Nota metodologica

1. Finalità e principali caratteristiche dell'indagine

Nel 2011-2012 l'Istat ha condotto, per la prima volta, la rilevazione statistica sulla "Condizione e integrazione sociale dei cittadini stranieri" attraverso la quale sono state rilevate informazioni su numerosi aspetti della vita e del processo di integrazione dei cittadini stranieri in Italia.

L'indagine è stata supportata con l'interesse e il contributo finanziario di varie istituzioni, tra le quali il Ministero dell'interno- Dipartimento per le libertà civili- Direzione centrale per le politiche dell'immigrazione e dell'asilo che ha stipulato con l'Istat una convenzione nell'ambito degli interventi finanziati attraverso il Fondo europeo per l'integrazione dei cittadini extracomunitari, per la realizzazione di un progetto volto a cogliere gli aspetti fondamentali del processo d'integrazione degli stranieri, tra cui la dimensione linguistica. Nell'ambito del progetto, inoltre, è stato introdotto un ampliamento campionario finalizzato ad approfondire le realtà territoriali dei comuni metropolitani di Milano, Roma e Napoli.

Oltre al tema della lingua, l'indagine affronta temi come la famiglia, i figli, i percorsi formativi, la storia migratoria, la storia lavorativa, l'attuale condizione di lavoro, gli stili di vita, le condizioni di salute e il ricorso ai servizi sanitari, l'appartenenza religiosa, le reti e le relazioni sociali, la partecipazione politica e sociale, le esperienze di discriminazione vissute, la sicurezza, le esperienze di vittimizzazione subite, le condizioni abitative. Per la varietà e ricchezza dei temi trattati, l'indagine segna un passaggio rilevante da parte della statistica ufficiale sulla conoscenza della presenza straniera in Italia, allargando il panorama delle informazioni disponibili anche su comportamenti, atteggiamenti e opinioni dei cittadini stranieri, a integrazione e complemento dei dati di fonte amministrativa, correntemente prodotti dall'Istat.

La rilevazione è stata condotta attraverso interviste dirette con tecnica *Computer assisted personal interview* (Capi) su un campione nazionale di circa 9.600 famiglie con almeno un cittadino straniero, residenti in 833 comuni italiani, per un totale di circa 21 mila stranieri residenti intervistati. Sono state, invece, circa 4 mila le famiglie interessate dal campione ampliato su Roma, Milano e Napoli. In ciascuna famiglia campione, individuata secondo specifiche regole di eleggibilità individuate per cogliere la molteplicità di famiglie con stranieri, sono stati intervistati gli individui, di qualunque età, di cittadinanza straniera oppure italiana per acquisizione (cittadini italiani al momento della rilevazione, ma stranieri alla nascita). Non sono stati intervistati, invece, gli individui italiani dalla nascita perché non appartengono alla popolazione di riferimento della rilevazione. Per questi, tuttavia, sono state raccolte informazioni sulle caratteristiche strutturali di tipo sociodemografico in modo da disporre di informazioni anche sulla composizione e stili di vita delle famiglie miste, cioè composte da italiani e stranieri. Gli individui al di sotto dei 14 anni sono stati intervistati in *proxy*, cioè per tramite di un genitore o componente maggiorenne della famiglia. Per facilitare il livello di comprensione delle domande, il questionario è stato tradotto in 10 lingue.

Altre informazioni sull'indagine sono consultabili sul sito web dell'Istat alla pagina http://www.istat.it/it/archivio/10825.

2. Strategia di campionamento

2.1 Disegno di campionamento

Il disegno campionario che è stato studiato per l'indagine sugli stranieri presenta le caratteristiche generali dei disegni utilizzati per le indagini Istat sulle famiglie condotte con intervista diretta e selezione dalle anagrafi comunali.

In particolare, si tratta di un disegno a due stadi di selezione, dove le unità di primo stadio sono i comuni e le unità di secondo stadio sono le famiglie. I comuni sono stratificati per regione e tipologia comunale e selezionati

con probabilità proporzionale alla loro popolazione straniera residente. Per garantire che le diverse nazionalità siano opportunamente rappresentate nel campione di comuni estratti al primo stadio, si è studiata la possibilità di procedere a una selezione bilanciata dei comuni sulla base delle nazionalità presenti.

La lista di selezione disponibile per le unità di primo stadio è l'archivio dei comuni italiani, che contiene per ciascun comune il numero degli stranieri residenti per sesso e cittadinanza.

Nella progettazione dello schema di selezione dei comuni si è dovuto tenere conto del fatto che la popolazione degli stranieri residenti presenta una distribuzione molto disomogenea sul territorio, concentrata soprattutto nelle regioni del Centro-Nord. In alcuni comuni non sono presenti stranieri (circa 80 comuni) e molti altri ne hanno in numero molto esiguo. Dal momento che nella progettazione di un disegno campionario a due stadi è necessario fissare il numero minimo di interviste da effettuare in ogni comune e tenendo anche conto dell'esigenza di prevedere un adeguato numero di famiglie sostitutive per le famiglie non rispondenti, si è stabilito di individuare una soglia al di sotto della quale un comune viene escluso dalla lista di selezione. Inoltre, la distribuzione molto disuguale della presenza straniera nei comuni del Centro-nord rispetto a quella del Sud e Isole ha fatto anche propendere per una soglia differenziata per area geografica.

Nella tabella 1 è riportata la copertura della popolazione di stranieri residenti che si ottiene includendo nella lista di selezione i comuni con popolazione superiore a diverse soglie pari a 50, 75, 80, 90 e 100 stranieri.

Sulla base dell'analisi di tale tabella, si è stabilito di differenziare la soglia per area geografica, dal momento che la soglia ottimale per il Centro-nord (intorno alle 100 unità) avrebbe comportato una copertura non accettabile per il Sud e le Isole (inferiore al 90%).

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 quattro aree che identificano la tipologia comunale[1], i comuni italiani sono suddivisi in due sottoinsiemi sulla base della popolazione straniera residente:

- l'insieme dei comuni Auto rappresentativi (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 sé 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 stranieri 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.

[1] La tipologia comunale è ottenuta suddividendo i comuni italiani in quattro classi formate in base a caratteristiche socio-economiche e demografiche: A1) comuni centro dell'area metropolitana: Torino, Milano, Venezia, Genova, Bologna, Firenze, Roma, Napoli, Bari, Palermo, Catania, Cagliari; A2) comuni che gravitano intorno ai comuni centro dell'area metropolitana; B1) comuni non appartenenti all'area metropolitana aventi fino a 10 mila abitanti; B2) comuni non appartenenti all'area metropolitana con oltre 10 mila abitanti.

Tabella 1. Distribuzione per regione, ripartizione e tipologia comunale della popolazione straniera e copertura % in termini di popolazione dei comuni con più di 50, 75, 80, 90, 100 stranieri

Livello territoriale	Numero comuni	Popolazione stranieri totale	% di copertura in corrispondenza delle soglie di popolazione straniera residente nei comuni				
			50	75	80	90	100
PIEMONTE	1.184	310.543	96,1	93,4	92,9	91,8	90,9
VALLE D'AOSTA	73	6.604	87,4	79,8	79,8	75,9	73,1
LOMBARDIA	1.533	815.335	99,1	98,2	98,0	97,5	97,2
BOLZANO	114	32.945	97,4	93,9	93,0	91,9	91,4
TRENTO	223	37.889	93,4	88,3	87,5	85,7	83,2
VENETO	581	403.985	99,7	99,2	99,1	98,9	98,6
FRIULI-VENEZIA GIULIA	218	83.306	98,0	96,5	96,1	95,3	94,7
LIGURIA	234	90.881	97,9	95,5	94,9	94,1	93,9
EMILIA-ROMAGNA	341	365.687	99,8	99,7	99,6	99,5	99,3
TOSCANA	287	275.149	99,8	99,6	99,5	99,2	99,1
UMBRIA	92	75.631	99,7	99,5	99,3	98,9	98,4
MARCHE	246	115.299	99,3	97,8	97,7	97,3	96,5
LAZIO	377	390.993	99,4	98,8	98,6	98,3	98,1
ABRUZZO	301	59.749	95,0	91,7	90,5	89,5	88,3
MOLISE	129	6.271	70,3	63,3	63,3	60,5	59,0
CAMPANIA	550	114.792	94,8	91,7	91,3	90,3	89,6
PUGLIA	257	63.868	96,9	94,7	93,4	91,6	90,7
BASILICATA	131	9.595	84,6	74,9	74,1	71,5	69,5
CALABRIA	406	50.871	90,4	86,5	85,9	84,0	82,2
SICILIA	389	98.152	96,2	93,9	93,4	92,1	91,0
SARDEGNA	357	25.106	85,1	79,9	79,3	77,6	74,5
NORD-OVEST	3.024	1.223.363	98,2	96,7	96,4	95,7	95,3
NORD-EST	1.477	923.812	99,2	98,5	98,3	98,0	97,7
CENTRO	1.002	857.072	99,5	99,0	98,8	98,5	98,2
SUD-EST	687	129.888	94,7	91,8	90,6	89,1	88,0
SUD-OVEST	1.087	175.258	92,9	89,2	88,8	87,5	86,3
ISOLE	746	123.258	94,0	91,0	90,5	89,1	87,7
Comuni metropolitani	12	679.551	100,0	100,0	100,0	100,0	100,0
Cintura metropolitana	483	342.367	99,7	99,2	99,0	98,7	98,5
Comuni fino a 10.000 ab.	6.586	936.155	93,8	89,3	88,3	86,4	84,9
Comuni oltre 10.000 ab.	942	1.474.578	100,0	100,0	99,9	99,9	99,8
ITALIA	**8.023**	**3.432.651**	**98,3**	**97,0**	**96,7**	**96,1**	**95,6**

2.1.1 Stratificazione e selezione delle unità campionarie

L'obiettivo della stratificazione è quello di formare gruppi (o strati) di unità caratterizzate, relativamente alle variabili oggetto d'indagine, da massima omogeneità interna agli strati e massima eterogeneità fra gli strati. Il raggiungimento di tale obiettivo si traduce in termini statistici in un guadagno nella precisione delle stime, ossia in una riduzione dell'errore campionario a parità di numerosità campionaria.

Nell'indagine in oggetto, i comuni vengono stratificati in base alla loro dimensione in termini di stranieri residenti e nel rispetto delle seguenti condizioni:

- autoponderazione del campione a livello regionale;
- scelta di un numero minimo di famiglie da intervistare in ciascun comune campione;
- scelta del numero, $\overline{n}$, di comuni campione da estrarre da ciascuno strato Nar: tale parametro è stato posto pari a 3;
- formazione di strati aventi ampiezza approssimativamente costante in termini di popolazione residente.

Il procedimento di stratificazione, attuato all'interno di ogni dominio territoriale individuato dalle quattro aree A_1, A_2, B_1 e B_2 di ciascuna regione geografica, si articola nelle seguenti fasi:

- ordinamento dei comuni del dominio in ordine decrescente secondo la loro dimensione demografica in termini di popolazione straniera residente;
- determinazione di una soglia di popolazione per la definizione dei comuni Ar, mediante la relazione:

$$_r\lambda = \frac{_r\overline{m} \; _r\delta}{_r f}$$

in cui per la generica regione geografica r si è indicato con: $_r\overline{m}$ il numero minimo di famiglie da intervistare in ciascun comune campione; $_r\delta$ il numero medio di componenti per famiglia (nel caso specifico è la dimensione media delle famiglie di stranieri); $_r f$ la frazione di campionamento, definita dal rapporto tra la dimensione campionaria e la popolazione straniera;

- suddivisione di tutti i comuni nei due sottoinsiemi Ar e Nar: i comuni di dimensione superiore o uguale a $_r\lambda$ sono definiti come comuni Ar e i rimanenti come Nar;
- suddivisione dei comuni dell'insieme Nar in strati aventi dimensione, in termini di popolazione straniera residente, approssimativamente costante e all'incirca pari $\overline{n}$ volte la soglia $_r\lambda$.

Effettuata la stratificazione, i comuni Ar sono inclusi con certezza nel campione; per quanto riguarda, invece, i comuni Nar, nell'ambito di ogni strato vengono estratti tre comuni campione con probabilità proporzionale alla dimensione demografica e seguendo lo schema di selezione bilanciata descritto nel paragrafo seguente.

Il numero minimo di interviste per comune, $_r\overline{m}$, è stato posto a 10 per le regioni del Centro-nord e a 8 per le regioni del Sud e delle Isole. Sulla base di questa scelta e tenendo in considerazione l'esigenza di disporre di un numero sufficiente di famiglie per le sostituzioni, la soglia di popolazione straniera per l'inclusione dei comuni è stata fissata a 100 per le regioni del Centro-nord e 80 per le regioni del Sud e delle Isole. In tal modo l'universo di selezione è costituito da 4.033 comuni, che garantiscono una copertura della popolazione degli stranieri residenti di circa il 96%.

2.1.2 Selezione dei comuni bilanciata rispetto alle nazionalità

Per l'estrazione dei comuni all'interno degli strati è stata studiata una *selezione bilanciata*, allo scopo di conseguire una maggiore rappresentatività delle nazionalità straniere presenti sul territorio in modo molto disomogeneo. Si è cercato di tenere conto in tal modo della distribuzione della presenza straniera nei comuni in termini di nazionalità, sebbene non fosse possibile effettuare una stratificazione per nazionalità. In altri termini, si è posto un vincolo sulla distribuzione risultante dei comuni selezionati, realizzando un campione bilanciato (Deville e Tillè 2004)[2].

[2] Deville, J.C. and Tillé, Y. (2004). *Efficient Balanced Sampling: The Cube Method. Biometrika*, 91, 893-912.

In particolare, in ogni generica area geografica *a*, $m_{(a)}$ comuni campione sono selezionati dagli $M_{(a)}$ comuni universo mediante un campione bilanciato, con probabilità di inclusione definite all'interno di ciascuno strato in modo proporzionale alla popolazione straniera residente. Le equazioni di bilanciamento impongono che le stime dirette dei totali di popolazione relativi alle *N* prefissate nazionalità presenti nei comuni coincidano con i corrispondenti totali noti; in simboli:

$$\sum_{c=1}^{m_{(a)}} \frac{\mathbf{x}_c}{\pi_c} = \sum_{c=1}^{M_{(a)}} \mathbf{x}_c$$

in cui π_c è la probabilità di inclusione del comune *c* e

$$\mathbf{x}'_c = \left({}_1P_c \ ,..., {}_n P_c \ ,..., {}_N P_c \ , \pi_c \right)$$

è il vettore di variabili ausiliarie riferito al generico comune *c*, in cui ${}_nP_c$ indica il numero di stranieri nazionalità *n* residenti nel comune *c*, nota dall'archivio dei comuni.

Poiché non era possibile tenere in considerazione tutte le nazionalità, anche quelle con pochissime presenze sul territorio nazionale, è stato scelto di realizzare un bilanciamento basato sulle nazionalità più numerose individuate a livello nazionale. Dopo un'analisi delle possibilità concrete di estrazione di un campione bilanciato, si è scelto di procedere ad un bilanciamento rispetto alle prime 15 nazionalità individuate a livello nazionale, che coprono circa il 72% della popolazione straniera totale, vincolate su tre ripartizioni geografiche (Nord, Centro, Sud e Isole).

2.1.3 Secondo stadio di campionamento: selezione delle famiglie

Una volta estratti i comuni campione, al secondo stadio di campionamento vengono selezionate le famiglie campione dalla lista delle famiglie in cui è presente almeno uno straniero residente. Tutti gli individui stranieri appartenenti a tali famiglie vengono intervistati poiché rappresentano la popolazione di riferimento. Gli individui con cittadinanza italiana, invece, non vengono intervistati poiché non appartengono alla popolazione di riferimento; per loro vengono rilevati soltanto dati relativi alle loro caratteristiche strutturali di tipo socio-demografico (sesso, età, cittadinanza, stato di nascita, titolo di studio, eccetera) che sono considerate come "attributi" degli stranieri intervistati.

È stata, inoltre, prevista la sostituzione delle famiglie non rispondenti mediante la costituzione di quartine di nominativi campione: ad ogni unità campione sono quindi state associate tre unità di riserva. La costituzione di quartine di nominativi campione consente di garantire che la numerosità campionaria individuata in fase di progettazione sia rispettata fino alla conclusione dell'indagine, evitando così che a seguito delle 'cadute' dei nominativi campione inizialmente estratti (per rinuncia delle famiglie a collaborare, trasferimenti delle famiglie in altri comuni o all'estero, errori di lista, ecc.) la dimensione campionaria effettiva, risultante a fine indagine, sia inferiore a quella teorica, individuata in fase di progettazione. Per ridurre il rischio di effetti distorsivi del campione, nella formazione delle quartine si è proceduto secondo un criterio di omogeneità per cittadinanza (del capofamiglia), per alfabetico di via e per dimensione anagrafica della famiglia.

2.2 La numerosità campionaria

La numerosità campionaria in termini di famiglie prevista per l'indagine è di circa 12 mila famiglie ed è stato definito un disegno campionario a due stadi, secondo le modalità sopra descritte, che ha portato alla selezione di 879 comuni campione.

La numerosità campionaria è tale da garantire l'attendibilità di stime di frequenze a livello dei domini di stima pianificati, ovvero l'intero territorio nazionale e le sei ripartizioni geografiche definite, per questa indagine, considerando separatamente il Sud-est e il Sud-ovest per le peculiarità della popolazione immigrata in queste aree; i domini pianificati di stima sono pertanto:

- Italia;
- Nord-ovest (Piemonte, Valle d'Aosta, Lombardia, Liguria)
- Nord-est (Trento, Bolzano, Veneto, Friuli-Venezia Giulia, Emilia-Romagna)
- Centro (Toscana, Umbria, Marche, Lazio)
- Sud-ovest (Campania, Basilicata e Calabria)
- Sud-est (Abruzzo, Molise e Puglia)
- Isole (Sicilia, Sardegna).

Si è inoltre tenuto sotto controllo il dominio di stima definito dalla tipologia comunale, aggregata nelle seguenti quattro modalità:

- A1, comuni metropolitani,
- A2, cintura dei comuni metropolitani,
- A3, altri comuni fino a 10 mila abitanti,
- A4, altri comuni con più di 10 mila abitanti.

È bene precisare che il livello territoriale per il quale è possibile produrre stime attendibili dipende dal livello delle stime stesse e dalla procedura di stima che è possibile mettere in atto sulla base dei risultati conseguiti sul campo.

Nella tabella 2 è illustrata l'allocazione del campione di famiglie e individui tra le regioni, le ripartizioni e le tipologie comunali relativamente alla popolazione degli stranieri residenti al 1° gennaio 2008. Poiché nell'archivio dei comuni le informazioni sugli stranieri residenti sono solamente in termini di individui, per calcolare il numero di famiglie sono stati utilizzati i dati sul numero di famiglie con almeno uno straniero residente, desunti dai bilanci demografici relativi agli stranieri residenti.

Tabella 2. Allocazione del campione tra i domini territoriali

Dominio territoriale	Popolazione straniera (*)	Campione FAMIGLIE
Piemonte	310.543	579
Valle D'Aosta	6.604	142
Lombardia	815.335	1.358
Bolzano	32.945	166
Trento	37.889	145
Veneto	403.985	779
Friuli-Venezia Giulia	83.306	237
Liguria	90.881	288
Emilia-Romagna	365.687	751
Toscana	275.149	666
Umbria	75.631	256
Marche	115.299	311
Lazio	390.993	1.008
Abruzzo	59.749	759
Molise	6.271	210
Campania	114.792	1.287
Puglia	63.868	742
Basilicata	9.595	209
Calabria	50.871	582
Sicilia	98.152	1.281
Sardegna	25.106	514
NORD-OVEST	1.223.363	2.367
NORD-EST	923.812	2.078
CENTRO	857.072	2.240
SUD-EST	129.888	1.711
SUD-OVEST	175.258	2.079
ISOLE	123.258	1.794
Comuni metropolitani	679.551	2.940
Cintura metropolitana	342.367	2.544
Comuni fino a 10.000 ab.	936.155	2.826
Comuni oltre 10.000 ab.	1.474.578	3.495
ITALIA	**3.432.651**	**12.269**

(*) Con riferimento all'anno 2009.

L'allocazione del campione è avvenuta in due fasi.

1. Allocazione del campione tra le sei ripartizioni geografiche. È stata definita un'allocazione del campione tra le ripartizioni geografiche in un'ottica di compromesso tra l'allocazione uniforme e l'allocazione proporzionale, attribuendo un peso pari a 0,8 a quella uniforme e 0,2 a quella proporzionale, nell'ottica di privilegiare l'attendibilità delle stime a livello di ripartizione.
2. Allocazione del campione delle ripartizioni tra le regioni. All'interno di ciascuna ripartizione l'allocazione tra le regioni è stata ottenuta nella stessa ottica della prima fase, ma con pesi invertiti, cioè attribuendo un peso pari a 0,2 all'allocazione uniforme e 0,8 a quella proporzionale. In tal modo le regioni con un numero esiguo di stranieri residenti si vedono assegnato un campione molto piccolo.

2.3 Disegno di campionamento e numerosità del campione ampliamento Milano, Roma, Napoli

Sul piano del disegno di campionamento, l'ampliamento non modifica le scelte effettuate per l'intera indagine nazionale. Di fatto, esso consiste in una maggiore numerosità di unità campionarie (famiglie con almeno uno

straniero) da intervistare nei tre comuni di interesse, garantendo quindi una rappresentatività del campione a livello di singolo comune.
Il numero complessivo di interviste aggiuntive è stato fissato in 4 mila famiglie, distribuite tra i tre comuni in modo tale da raggiungere, a livello comunale, una precisione delle stime basate sul campione complessivo dello stesso livello, garantito per le ripartizioni geografiche dalla dimensione del campione dell'indagine nazionale.
In particolare, nella tabella 3 sono riportare le dimensioni del campione base, dell'ampliamento e del campione finale nei comuni di Milano, Roma e Napoli.

Tabella 3. Dimensione campionaria dei comuni di Milano, Roma e Napoli

Comune	Stranieri residenti	Numerosità del campione nazionale (famiglie)	Numerosità del campione ampliamento (famiglie)	Numerosità del campione finale (famiglie)
Milano	199.322	301	1.469	1.770
Roma	268.996	574	946	1.520
Napoli	27.481	264	1.585	1.849
Totale	**495.799**	**1.139**	**4.000**	**5.139**

3. Livello di precisione delle stime

3.1 Le stime campionarie

L'indagine deve produrre le stime riferite al numero di individui che nella popolazione di riferimento possiedono una certa caratteristica o il livello di una quantità misurata sugli individui. Per il calcolo dei coefficienti di riporto all'universo si utilizza una procedura generalizzata di stima, basata sull'uso di una famiglia di stimatori, noti in letteratura come *calibration estimator* (stimatori di ponderazione vincolata). La 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, e correlati alle principali variabili oggetto di indagine.
La famiglia di stimatori di ponderazione vincolata coincide asintoticamente con lo stimatore di regressione generalizzato: per campioni sufficientemente grandi, quindi, tali stimatori hanno approssimativamente le stesse proprietà, ovvero sono corretti, consistenti e con la stessa varianza campionaria [3].
La strategia adottata per la costruzione dei coefficienti di riporto all'universo si sviluppa attraverso le fasi tipiche utilizzate per la costruzione degli stimatori nelle varie indagini campionarie dell'Istituto. In particolare possiamo distinguere:

- la determinazione della probabilità di inclusione di ogni unità statistica e del relativo peso diretto, pari all'inverso della probabilità di inclusione;
- calcolo dei coefficienti di correzione per mancata risposta totale;
- determinazione dei coefficienti di riporto all'universo finali vincolati ai totali noti desunti da fonti esterne all'indagine.

3.2 La probabilità di inclusione e il peso diretto

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 stesso. A tale scopo, ad ogni

[3] La metodologia è illustrata da Deville, J.C. e Särndal, C.E. in *Calibration Estimation in Survey Sampling*, Journal of the American Statistical Association, Vol. 87, n.418, 1992.

unità campionaria viene attribuito un peso, o coefficiente di riporto all'universo, che indica quante unità della popolazione sono rappresentate, rispettivamente, da ogni unità presente nel campione.
Senza perdere di generalità, definiamo la seguente simbologia:

- U popolazione di riferimento oggetto di indagine;
- y_k valore della variabile Y assunto dalla k-esima osservazione della popolazione;
- y_j valore della variabile Y assunto dalla j-esima osservazione della popolazione;
- πj probabilità, assegnata dal disegno di campionamento, che l'unità j-esima sia inclusa nel campione S.

Il totale di una generica variabile *Y*, calcolato sull'intera popolazione, assume la seguente forma:

$$Y = \sum_{k \in U} y_k \qquad (1)$$

Il disegno di campionamento assegna le probabilità di inclusione ad ogni unità del campione in modo tale che

$$\hat{Y} = \sum_{j \in s} y_j \frac{1}{\pi_j} \qquad (2)$$

sia uno stimatore corretto della (1).

Nel disegno di campionamento di questa indagine, la probabilità di inclusione di un generico individuo è data: dalla probabilità di estrazione del comune si residenza (direttamente proporzionale all'ampiezza demografica dei comuni all'interno dello strato); e dalla probabilità di estrazione della famiglia di appartenenza tra le famiglie eleggibili del comune.
Per una generica famiglia eleggibile *j*, nel comune *i* dello strato *h*, il peso diretto d_{hij}, inverso della probabilità di inclusione π_{hij}, assume la seguente forma:

$$d_{hij} = \frac{1}{\pi_{hij}} = \frac{1}{c_h} \frac{P_h}{P_{hi}} \frac{M_{hi}}{m_{hi}} \qquad (3)$$

dove :

- h denota l'indice di strato;
- i è l'indice di comune;
- j denota l'indice della famiglia;
- c_h indica il numero di comuni campione dello strato *h;*
- P_h indica il totale della popolazione residente nello strato *h*;
- P_{hi} il totale della popolazione residente nel comune *i* dello strato *h*;
- M_{hi} indica il totale di famiglie eleggibili nel comune *i* dello strato *h*;
- m_{hi} indica il numero di famiglie campione nel comune *i* dello strato *h.*

Nei comuni di Roma, Milano e Napoli, oggetto di specifico ampliamento campionario, la probabilità di inclusione di un generico individuo è data semplicemente dalla probabilità di estrazione della famiglia di appartenenza tra le famiglie eleggibili del comune (pari quindi al numero di famiglie estratte e numero di famiglie totali nel comune). In questo caso, quindi, per una generica famiglia eleggibile *j*, nel comune *i*, il peso diretto d_{ij}, inverso della probabilità di inclusione π_{ij}, assume quindi la seguente forma:

$$d_{ij} = \frac{1}{\pi_{ij}} = \frac{M_i}{m_i} \qquad (3 \text{ bis})$$

dove :

- i è l'indice di comune;
- j denota l'indice della famiglia;
- M_i indica il totale di famiglie eleggibili nel comune *i*;
- m_i indica il numero di famiglie campione nel comune *i.*

3.3 La correzione per mancata risposta

Nel corso della fase di raccolta delle informazioni presso le unità che formano il campione, come accade per tutte le indagini statistiche, alcune di queste si trovano nell'impossibilità di partecipare all'indagini. Questo aspetto comporta che al termine della rilevazione, la numerosità campionaria teorica individuata nel disegno e quella effettiva differiscono numericamente (mancata risposta totale). Nell'indagine in questione, l'utilizzo delle quartine[4] ha fatto sì che il problema della mancata risposta totale si mantenesse a livelli bassi: il campione finale è infatti pari a 9.553 unità rispetto alle 12.269 previste in fase di definizione del disegno nel caso del campione nazionale; a 3.655 anziché 5.139 nel caso dell'ampliamento. Per ovviare alla mancata partecipazione di alcune unità del campione teorico all'indagine, nella fase di calcolo dei coefficienti di riporto all'universo viene introdotto, come di norma, un correttore per mancata risposta che, sotto l'ipotesi che il comportamento dei rispondenti sia simile a quello dei non rispondenti all'interno dello stesso strato, assume la forma dell'inverso del tasso di risposta (δ_h):

$$\frac{1}{\delta_h} = \frac{m_h}{m_h^r} \qquad (5)$$

in cui m_h^r rappresenta il numero di famiglie rispondenti nello strato *h.*

In questa maniera, il coefficiente di riporto all'universo corretto per mancata risposta, da assegnare al campione rispondente, risulta essere:

$$k_{hij} = d_{hij}\frac{1}{\delta_h} = \frac{1}{c_h}\frac{P_h}{P_{hi}}\frac{M_{hi}}{m_{hi}}\frac{m_h}{m_h^r} \qquad (6)$$

Nel caso dell'ampliamento per Roma, Milano e Napoli, essendo gli strati coincidenti con i comuni stessi, la correzione per mancata risposta assume la forma semplificata:

$$\frac{1}{\delta_i} = \frac{m_i}{m_i^r} \qquad (5\text{ bis})$$

in cui m_i^r rappresenta il numero di famiglie rispondenti nel comune i.

Di conseguenza, il coefficiente di riporto corretto per mancata risposta assume la forma:

$$k_{ij} = d_{ij}\frac{1}{\delta_i} = \frac{M_i}{m_i}\frac{m_i}{m_i^r} \qquad (6\text{ bis})$$

[4] Per ogni famiglia estratta ne sono state estratte altre tre di profilo simile che potessero eventualmente sostituirla. Una unità campionaria ha dato luogo ad una mancata risposta totale, vale a dire non ha contribuito all'indagine, se nessuna delle famiglie della quartina ha realizzato l'intervista (per irreperibilità, rifiuto, interruzione definitiva, eccetera).

3.4 La calibrazione a fonti esterne

Per il calcolo dei coefficienti di riporto all'universo finali si adottano gli stimatori *calibration estimator*. La metodologia si basa sull'utilizzo di opportune informazioni ausiliarie, sintetizzate in totali noti, che, correlate con le variabili principali oggetto di indagine, hanno la funzione di aumentare l'accuratezza delle stime. I pesi finali si ottengono risolvendo un problema di minimo vincolato, in cui la funzione da minimizzare è una funzione di distanza tra i pesi diretti corretti per la mancata risposta (**k**) e i pesi finali (**w**) delle famiglie del campione rispondente (S^r), e i vincoli sono proprio le condizioni di uguaglianza delle stime campionarie di alcune variabili ausiliarie con i rispettivi totali noti desunti da fonti esterne all'indagine[5].

$$\begin{cases} Min\left\{\sum_{j\epsilon S^r} dist(k_j, w_j)\right. \\ \sum_{j\epsilon S^r} \boldsymbol{x}_j * w_j = \boldsymbol{t} \end{cases}$$

dove **t** è il vettore dei totali noti e x_j è il vettore delle variabili ausiliarie osservate sulla j-esima unità campionaria appartenente al campione rispondente (S^r). La funzione di distanza utilizzata è la logaritmica troncata.

I totali noti introdotti come vincoli nel calcolo dei pesi finali consentono di migliorare l'accuratezza delle stime, poichè quanto più le variabili ausiliarie considerate sono correlate con le variabili oggetto d'indagine, tanto più si riduce la distorsione delle stime. Nello specifico dell'indagine Istat "Condizione e integrazione sociale dei cittadini stranieri residenti in Italia" le stime campionarie sono state vincolate ai seguenti totali noti:

- stranieri residenti in Italia secondo le 15 cittadinanze più rappresentate (Romania, Albania, Marocco, Repubblica Popolare Cinese, Ucraina, Filippine, Tunisia, Polonia, India, Moldavia, Macedonia- ex Rep. Jugoslavia, Equador, Perù, Egitto, Bangladesh)
- stranieri residenti per ripartizione (Nord–ovest, Nord–est, Centro, Sud–ovest, Sud–est, Isole)) e 8 gruppi di cittadinanze (Unione europea; Europa Centro-orientale; Africa settentrionale; Africa occidentale; Asia Centro-meridionale; Asia orientale; America Centro-meridionale; altro)
- stranieri residenti per ripartizione, sesso e classi d'età (0-5, 6-15, 16-24, 25-34, 35-44, 45-64, 65 e più);
- stranieri residenti per ripartizione e tipologia del comune di residenza (comuni metropolitani, comuni periferia dei comuni metropolitani, altri comuni fino a 10 mila abitanti , altri comuni con più di 10 mila abitanti).

Lo stesso criterio di calibrazione è stato adottato per il campione ampliato nei comuni di Milano, Roma, Napoli e le stime campionarie , in questo caso, risultano vincolate ai seguenti totali noti:

- stranieri residenti in ciascuno dei tre comuni dell'ampliamento secondo i 5 gruppi di cittadinanze più rappresentate (Per Milano: Africa settentrionale; Asia orientale; America Centro-meridionale; Ue; Asia Centro-meridionale. Per Roma: Ue; Europa Centro-orientale; Asia orientale; America Centro-meridionale; Asia Centro-meridionale. Per Napoli: Ue; Europa Centro-orientale; Asia Centro-meridionale; Asia orientale; America Centro-meridionale).
- Stranieri residenti in ciascuno dei tre comuni dell'ampliamento secondo le cinque cittadinanze più rappresentate (Per Milano: Filippine; Egitto; Cina; Perù; Equador. Per Roma: Filippine; Romania; Cina; Bangladesh; Polonia. Per Napoli: Ucraina; Romania; Cina; Sri Lanka; Polonia).
- stranieri residenti in ciascuno dei tre comuni dell'ampliamento per sesso e 3 classi d'età (0-15, 16-34, 35 e più);
- stranieri residenti in ciascuno dei tre comuni dell'ampliamento per 7 classi d'età (0-5, 6-15, 16-24, 25-34, 35-44, 45-64, 65 e più.);

[5] La calibrazione è una calibrazione integrata, vale a dire che tutti gli individui della medesima famiglia hanno lo stesso coefficiente di riporto all'universo.

3.5 Valutazione del livello di precisione delle stime

Le stime prodotte da un'indagine campionaria sono sempre affette da errore. Questo si distingue in errore campionario, che deriva proprio dall'incertezza derivante dall'aver osservato la variabile di interesse solo su una parte (campione) della popolazione; ed errore non campionario, che deriva essenzialmente da errori nelle liste della popolazione utilizzate per selezionare le unità del campione; mancate risposte parziali dovute a risposte mancanti o non ammissibili a causa di errori di rilevazione o di registrazione; in generale, da tutto ciò che ha a che fare con le tecniche di indagine utilizzate e i comportamenti dei rilevatori.

In questo paragrafo si descrivono le metodologie e le tecniche utilizzate per la valutazione dell'errore campionario associato alle stime prodotte.

Le principali statistiche per valutare l'errore campionario sono l'errore di campionamento assoluto e l'errore di campionamento relativo. La stima dell'errore di campionamento assoluto e relativo di una generica stima $\hat{Y}$ sono definite dalle seguenti espressioni:

$$\hat{\sigma}(\hat{Y}) = \sqrt{\widehat{Var}(\hat{Y})}$$

$$\hat{\epsilon}(\hat{Y}) = \frac{\hat{\sigma}(\hat{Y})}{\hat{Y}}$$

Conoscendo la stima $\hat{Y}$ di un parametro Y della popolazione e la stima dell'errore assoluto $\hat{\sigma}(\hat{Y})$ ad essa associato, è possibile costruire un intervallo di confidenza che, con livello di fiducia α, contiene al suo interno il valore del parametro Y oggetto di stima; tale intervallo è:

$$\{\hat{Y} - k_\alpha \hat{\sigma}(\hat{Y}) \leq Y \leq \hat{Y} + k_\alpha \hat{\sigma}(\hat{Y})\}$$

dove il valore di k_α dipende dalla forma della distribuzione campionaria dello stimatore e dal valore scelto per il livello di confidenza α; per grandi campioni si fa comunemente riferimento alla distribuzione normale e si ha ad esempio, per α=0,05, che k=1,96.

3.6 Presentazione sintetica degli errori campionari

Ad ogni stima generica stima $\hat{Y}$ corrisponde una stima dell'errore campionario relativo che consente di valutarne la precisione; pertanto, per consentire una corretta interpretazione delle stime prodotte, sarebbe necessario presentare contestualmente a ciascuna stima anche il corrispondente errore campionario stimato. Ciò, tuttavia, non è possibile quando le stime prodotte sono in numero molto elevato. Per questi motivi si ricorre frequentemente ad una presentazione sintetica delle stime degli errori campionari, basata sul metodo dei modelli regressivi. Questo metodo si basa sulla determinazione di una semplice funzione matematica che mette in relazione ciascuna stima con il proprio errore campionario relativo stimato.
Il modello utilizzato per le stime di frequenze assolute e relative riferite agli individui è il seguente:

$$\log(\hat{\epsilon}^2(\hat{Y})) = a + b * \log(\hat{Y})$$

dove i parametri a e b sono stimati con il metodo dei minimi quadrati. I modelli regressivi del tipo descritto, che permettono la presentazione sintetica degli errori di campionamento, sono stati ottenuti tramite un software generalizzato messo a punto dall'Istat.

Nel prospetto 1 sono riportati i valori dei coefficienti a e b e del coefficiente di determinazione R^2 dei modelli stimati per l'interpolazione degli errori campionari relativi delle stime di frequenze assolute e relative per il totale Italia e per le diverse ripartizioni geografiche; similmente il prospetto 2 riporta i valori riferiti alle stime per il campione ampliato per Milano, Roma e Napoli.
Utilizzando gli opportuni coefficienti è possibile calcolare una stima dell'errore campionario relativo di una generica stima di una frequenza $\hat{Y}$ applicando la seguente formula:

$$\hat{\epsilon}(\hat{Y}) = \sqrt{\exp(a + b * \log(\hat{Y}))}$$

Prospetto 1 - Valori dei coefficienti a, b e R^2 delle funzioni utilizzate per l'interpolazione degli errori campionari delle stime - Campione nazionale

	a	b	R^2
ITALIA	9.513502	-1.29689	97.22
RIPARTIZIONE GEOGRAFICA			
Nord–ovest	9.387337	-1.28158	97.60
Nord–est	8.41149	-1.24568	95.74
Centro	9.511312	-1.3472	95.21
Sud–ovest	6.976532	-1.28023	91.45
Sud–est	6.702995	-1.29117	93.73
Isole	6.241939	-1.23959	87.72
TIPO DI COMUNE			
Comuni metropolitani	9.712262	-1.39297	93.05
Comuni periferia dei comuni metropolitani	9.022745	-1.32295	94.66
Altri comuni fino a 10.000 abitanti	8.896622	-1.25646	96.88
Altri comuni con più di 10.000 abitanti	9.04692	-1.28708	97.02

Prospetto 2 - Valori dei coefficienti a, b e R^2 delle funzioni utilizzate per l'interpolazione degli errori campionari delle stime – Ampliamento campionario per Roma, Milano e Napoli

	a	b	R^2
COMUNE			
Milano	5.08642	-1.04405	98.48
Roma	5.25919	-1.05239	98.93
Napoli	2.39128	-0.97204	98.52

Infine, i prospetti 3 e 4 hanno lo scopo di rendere più agevole e immediata la valutazione degli errori campionari. In testata sono elencati valori crescenti di stima di frequenze relative (0.005, 0.010, 0.020, …,0.400, 0.500); in fiancata sono riportati i domini di riferimento delle stime; le celle interne contengono gli errori campionari relativi percentuali stimati mediante la formula precedente. Consultando queste tavole è possibile disporre di una valutazione immediata (anche se meno precisa rispetto all'applicazione della formula precedente), dell'errore campionario di una generica stima di una frequenza relativa (o assoluta, ricavabile moltiplicando la frequenza relativa al totale degli stranieri nel dominio di riferimento), cercando nella testata il valore che più si avvicina alla stima di interesse e in fiancata il dominio di riferimento.

Prospetto 3 - Valori interpolati degli errori relativi percentuali delle stime - Campione nazionale

	STIME DI FREQUENZA RELATIVA								
	0.005	**0.01**	**0.02**	**0.05**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
ITALIA	15.5	9.9	6.3	3.5	2.2	1.4	1.1	0.9	0.8
RIPARTIZIONE GEOGRAFICA									
Nord–ovest	30.2	19.4	12.4	6.9	4.4	2.8	2.2	1.8	1.6
Nord–est	26.4	17.2	11.2	6.3	4.1	2.7	2.1	1.7	1.5
Centro	31.6	19.8	12.4	6.7	4.2	2.6	2.0	1.7	1.4
Sud-ovest	29.7	19.0	12.2	6.8	4.4	2.8	2.2	1.8	1.6
Sud-est	31.6	20.2	12.9	7.1	4.6	2.9	2.3	1.9	1.6
Isole	30.0	19.6	12.7	7.2	4.7	3.1	2.4	2.0	1.7
TIPO DI COMUNE									
Comuni metropolitani	37.3	23.0	14.2	7.5	4.6	2.9	2.2	1.8	1.5
Comuni periferia dei comuni metropolitani	47.7	30.2	19.1	10.4	6.6	4.2	3.2	2.6	2.3
Altri comuni fino a 10.000 abitanti	30.3	19.6	12.7	7.1	4.6	3.0	2.3	1.9	1.7
Altri comuni con più di 10.000 abitanti	21.8	14.0	8.9	5.0	3.2	2.0	1.6	1.3	1.1

Prospetto 4 - Valori interpolati degli errori relativi percentuali delle stime - Ampliamento campionario per Roma, Milano e Napoli

	STIME DI FREQUENZA RELATIVA								
	0.005	**0.01**	**0.02**	**0.05**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
COMUNE									
Milano	32.2	22.5	15.6	9.7	6.8	4.7	3.8	3.3	2.9
Roma	30.3	21.1	14.6	9.0	6.3	4.4	3.5	3.0	2.7
Napoli	24.2	17.3	12.4	7.9	5.7	4.0	3.3	2.9	2.6