3 mesi dalla data d'intervista;

- la **condizione soggettiva** è quella dichiarata come unica o prevalente dalle persone di 21 anni e più;

- la **posizione nella professione** è quella dichiarata come unica o prevalente dagli occupati di 21 anni e più che viene aggregata nel modo seguente:

  dirigenti, imprenditori, liberi professionisti;

  direttivi, quadri, impiegati, intermedi (appartenenti alle categorie speciali);

  capo operai, operai, subalterni (inclusi apprendisti, lavoratori a domicilio per conto di imprese);

  lavoratori in proprio, coadiuvanti (inclusi soci di cooperative di produzione di beni e/o prestazioni di servizio, collaboratori coordinati e continuativi e prestatori d'opera occasionali);

- le **ripartizioni geografiche** costituiscono una suddivisione geografica del territorio e sono così articolate:

  Nord-ovest: Piemonte, Valle d'Aosta, Lombardia, Liguria;

  Nord-est: Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia-Romagna;

  Centro: Toscana, Umbria, Marche, Lazio;

  Sud: Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria;,

  Isole: Sicilia, Sardegna;

- il **tipo di comune**

  i comuni italiani sono suddivisi nelle seguenti tipologie:

  comuni delle aree metropolitane: Sono inclusi i grandi comuni:Torino, Milano, Venezia, Genova, Bologna, Firenze, Roma, Napoli, Bari, Palermo, Catania, Cagliari; e comuni che gravitano intorno al centro dell'area metropolitana, come definiti sulla base dei risultati del Censimento della Popolazione;

  altri comuni: suddivisi in base alla dimensione demografica (fino a 10.000 abitanti, oltre 10.000 abitanti);

- **famiglia di fatto** e **nucleo familiare**

  la **famiglia di fatto** è costituita dall'insieme delle persone che:
  - hanno la loro dimora abituale nella stessa abitazione del capofamiglia anagrafico;
  - hanno con tale persona una relazione di matrimonio, parentela, affinità, adozione, tutela o affettiva;

  per **nucleo familiare** si intende l'insieme delle persone che formano una coppia con figli celibi o nubili, una coppia senza figli, un genitore solo con figli celibi o nubili;

  Una famiglia può coincidere con un nucleo, può essere formata da un nucleo più altri membri aggregati, da più nuclei (con o senza membri aggregati), o da nessun nucleo (persone sole, famiglie composte ad esempio da due sorelle, da un genitore con figlio separato, divorziato o vedovo, ecc.).

Strategia di campionamento e livello di precisione dei risultati

C.1. Obiettivi dell'indagine di ritorno

La *popolazione di interesse* dell'indagine in oggetto, ossia l'insieme delle unità statistiche intorno alle quali si intende investigare, è costituita dagli individui di età compresa tra 18 e 64 anni intervistati nell'ambito dell'indagine "Famiglia e soggetti sociali" del 2003 e appartenenti alle famiglie residenti in Italia, al netto dei membri permanenti delle convivenze. La famiglia è intesa come *famiglia di fatto*, ossia un insieme di persone coabitanti e legate da vincoli di matrimonio, parentela, affinità, adozione, tutela o affettivi.

L'obiettivo principale dell'indagine è quello di verificare, a distanza di 3 anni, quali sono stati i percorsi lavorativi e familiari degli individui; pertanto l'indagine fa riferimento alla popolazione *compresente* tra il 2003 e il 2007 ed è stata effettuata a partire dal campione intervistato per l'indagine "Famiglia e soggetti sociali" nell'anno 2003 (circa 19.000 famiglie e 50.000 individui); da questo è stato estratto un campione di circa 10.000 individui per i quali, quindi, sono disponibili le informazioni riferite sia al 2003 che al 2007.

Il *periodo di riferimento* è prevalentemente il momento dell'intervista, ossia il 2007, con alcuni confronti con quanto dichiarato nel 2003 e accaduto nel triennio.

I *domini di studio*, ossia gli ambiti rispetto ai quali sono riferiti i parametri di popolazione oggetto di stima, sono:

- l'intero territorio nazionale;
- le cinque ripartizioni geografiche (Italia Nord-Occidentale, Italia Nord-Orientale, Italia Centrale, Italia Meridionale, Italia Insulare);
- la tipologia comunale ottenuta suddividendo i comuni italiani in tre classi formate in base a caratteristiche socio-economiche e demografiche:
  A)comuni appartenenti all'area metropolitana
  B)comuni non appartenenti all'area metropolitana suddivisi in:
    comuni aventi fino a 10.000 abitanti;
    comuni con oltre 10.000 abitanti.

Il disegno di campionamento è un disegno complesso, definito in due fasi:

- la prima fase è costituita dal disegno campionario dell'indagine "Famiglia e soggetti sociali" del 2003 (descritto nel seguito);
- la seconda fase ha visto l'estrazione degli individui facenti parte dei 18-64enni che costituivano il campione dell'indagine "Famiglia e soggetti sociali" del 2003.

C.2. Prima fase di selezione del campione: l'indagine "Famiglia e soggetti sociali"[3]

C.2.1 Obiettivi conoscitivi dell'indagine

La *popolazione di interesse* dell'indagine "Famiglia e soggetti sociali", ossia l'insieme delle unità statistiche intorno alle quali si intende investigare, è costituita dalle famiglie residenti in Italia e dagli individui ad esse appartenenti, al netto dei membri permanenti delle convivenze. La famiglia è intesa come *famiglia di fatto*, ossia un insieme di persone coabitanti e legate da vincoli di matrimonio, parentela, affinità, adozione, tutela o affettivi.

Il *periodo di riferimento* è prevalentemente costituito dai dodici mesi che precedono l'intervista, anche se per alcuni quesiti il riferimento è il momento dell'intervista (fine 2003).

I *domini di studio*, ossia gli ambiti rispetto ai quali sono riferiti i parametri di popolazione oggetto di stima, sono:

- l'intero territorio nazionale;
- le cinque ripartizioni geografiche (Italia Nord-Occidentale, Italia Nord-Orientale, Italia Centrale, Italia Meridionale, Italia Insulare);
- le regioni geografiche (ad eccezione del Trentino-Alto Adige le cui stime sono prodotte separatamente per le province di Bolzano e Trento);
- la tipologia comunale.

C.2.2 Descrizione generale del disegno di campionamento

Il disegno di campionamento è di tipo complesso e si avvale di due differenti schemi di campionamento. Nell'ambito di ognuno dei domini definiti dall'incrocio della regione geografica con le sei aree A_1, A_2, B_1, B_2, B_3 e B_4, i comuni italiani sono suddivisi in due sottoinsiemi sulla base della popolazione residente:

- l'insieme dei comuni Auto Rappresentativi (che indicheremo d'ora in avanti come comuni AR) costituito dai comuni di maggiore dimensione demografica;
- l'insieme dei comuni Non Auto Rappresentativi (o NAR) costituito dai rimanenti comuni.

Nell'ambito dell'insieme dei comuni AR, ciascun comune viene considerato come uno strato a se stante e viene adottato un disegno noto con il nome di *campionamento a grappoli*. Le unità primarie di campionamento sono rappresentate dalle famiglie anagrafiche, estratte in modo sistematico dall'anagrafe del comune stesso; per ogni famiglia anagrafica inclusa nel campione vengono rilevate le caratteristiche oggetto di indagine di tutti i componenti di fatto appartenenti alla famiglia medesima.

Nell'ambito dei comuni NAR viene adottato un disegno a due stadi con stratificazione delle unità primarie. Le Unità Primarie (UP) sono i comuni, le Unità Secondarie sono le famiglie anagrafiche; per ogni famiglia anagrafica inclusa nel campione vengono rilevate le caratteristiche oggetto di indagine di tutti i componenti di fatto appartenenti alla famiglia medesima.

I comuni vengono selezionati con probabilità proporzionali alla loro dimensione demografica e senza reimmissione, mentre le famiglie vengono estratte con probabilità uguali e senza reimmissione.

C.2.3 Definizione della dimensione campionaria

Per un'indagine ad obiettivi plurimi, come quella in esame, è poco realistico pensare di poter disegnare una strategia campionaria che assicuri prefissati livelli di precisione di tutte le stime prodotte. La questione è complicata dal fatto che l'indagine ha la finalità di determinare stime per livelli territoriali differenti, il che comporta l'adozione di soluzioni di tipo ottimale diverse e contrastanti. Ad esempio, se l'unico ambito territoriale di pubblicazione delle stime fosse quello

[3] Per approfondimenti si veda la collana Informazioni n. 18 anno 2006 "*Strutture familiari e opinioni su famiglia e figli*" Indagine multiscopo sulle famiglie "*Famiglia e soggetti sociali*", Anno 2003.

nazionale, una soluzione approssimativamente ottimale sarebbe quella di determinare la numerosità nazionale e ripartirla tra le regioni in modo proporzionale alla loro dimensione demografica; viceversa, avendo la finalità di produrre stime con uguale attendibilità a livello regionale, una soluzione approssimativamente ottimale sarebbe quella di selezionare un campione uguale in tutte le regioni. Quest'ultima soluzione, però, è poco efficiente per le stime a livello nazionale. Per affrontare questo problema, conformemente a quanto fatto in altri paesi, si è fatto ricorso ad una strategia che perviene alla definizione della numerosità campionaria attraverso approssimazioni successive.

In base alle considerazioni precedenti si è deciso di adottare un'ottica mista basata sia su criteri di costo ed organizzativi, sia su una valutazione degli errori campionari delle principali stime a livello nazionale e con riferimento a ciascuno dei domini territoriali di interesse.

I criteri seguiti possono essere sintetizzati nei seguenti punti:

- la dimensione del campione teorico in termini di famiglie, prefissata a livello nazionale essenzialmente in base a criteri di costo ed operativi, è pari a circa 20.000;
- il numero di comuni campione interessati non deve essere superiore a 900 in modo da consentire un buon lavoro di controllo e supervisione.

L'allocazione del campione di famiglie e di comuni tra le varie regioni è stata poi definita adottando un criterio di compromesso tale da garantire sia l'affidabilità delle stime a livello nazionale che quella delle stime a livello di ciascuno dei domini territoriali descritti nel paragrafo C.2.1.

C.2.4 Stratificazione e selezione delle unità campionarie

Nell'indagine i comuni vengono stratificati in base alla loro dimensione demografica e nel rispetto delle seguenti condizioni:

- autoponderazione del campione a livello regionale;
- selezione di un comune campione nell'ambito di ciascuno strato definito sui comuni dell'insieme NAR;
- scelta di un numero minimo di famiglie da intervistare in ciascun comune campione; per l'indagine in oggetto tale numero è stato posto pari a 23;
- formazione di strati aventi ampiezza approssimativamente costante in termini di popolazione residente.

Effettuata la stratificazione, i comuni AR sono inclusi con certezza nel campione; per quanto riguarda, invece, i comuni NAR, nell'ambito di ogni strato viene estratto un comune campione con probabilità proporzionale alla dimensione demografica, mediante la procedura di selezione sistematica proposta da Madow[4].

La selezione delle famiglie da intervistare in ogni comune campione viene effettuata dalla lista anagrafica di ciascun comune senza reimmissione e con probabilità uguali.

C.3. Seconda fase di selezione del campione: l'indagine "Criticità dei percorsi lavorativi in un'ottica di genere"

A partire dal campione di individui in età compresa tra 18-64 anni selezionato nel 2003 con l'indagine "Famiglia e soggetti sociali", è stato estratto un campione casuale semplice di 10.000 individui.

La popolazione di riferimento dell'indagine (costituita dagli individui del 2003 che ancora sono residenti in Italia nel 2007) è stata definita a partire dalla popolazione del 2003, depurandola dalle uscite (morti e migrazioni) stimate, sulla base delle fonti ufficiali, nel triennio relativo. Le fonti ufficiali di cui ci si è avvalsi sono:

- i Bilanci demografici per il triennio in esame, per conoscere la popolazione residente e i decessi nel triennio;

[4] Madow, W.G. (1949) "On the theory of systematic sampling II", Ann. Math. Stat., 20, 333-354.

- l'indagine campionaria sulle Cause di morte (anno 2004), per stimare la distribuzione dei morti per le sole classi di età 18-64 nel triennio[5];
- i Trasferimenti di residenza per l'estero nel 2002-2004, per stimare la distribuzione degli usciti in età 18-64 dalla popolazione residente.

Nel prospetto C1 viene riportata la distribuzione per ripartizione dell'universo e del campione degli individui.

Prospetto C.1 - Distribuzione per ripartizione geografica degli individui nell'universo e nel campione - Anno 2007

RIPARTIZIONI	Universo (a)	Campione
Nord-Ovest	9.563.468	2.386
Nord-Est	6.830.680	2.403
Centro	6.888.013	1.906
Sud	8.683.617	2.488
Isole	4.128.546	814
Italia	36.094.324	9.997

(a) Stima della popolazione compresente 2003-2007

C.4. Procedimento per il calcolo delle stime

Le stime prodotte dall'indagine sono essenzialmente stime di frequenze assolute e relative, riferite agli individui. Le stime dell'indagine sono state ottenute mediante uno stimatore di ponderazione vincolata, che è il metodo di stima adottato per la maggior parte delle indagini ISTAT sulle imprese e sulle famiglie.

Il principio su cui è basato ogni metodo di stima campionaria è che le unità appartenenti al campione rappresentino anche le unità della popolazione che non sono incluse nel campione.

Questo principio viene realizzato attribuendo a ogni unità campionaria un peso che indica il numero di unità della popolazione rappresentate dall'unità medesima. Se, per esempio, a un'unità campionaria viene attribuito un peso pari a 30, allora questa unità rappresenta se stessa e altre 29 unità della popolazione che non sono state incluse nel campione.

Al fine di rendere più chiara la successiva esposizione, introduciamo la seguente simbologia:

- d, indice di livello territoriale di riferimento delle stime;
- i, indice di comune;
- j, indice di individuo;
- h, indice di strato di comuni;
- y, generica variabile oggetto di indagine;
- $Y_{hij,}$ valore di y osservato sull'individuo j del comune i dello strato h;
- M_{hi}, numero di individui nel comune i dello strato h;
- m_{hi}, campione di individui nel comune i dello strato h;
- N_h, totale di comuni nello strato h;
- n_h, numero di comuni campione nello strato h;
- H_d, numero totale di strati nel generico dominio territoriale d.

Ipotizziamo di voler stimare, con riferimento ad un generico dominio d, il totale della generica variabile y oggetto di indagine, espresso dalla seguente relazione

$$Y_d = \sum_{h=1}^{H_d} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{hij} \,. \qquad (1)$$

[5] Le classi di età considerate sono: 18-24, 25-34, 35-44, 45-54, 55-64.

La stima del totale (1) è data da

$$\hat{Y}_d = \sum_{h=1}^{H_d} \hat{Y}_h \ , \quad \text{essendo} \quad \hat{Y}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hij} Y_{hij} \ , \qquad (2)$$

in cui W_{hij} è il peso finale da attribuire agli individui j del comune i dello strato h.

Dalla precedente relazione si desume, quindi, che per ottenere la stima del totale Y (1) occorre moltiplicare il valore della variabile y assunto da ciascuna unità campionaria per il peso di tale unità ed effettuare, a livello del dominio di interesse, la somma dei prodotti così ottenuti.

Il peso da attribuire alle unità campionarie è ottenuto per mezzo di una procedura complessa che:

- corregge l'effetto distorsivo della mancata risposta totale dovuta all'impossibilità di intervistare alcuni individui selezionati per irreperibilità o per rifiuto all'intervista;
- tiene conto della conoscenza di totali noti di importanti variabili ausiliarie (disponibili da fonti esterne all'indagine), nel senso che le stime campionarie dei totali noti delle variabili ausiliarie devono coincidere con i valori noti degli stessi.

Nell'indagine in oggetto vengono definiti per ciascuna ripartizione geografica 10 totali noti, che si riferiscono alla distribuzione della popolazione per sesso e cinque classi di età. Indicando, quindi, con ${}_kX$ (k=1,…,10) il totale noto della k-esima variabile ausiliaria per la generica ripartizione geografica e con ${}_kX_{hij}$ il valore assunto dalla k-esima variabile ausiliaria per l'individuo rispondente hij, la condizione sopra descritta è espressa dalla seguente uguaglianza

$${}_k X = {}_k \hat{X} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hij\,k} X_{hij} \quad (k=1,\ldots., 10)$$

in cui H indica il numero complessivo di strati definiti nella ripartizione.

Poiché la popolazione di riferimento dell'indagine è costituita dagli individui in età 18-64 anni del 2003, che ancora sono residenti in Italia nel 2007, il calcolo dei pesi finali per l'indagine ha richiesto la conoscenza dei totali noti riferiti a tale popolazione. In base alle fonti ufficiali disponibili è stato possibile ricostruire i tali totali noti a livello di ripartizione, sesso e classi di età[6].

Il calcolo dei *pesi finali* da attribuire alle unità campionarie intervistate nel 2007 avviene per due passi successivi:

- passo 1: calcolo dei pesi finali dell'indagine 2003
- passo 2: calcolo dei pesi finali dell'indagine 2007 a partire dai pesi finali dell'indagine 2003.

Passo 1. Calcolo dei *pesi finali* da attribuire alle unità campionarie rispondenti nel *2003:*

1) si calcolano i *pesi diretti* come reciproco della probabilità di inclusione delle unità;
2) si calcolano i fattori correttivi per mancata risposta totale, come l'inverso del tasso di risposta del comune cui ciascuna unità appartiene;
3) si ottengono i *pesi base*, o pesi corretti per mancata risposta totale, moltiplicando i pesi diretti per i corrispondenti fattori correttivi per mancata risposta totale;
4) si costruiscono i fattori correttivi che consentono di soddisfare, a livello regionale, la condizione di uguaglianza tra i totali noti delle variabili ausiliarie e le corrispondenti stime campionarie;
5) si calcolano, infine, i *pesi finali* mediante il prodotto dei pesi base per i fattori correttivi ottenuti al punto 4 del passo 1.

[6] Le classi di età considerate sono: 18-24, 25-34, 35-44, 45-54, 55-64.

Passo 2. Calcolo dei *pesi finali* da attribuire alle unità campionarie rispondenti dell'indagine 2007:

6) i *pesi diretti* sono posti uguale ai *pesi finali* definiti nel passo1 (indagine 2003);
7) si costruiscono i fattori correttivi che consentono di soddisfare, a livello di ripartizione, sesso e classe di età, la condizione di uguaglianza tra i totali noti (calcolati sulla popolazione compresente 2003-2007) delle variabili ausiliarie e le corrispondenti stime campionarie, attraverso la risoluzione di un problema di minimo vincolato;
8) si calcolano, infine, i *pesi finali* mediante il prodotto dei pesi base per i fattori correttivi ottenuti al punto 7 del passo 2.

I fattori correttivi dei passi 4 e 7 sono ottenuti dalla risoluzione di un problema di minimo vincolato, in cui la funzione da minimizzare è una funzione di distanza (opportunamente prescelta) tra i pesi base e i pesi finali e i vincoli sono definiti dalla condizione di uguaglianza tra stime campionarie dei totali noti di popolazione e valori noti degli stessi. La funzione di distanza prescelta è la funzione logaritmica troncata; l'adozione di tale funzione garantisce che i pesi finali siano positivi e contenuti in un predeterminato intervallo di valori possibili, eliminando in tal modo i pesi positivi estremi (troppo grandi o troppo piccoli).

Tutti i metodi di stima che scaturiscono dalla risoluzione di un problema di minimo vincolato del tipo sopra descritto rientrano in una classe generale di stimatori nota come stimatori di ponderazione vincolata[7]. Un importante stimatore appartenente a tale classe, che si ottiene utilizzando la funzione di distanza euclidea, è lo *stimatore di regressione generalizzata*. Tale stimatore riveste un ruolo centrale in quanto è possibile dimostrare[8] che tutti gli stimatori di ponderazione vincolata convergono asintoticamente, all'aumentare della numerosità campionaria, allo stimatore di regressione generalizzata.

C.5. Valutazione del livello di precisione delle stime

C.5.1 Metodologia di calcolo degli errori campionari

Le principali statistiche di interesse per valutare la variabilità campionaria delle stime prodotte da un'indagine sono l'errore di campionamento assoluto e l'errore di campionamento relativo. Indicando con $\hat{V}ar(\hat{Y}_d)$ la stima della varianza della generica stima $\hat{Y}_d$, la stima dell'errore di campionamento assoluto di $\hat{Y}_d$ si può ottenere mediante la seguente espressione

$$\hat{\sigma}(\hat{Y}_d) = \sqrt{\hat{V}ar(\hat{Y}_d)}\ ; \qquad (3)$$

la stima dell'errore di campionamento relativo di $\hat{Y}_d$ è invece definita dall'espressione

$$\hat{\varepsilon}(\hat{Y}_d) = \frac{\hat{\sigma}(\hat{Y}_d)}{\hat{Y}_d}. \qquad (4)$$

Come è stato descritto nel paragrafo C.4., le stime prodotte dall'indagine sono state ottenute mediante uno stimatore di ponderazione vincolata definito in base ad una funzione di distanza di tipo logaritmico troncato. Poiché, lo stimatore adottato non è funzione lineare dei dati campionari, per la stima della varianza $\hat{V}ar(\hat{Y}_d)$ si è utilizzato il metodo proposto da Woodruff; in base a tale metodo, che ricorre all'espressione linearizzata in serie di Taylor, è possibile ricavare la varianza di ogni stimatore non lineare (funzione regolare di totali) calcolando la varianza dell'espressione linearizzata ottenuta. In particolare, per la definizione dell'espressione linearizzata dello stimatore ci si è riferiti

[7] Nella letteratura in lingua anglosassone sull'argomento tali stimatori sono noti come *calibration* estimators.

[8] Deville J.C., Sarndal C.E. (1992) "Calibration Estimators in Survey Sampling", Journal of the American Statistical Association, vol. 87, pp. 376-382.

allo stimatore di regressione generalizzata, sfruttando la convergenza asintotica di tutti gli stimatori di ponderazione vincolata a tale stimatore, poiché nel caso di stimatori di ponderazione vincolata che utilizzano funzioni distanza differenti dalla distanza euclidea (che conduce allo stimatore di regressione generalizzata) non è possibile derivare l'espressione linearizzata dello stimatore. L'espressione linearizzata dello stimatore (2) è data, quindi, da

$$\hat{Y}_d \cong \hat{Z}_d = \sum_{h=1}^{H_d} \hat{Z}_h \quad , \quad \text{essendo} \quad \hat{Z}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} Z_{hij} W_{hij} \tag{5}$$

dove Z_{hij} è la variabile linearizzata espressa come $Z_{hij} = Y_{hij} - \mathbf{X}'_{\mathbf{hij}}\beta$, essendo $\mathbf{X}_{\mathbf{hij}} = \left({}_1X_{hij},\ldots,{}_kX_{hij},\ldots{}_KX_{hij}\right)'$ il vettore contenente i valori delle K variabili ausiliarie, osservati per il generico individuo hij e $\hat{\beta}$, il vettore dei coefficienti di regressione del modello lineare che lega la variabile di interesse y alle K variabili ausiliarie x. In base alla (5), si ha, quindi, che la stima della varianza della stima $\hat{Y}_d$ è ottenuta mediante la seguente relazione

$$\hat{\mathrm{V}}\mathrm{ar}\left(\hat{Y}_d\right) \cong \hat{\mathrm{V}}\mathrm{ar}\left(\hat{Z}_d\right) = \sum_{h=1}^{H_d} \hat{\mathrm{V}}\mathrm{ar}\left(\hat{Z}_h\right). \tag{6}$$

Dalla (6) risulta che la stima della varianza della stima $\hat{Y}_d$ viene calcolata come somma della stima delle varianze dei singoli strati, AR e NAR, appartenenti al dominio d. La formula di calcolo della varianza, $\hat{\mathrm{V}}\mathrm{ar}\left(\hat{Z}_h\right)$, della stima $\hat{Z}_h$ è differente a seconda che lo strato sia AR oppure NAR. Possiamo, quindi scomporre come segue

$$\hat{\mathrm{V}}\mathrm{ar}\left(\hat{Y}_d\right) \cong \hat{\mathrm{V}}\mathrm{ar}\left(\hat{Z}_d\right) = \sum_{h=1}^{H_{AR}} \hat{\mathrm{V}}\mathrm{ar}\left(\hat{Z}_h\right) \; + \sum_{h=1}^{H_{NAR}} \hat{\mathrm{V}}\mathrm{ar}\left(\hat{Z}_h\right) \quad , \tag{7}$$

in cui H_{AR} e H_{NAR} indicano rispettivamente il numero di strati AR e NAR appartenenti al dominio d.

Negli strati AR (in cui ciascun comune fa strato a sé e $N_h = n_h = 1$, l'indice i di comune diviene superfluo e viene omesso) la varianza è stimata mediante la seguente espressione

$$\sum_{h=1}^{H_{AR}} \hat{\mathrm{V}}\mathrm{ar}\left(\hat{Z}_h\right) = \sum_{h=1}^{H_{AR}} M_h^2 \frac{(M_h - m_h)}{m_h(m_h - 1)} \sum_{j=1}^{m_h} \left(Z_{hj} - \overline{Z}_h\right)^2 , \tag{8}$$

dove si è posto $M_h = M_{hi}$, $m_h = m_{hi}$, $Z_{hj} = Z_{hij}$ e $\overline{Z}_h = \frac{1}{m_h} \sum_{j=1}^{m_{h\cdot}} Z_{hj}$.

Negli strati NAR, in cui viene estratto un solo comune campione da ogni strato, per stimare la varianza di campionamento si ricorre alla *tecnica di collassamento degli strati*. Questa tecnica consiste nel formare G gruppi contenenti ciascuno L_g $(L_g \geq 2)$ strati; la varianza viene stimata mediante la formula seguente

$$\sum_{h=1}^{H_{NAR}} \hat{\mathrm{V}}\mathrm{ar}\left(\hat{Z}_h\right) = \sum_{g=1}^{G} \hat{\mathrm{V}}\mathrm{ar}\left(\hat{Z}_g\right) = \sum_{g=1}^{G} \frac{L_g}{L_g - 1} \sum_{h=1}^{L_g} \left(\hat{Z}_{hg} - \frac{\hat{Z}_g}{L_g} \right)^2 \tag{9}$$

dove le quantità sono espresse come

$$\hat{Z}_{hg} = \sum_{j=1}^{m_{hi}} Z_{hij} W_{hij} \qquad \text{e} \qquad \hat{Z}_g = \sum_{h=1}^{L_g} \sum_{j=1}^{m_{hi}} Z_{hij} W_{hij} \, .$$

Utilizzando le espressioni (8) e (9) è possibile, infine, calcolare la varianza di campionamento, $\hat{\mathrm{V}}\mathrm{ar}\left(\hat{Y}_d\right)$ in base alla (7) e calcolare, quindi, in base alla (3) ed alla (4) rispettivamente l'errore di campionamento assoluto e l'errore di campionamento relativo.

Gli errori campionari espressi dalla (3) e dalla (4) consentono di valutare il grado di precisione delle stime; inoltre, l'errore assoluto permette di costruire un intervallo di confidenza, che, con livello di fiducia P contiene il parametro oggetto di stima, l'intervallo viene espresso come

$$\left\{\hat{Y}_d - k_p\hat{\sigma}(\hat{Y}_d) \le Y_d \le \hat{Y}_d + k_p\hat{\sigma}(\hat{Y}_d)\right\} \qquad (10)$$

Nella (10) il valore di k_P dipende dal valore fissato per la probabilità P; ad esempio, per P=0.95 si ha k=1.96.

C.5.2 Presentazione sintetica degli errori campionari

Ad ogni stima $\hat{Y}_d$ corrisponde un errore di campionamento relativo $\hat{\varepsilon}(\hat{Y}_d)$; ciò significa che per consentire una lettura corretta delle tabelle pubblicate sarebbe necessario presentare per ogni stima pubblicata il corrispondente errore di campionamento relativo. Ciò, tuttavia, non è possibile sia per limiti di tempo e di costi di elaborazione, sia perché le tavole della pubblicazione risulterebbero appesantite e di non facile consultazione per l'utente finale. Inoltre, non sarebbero comunque disponibili gli errori delle stime non pubblicate, che l'utente può ricavare in modo autonomo.

Per le ragioni sopra esposte, si ricorre frequentemente ad una presentazione sintetica degli errori relativi, basata sul *metodo dei modelli regressivi*. Questo metodo si basa sulla determinazione di una funzione matematica che mette in relazione ciascuna stima con il proprio errore relativo.

Nella presente indagine, il modello utilizzato per le stime di frequenze assolute e relative, è del tipo seguente:

$$\log\left(\hat{\varepsilon}^2(\hat{Y}_d)\right) = a + b\log(\hat{Y}_d) \qquad (11)$$

dove i parametri a e b vengono stimati utilizzando il metodo dei minimi quadrati.

Nel prospetto C.2 sono riportati i valori dei coefficienti a e b e dell'indice di determinazione R^2 del modello utilizzato per l'interpolazione degli errori campionari di stime di frequenze assolute e relative, per totale Italia, ripartizione geografica e tipologia comunale.

Sulla base delle informazioni contenute in tale prospetto, è possibile calcolare la stima dell'errore di campionamento relativo di una determinata stima di frequenza assoluta $\hat{Y}_d$ mediante la formula:

$$\hat{\varepsilon}(\hat{Y}_d) = \sqrt{\exp\left(a + b\log(\hat{Y}_d)\right)} \qquad (12)$$

che si ricava facilmente dalla (11).

Se, per esempio, la stima $\hat{Y}_d$ si riferisce alle persone dell'Italia Nord Occidentale, l'errore relativo corrispondente si ottiene introducendo nella (12) i valori dei parametri a e b riportati nella seconda riga del prospetto C.2 (a = 9,51202, b = -1,07702).

Il prospetto C.3, presentato in aggiunta, consente di rendere più agevole il calcolo degli errori campionari. Esso riguarda gli individui ed ha la seguente struttura: a) in fiancata sono elencati i valori crescenti di stima (20.000, 30.000, …, 25.000.000); b) le colonne successive contengono gli errori di campionamento relativo, per ciascun dominio territoriale di interesse, calcolati mediante la formula (12), corrispondenti alle stime di frequenze assolute della prima colonna.

Le informazioni contenute in tali prospetti permettono di calcolare l'errore relativo di una generica stima di frequenza assoluta (o relativa) mediante due procedimenti che risultano di facile applicazione, anche se conducono a risultati meno precisi di quelli ottenibili mediante l'espressione (12). Il primo metodo consiste nell'individuare, nella prima colonna del prospetto, il livello di stima che più si avvicina alla stima di interesse e nel considerare come errore relativo il valore che si trova sulla stessa riga, nella colonna corrispondente al domino territoriale di riferimento.

Con il secondo metodo, l'errore campionario della stima $\hat{Y}_d$ si ricava mediante la seguente espressione:

$$\hat{\varepsilon}(\hat{Y}_d) = \hat{\varepsilon}(\hat{Y}_d^{k-1}) - \frac{\hat{\varepsilon}(\hat{Y}_d^{k-1}) - \hat{\varepsilon}(\hat{Y}_d^{k})}{\hat{Y}_d^{k} - \hat{Y}_d^{k-1}}(\hat{Y}_d - \hat{Y}_d^{k-1}) \qquad (13)$$

dove $\hat{Y}_d^{k-1}$ e $\hat{Y}_d^k$ sono i valori delle stime, riportati nella prima colonna, entro i quali è compresa la stima di interesse $\hat{Y}_d$, ed $\hat{\varepsilon}(\hat{Y}_d^{k-1})$ e $\hat{\varepsilon}(\hat{Y}_d^k)$ i corrispondenti errori relativi.

Prospetto C.2 - Valori dei coefficienti a, b e dell'indice di determinazione R^2 (%) delle funzioni utilizzate per le interpolazioni degli errori campionari delle stime riferite agli individui per totale Italia e ripartizione geografica

ZONA TERRITORIALE	PERSONE		
	a	b	R^2 (%)
ITALIA	**10,38642**	**-1,13804**	**96,8**
RIPARTIZIONI GEOGRAFICHE (a)			
Nord-ovest	9,51202	-1,07702	95,3
Nord-est	10,14542	-1,15872	95,5
Centro	10,11610	-1,13938	95,2
Sud	10,10136	-1,13911	93,9
Isole	10,03172	-1,11533	95,1

(a) Italia nord-occidentale: Piemonte, Valle d'Aosta, Lombardia, Liguria; Italia nord-orientale: Bolzano, Trento, Veneto, Friuli-Venezia Giulia, Emilia Romagna; Italia centrale: Toscana, Umbria, Marche, Lazio; Italia meridionale: Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria; Italia insulare: Sicilia, Sardegna.

(b) Comuni tipo A1: Area urbana centro; Tipo A2: Area urbana periferia; Tipo B1: comuni fino a 2.000 abitanti; Tipo B2: da 2.001 a 10.000 abitanti; Tipo B3: da 10.001 a 50.000 abitanti; Tipo B4: oltre 50.000 abitanti.

Prospetto C.3 - Valori interpolati degli errori campionari relativi percentuali delle stime riferite agli individui per totale Italia e ripartizione geografica

STIME	Italia	Nord-ovest	Nord-est	Centro	Sud	Isole
20.000	64,3	56,2	51,4	55,8	55,4	60,2
50.000	38,2	34,3	30,2	33,1	32,9	36,1
60.000	34,4	31,1	27,2	29,8	29,7	32,6
70.000	31,5	28,6	24,9	27,3	27,2	30,0
80.000	29,2	26,6	23,0	25,3	25,2	27,8
90.000	27,3	25,0	21,5	23,7	23,5	26,0
100.000	25,7	23,6	20,2	22,3	22,2	24,5
200.000	17,3	16,3	13,5	15,0	14,9	16,7
300.000	13,8	13,1	10,7	11,9	11,9	13,3
400.000	11,7	11,2	9,1	10,1	10,1	11,3
500.000	10,3	9,9	8,0	8,9	8,9	10,0
750.000	8,2	8,0	6,3	7,1	7,0	8,0
1.000.000	6,9	6,8	5,3	6,0	6,0	6,8
2.000.000	4,7	4,7	3,6	4,0	4,0	4,6
3.000.000	3,7	3,8	2,8	3,2	3,2	3,7
4.000.000	3,2	3,2	2,4	2,7	2,7	3,1
5.000.000	2,8	2,9	2,1	2,4	2,4	2,8
7.500.000	2,2	2,3	1,7	1,9	1,9	2,2
10.000.000	1,9	2,0	1,4	1,6	1,6	1,9
15.000.000	1,5	1,6	1,1	1,3	1,3	1,5