Nota metodologica

Premessa

L'indagine europea sulla salute (Ehis) è condotta in tutti gli Stati dell'Unione europea con l'obiettivo di costruire indicatori di salute confrontabili a livello europeo sui principali aspetti delle condizioni di salute della popolazione, il ricorso ai servizi sanitari e i determinanti di salute. L'indagine è prevista dal regolamento (Ue) n. 255/2018 della Commissione, del 19 febbraio 2018 (che attua il regolamento (Ce) n. 1338/2008 del Parlamento europeo e del Consiglio relativo alle statistiche comunitarie in materia di sanità pubblica e di salute e sicurezza sul luogo di lavoro) ed è inserita nel Programma statistico nazionale 2017-2019 (cod. IST 02565).
Per l'Italia rappresenta la seconda edizione dopo la prima realizzata nel 2015, mentre per la maggior parte dei paesi europei (17 paesi dell'UE) quella del 2019 costituisce la terza edizione dell'indagine (Ehis wave 3). L'Italia ha comunque partecipato attivamente al lungo processo di armonizzazione per la definizione degli strumenti di rilevazione sin dagli inizi del 2000, nonché alle attività di un'apposita Task Force (TF) in ambito europeo per la realizzazione di quest'ultima edizione.
Allo scopo di garantire il più possibile la comparabilità con l'edizione del 2015, il mandato principale della TF per la wave 3 era di introdurre pochi interventi migliorativi al fine di non stravolgere il modello di rilevazione utilizzato nella seconda wave, e al contempo di arricchire il quadro informativo comunitario, inserendo su base volontaria alcuni moduli aggiuntivi. L'Italia, tra i moduli aggiuntivi, ha selezionato quelli sulla partecipazione sociale delle persone con disabilità (*Disability module*) e quello sulla valutazione delle prestazioni sanitarie ricevute (*Patient Experience module*). Contestualmente è stato quindi aggiornato il manuale metodologico predisposto da Eurostat, che fornisce tutte le raccomandazioni e le istruzioni per implementare al meglio l'Indagine: European Health Interview Survey (EHIS.wave 3) — Methodological manual: https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-02-18-240).
In questa edizione, sono inoltre stati inseriti alcuni quesiti per soddisfare bisogni informativi nazionali, al fine di poter monitorare alcuni rilevanti aspetti più specifici del nostro paese o per dare continuità informativa, ove possibile, ad alcuni temi indagati nelle precedenti indagini sulla salute: il fenomeno dell'out of pocket nel ricorso ai servizi sanitari, la prevenzione dei tumori femminili, il ricorso a metodi contraccettivi, o anche la prevalenza di patologie particolarmente diffuse nella popolazione molto anziana (demenze senili, Parkinson, ecc.), nonché alcuni aspetti della salute nei bambini (diffusione dell'allattamento, eccesso di peso nei minori, livelli adeguati di attività fisica).

Finalità e caratteristiche dell'indagine

In Italia l'indagine Ehis (wave 3) è stata condotta dall'Istat nel 2019, suddividendo il campione di famiglie in due periodi di rilevazione: il primo da aprile a giugno e il secondo da settembre a dicembre (anche per ottemperare al periodo di riferimento previsto dal regolamento comunitario).
Per la gran parte dei quesiti le interviste sono state condotte secondo la tecnica *Pen and paper interview* (Papi) - tecnica di rilevazione che prevede l'utilizzo delle interviste faccia-a-faccia da parte di un rilevatore comunale, appartenente alla rete comunale, adeguatamente e prioritariamente formato dall'Istat. Per un'altra parte di quesiti, più esigua, è stata prevista l'autocompilazione del questionario. Nell'intervista diretta è stato somministrato un questionario familiare e tante schede individuali quanti sono i membri della famiglia.
Il campione realizzato è di circa 22.800 famiglie residenti in 835 comuni di diversa ampiezza demografica, distribuiti su tutto il territorio nazionale. Il disegno campionario (cfr. il parag. "Strategia di campionamento e livello di precisione dei risultati") è a due stadi con stratificazione delle unità di primo stadio (comuni). Le unità di secondo stadio sono le famiglie, estratte con criterio di scelta casuale dalle liste anagrafiche per i comuni campione con meno di 1000 abitanti e dalla lista delle famiglie selezionate per il Censimento Permanente del 2018 per i comuni campione con 1000 abitanti e oltre, in modo da costituire un campione statisticamente rappresentativo della popolazione residente.

L'unità di rilevazione è costituita dalla famiglia di fatto (ff) associata alla famiglia anagrafica (fa) campionata. La famiglia di fatto è definita come l'insieme di persone che dimorano abitualmente nella stessa abitazione e sono legate da vincoli di parentela, affinità, affettività o amicizia.
Le tematiche trattate riguardano tre macro aree: lo stato di salute, i determinanti di salute e l'accesso ed utilizzo dei servizi sanitari, indagati insieme al contesto socio-demografico di ciascun individuo delle famiglie intervistate. Più nello specifico i contenuti informativi, corrispondenti alle sezioni tematiche in cui sono suddivisi i questionari per l'Italia, sono i seguenti:

- Dati anagrafici
- Condizioni generali di salute
- Salute orale
- Peso e altezza
- Consumo di frutta e verdura
- Attività fisica
- Malattie e condizioni croniche
- Infortuni e lesioni
- Limitazioni funzionali fisiche e sensoriali
- Attività di cura della persona
- Attività domestiche
- Dolore
- Benessere psicologico
- Assistenza ospedaliera
- Visite mediche di medicina generale
- Visite mediche specialistiche
- Accertamenti diagnostici
- Assistenza domiciliare e altri servizi
- Consumo di farmaci
- Prevenzione
- Difficoltà di accesso a prestazioni sanitarie
- Sostegno sociale
- Cure o assistenza fornite
- Situazione lavorativa
- Assenze dal lavoro per motivi di salute
- Consumo di tabacco
- Consumo di bevande
- Stato di salute percepito
- Dolore cronico
- Partecipazione alla vita sociale
- Valutazione delle prestazioni sanitarie
- Allattamento al seno
- Metodi contraccettivi
- L'abitazione in cui vive
- Aiuti e situazione economica della famiglia
- Altre persone coabitanti nella famiglia

Per prendere visione della varietà di informazioni e della formulazione degli specifici quesiti contenuti nei questionari, si suggerisce di consultare la pagina dedicata all'indagine sul sito dell'Istat, disponibile all'indirizzo: http://www.istat.it/it/archivio/167485. La maggior parte delle sezioni dei modelli di rilevazione fa riferimento alla popolazione di 15 anni e oltre, come richiesto dal Regolamento europeo, fanno eccezione le prime sezioni del questionario per intervista, dove si raccolgono informazioni anche sui minori di 15 anni, per soddisfare bisogni informativi nazionali, rilasciate su base volontaria da un genitore o un adulto di riferimento della famiglia.

Strategia di campionamento e livello di precisione dei risultati[1]

1. Obiettivi conoscitivi

La *popolazione di interesse* dell'indagine in oggetto, ossia l'insieme delle unità statistiche intorno alle quali si intende investigare, è costituita dalle famiglie residenti in Italia e dai membri che le compongono; sono pertanto esclusi i 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.

Per soddisfare i bisogni informativi a livello territoriale e consentire stime regionali utili alla programmazione sanitaria locale, nonché ove possibile stime sub-regionali, i domini di studio, ossia gli ambiti territoriali ai quali sono riferiti i parametri di popolazione oggetto di stima sono:

[1] Lo studio del disegno campionario, il calcolo dei coefficienti di riporto all'universo e degli errori campionari è stato condotto mediando le esigenze informative a livello nazionale e quelle europee, nel rispetto delle raccomandazioni definite da Eurostat nel manuale d'indagine Ehis (wave3). Il lavoro è il frutto della collaborazione con i colleghi metodologi coordinati da C. De Vitiis (F.Inglese, A.Guandalini, M.D.Terribili, D.Moretti).

- le cinque ripartizioni geografiche (Italia Nord-Occidentale, Italia Nord-Orientale, Italia Centrale, Italia Meridionale, Italia Insulare);
- le regioni (ad eccezione del Trentino Alto Adige le cui stime sono prodotte distintamente per le province autonome di Bolzano e Trento).

I domini sub-regionali, indicati come Aree Vaste, costituiti da aggregati territoriali di interesse per la programmazione sanitaria a livello locale e definiti in relazione allo specifico contesto informativo dell'indagine sulle condizioni di salute, sono utilizzati per la stratificazione del campione.
Questi ultimi domini sono stati definiti partendo dalla considerazione che, sebbene le unità amministrative territoriali di prevalente interesse per la programmazione sanitaria sono le Aziende Sanitarie Locali (ASL), tuttavia non potendo progettare, per vincoli di costo, un disegno campionario che garantisse stime attendibili a tale livello di dettaglio, si è provveduto a considerare tali aggregati di ASL nella stratificazione del campione. In tal modo risulta più agevole calcolare, eventualmente, stime indirette riferite alle ASL e alle Aree Vaste. La dimensione media di popolazione delle Aree Vaste è pari a circa 850.000 abitanti.

2. Struttura generale del disegno

Il disegno di campionamento ha una struttura generale che ricalca quella degli schemi campionari della maggior parte delle indagini sulle famiglie dell'Istat, ossia un disegno a più stadi comuni-famiglie, con stratificazione dei comuni. Per l'indagine EHIS 2019 il campione è stato integrato con il disegno campionario seguito per il Master Sample del Censimento permanente. Nel caso specifico, i comuni campione sono stati individuati come sotto-campione del campione di 2850 comuni del Master Sample utilizzato per il Censimento 2018. A tale scopo, lo schema campionario classico utilizzato per le indagini sulle famiglie, di seguito descritto, è stato implementato sul sotto-universo dei comuni rilevati per il Censimento Permanente a ottobre del 2018.
Nell'ambito di ogni Area Vasta i comuni universo sono stati suddivisi in due sottoinsiemi: i comuni di maggiore dimensione demografica costituiscono strato a sé stante e vengono definiti Auto Rappresentativi (AR); i rimanenti comuni sono definiti Non Auto Rappresentativi (NAR) e sono suddivisi, sulla base della dimensione demografica, in strati di uguale ampiezza; da tali strati i comuni campione (due per ogni strato) sono stati selezionati con probabilità proporzionali alla loro dimensione.
Per ognuno dei comuni coinvolti nell'indagine (AR e NAR), viene effettuato un campionamento a grappoli: i grappoli - le famiglie - sono selezionati in maniera casuale dalla lista anagrafica e tutti i componenti che appartengono alla famiglia di fatto vengono sottoposti a rilevazione. La numerosità minima di famiglie campione per ciascun comune è stata posta pari a 22.
Le famiglie sono selezionate per ciascun comune campione a partire dal campione teorico del Master Sample; per ogni famiglia inclusa nel campione vengono rilevate le caratteristiche oggetto di indagine di tutti i componenti di fatto appartenenti alla famiglia medesima.
Il campione finale teorico adottato per l'indagine europea EHIS 2019 comprende 837 comuni e 23.700 famiglie.

3. Stratificazione e selezione delle unità campionarie

L'obiettivo della stratificazione è, in generale, 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 esame, i comuni vengono stratificati in base alla loro dimensione demografica e nel rispetto delle seguenti condizioni:

- autoponderazione del campione a livello regionale;
- selezione di due comuni 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; tale numero è stato posto pari a 22;
- formazione di strati aventi ampiezza approssimativamente costante in termini di popolazione residente.

Il procedimento di stratificazione, attuato all'interno di ogni area vasta 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 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; $_r f$ la frazione di campionamento;

- suddivisione di tutti i comuni nei due sottoinsiemi AR e NAR: i comuni di dimensione superiore o uguale alla soglia sono definiti come comuni AR e i rimanenti come NAR;
- suddivisione dei comuni dell'insieme NAR in strati aventi dimensione, in termini di popolazione residente, approssimativamente costante e all'incirca pari a due volte la soglia $_r\lambda$; il numero due è connesso con il fatto che da ciascuno strato si selezionano due comuni campione.

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 quindi due comuni campione con probabilità proporzionale alla dimensione demografica, mediante la procedura di selezione sistematica proposta da Madow[2].

3.1 Definizione della dimensione campionaria

Il campione teorico adottato per l'indagine europea EHIS 2019, come detto sopra, ha una dimensione in termini di famiglie pari a 23.700[3]. Al fine di tenere sotto controllo gli errori campionari sia delle stime nazionali sia di quelle regionali, è stata utilizzata un'allocazione di compromesso tra l'allocazione uniforme e quella proporzionale alla popolazione. All'interno delle regioni il campione è stato distribuito in modo proporzionale tra le aree vaste sub-regionali.

Per fare fronte alla prevista caduta delle risposte dovuta ai rifiuti o alla irreperibilità delle famiglie da intervistare (errori di lista) la dimensione campionaria per ciascun comune campione è stata incrementata utilizzando i tassi di mancata risposta osservati in precedenti indagini simili, portando a selezionare un campione di famiglie di numerosità pari a 30.142 famiglie.

La dimensione finale effettiva del campione di famiglie rispondenti è risultata pari a quasi 22.800.

Nel prospetto 1 viene riportata la distribuzione regionale dell'universo e del campione dei comuni, delle famiglie (selezionate complessivamente per sovra campionamento e rispondenti) e degli individui.

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

[3] La dimensione campionaria di famiglie è stata aumentata rispetto a quella definita nel campione minimo indicato da Eurostat poiché si è tenuto conto di esigenze di stima nazionali e territoriali (regioni e aree vaste).

Prospetto 1 – Distribuzione regionale dei comuni, delle famiglie e degli individui nell'universo e nel campione.

REGIONI	Comuni			Famiglie			Individui	
	Universo	Campione selezionato	Campione effettivo	Universo (a)	Campione selezionato	Campione effettivo	Universo (a)	Campione effettivo
Piemonte	1.197	61	60	2.019.357	2.475	1.722	4.312.064	3.678
Valle d'Aosta/ Vallée d'Aoste	74	22	22	59.027	798	621	124.575	1.338
Liguria	234	27	27	764.819	1.458	1.095	1.535.091	2.245
Lombardia	1.516	86	86	4.335.560	3.088	2.182	9.995.199	4.973
Trentino-Alto Adige	292	48	48	444.127	1.691	1.233	1.059.996	2.852
Bolzano - Bozen	116	24	24	209.708	844	581	524.308	1.344
Trento	176	24	24	234.420	847	652	535.688	1.508
Veneto	574	55	55	2.047.541	1.652	1.332	4.866.504	3.128
Friuli-Venezia Giulia	217	31	31	538.803	1.089	846	1.205.068	1.815
Emilia-Romagna	331	48	48	1.959.349	1.697	1.330	4.426.195	2.912
Toscana	274	53	53	1.625.724	1.922	1.473	3.704.651	3.358
Umbria	92	22	21	370.966	841	580	874.547	1.321
Marche	229	35	35	631.523	1.141	903	1.516.308	2.201
Lazio	378	33	33	2.566.713	2.304	1.434	5.836.692	3.142
Abruzzo	305	35	35	536.496	1.100	819	1.306.975	1.885
Molise	136	20	20	122.517	661	535	302.640	1.261
Campania	550	55	55	2.136.736	1.683	1.403	5.774.695	3.787
Puglia	258	47	47	1.599.910	1.436	1.228	4.013.977	2.947
Basilicata	131	25	25	228.354	737	620	559.321	1.419
Calabria	405	41	41	776.114	1.236	1.029	1.937.913	2.486
Sicilia	390	53	53	2.015.104	1.870	1.454	4.970.108	3.515
Sardegna	377	40	40	707.147	1.263	957	1.632.299	2.179
ITALIA	**7.960**	**837**	**835**	**25.485.888**	**30.142**	**22.796**	**59.954.817**	**52.442**

(a) Stima Indagine europea sulla salute, dati in migliaia (Ehis 2019)

4. Procedimento per il calcolo delle stime

Le stime prodotte dall'indagine sono essenzialmente stime di frequenze assolute e relative, riferite alle famiglie e agli individui.

Le stime sono 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 rappresentata dall'unità medesima.

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 famiglia; *p*, indice di componente della famiglia; *h*, indice di strato di comuni; *y*, generica variabile oggetto di indagine; y_{hijp}, valore di *y* osservato sul componente *p* della famiglia *j* del comune *i* dello strato *h*; P_{hij}, numero di componenti della famiglia *j* del comune *i* dello strato *h*; $Y_{hij} = \sum_{p=1}^{P_{hij}} y_{hijp}$, totale della variabile *y* osservato sulla famiglia *j* del comune *i* dello strato *h*; M_{hi}, numero di famiglie residenti nel comune i dello strato *h*; m_{hi}, campione di famiglie nel comune i dello strato *h*; N_h, totale di comuni nello strato *h*; n_h, numero di comuni campione nello strato *h* (nell'indagine in oggetto si ha $n_h = 1$); 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

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

La stima del totale (1) è data da

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

in cui w_{hij} è il peso finale da attribuire a tutti i componenti della famiglia *j* del comune *i* dello strato *h*.

Dalla precedente relazione si desume, quindi, che per ottenere la stima del totale (1) occorre moltiplicare il valore della variabile *y* assunto da ciascuna unità campionaria per il peso di tale unità[4] ed effettuare, a livello del dominio di interesse, la somma dei prodotti così ottenuti.

[4] Al fine di ottenere stime coerenti per individui e famiglie i pesi finali sono definiti in modo tale che a ciascuna famiglia *hij* e a tutti i componenti della stessa sia assegnato un medesimo peso finale w_{hij}.

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 alcune delle famiglie selezionate 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 sono stati definiti i seguenti vincoli (totali noti di popolazione):

- il primo si riferisce alla distribuzione della popolazione nelle 21 regioni italiane per sesso e sette classi d'età[5];
- il secondo si riferisce alla distribuzione della popolazione nelle 5 ripartizioni territoriali per sesso e nove classi d'età[6];
- il terzo riguarda la distribuzione delle famiglie per cittadinanza (famiglia di cittadini italiani, famiglia di cittadini stranieri, famiglia mista) nelle 5 ripartizioni territoriali;
- il quarto riguarda il numero di famiglie monocomponente per ripartizione territoriale.

La procedura che consente di costruire i *pesi finali* da attribuire alle unità campionarie rispondenti, è articolata nelle seguenti fasi:

1) si calcolano i *pesi diretti*, d_{hij} , come reciproco della probabilità di inclusione delle unità;
2) si calcolano i fattori correttivi per mancata risposta totale;
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 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 passo 4.

I fattori correttivi per mancata risposta totale del passo 3, sono ottenuti tramite il metodo noto come *response propensity*. L'applicazione del metodo si basa su un modello di risposta che sfrutta variabili ausiliarie note per le unità rispondenti e non rispondenti. Per l'indagine in oggetto il modello utilizzato è di tipo logit[7]; le propensioni di risposta stimate a livello familiare sono utilizzate per la costruzione di strati (celle di ponderazione) definiti sulla base dei quartili della distribuzione di probabilità; in ogni strato il fattore correttivo è calcolato come inverso del tasso di risposta osservato (*response propensity stratification*).

I fattori correttivi del passo 4 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 euclidea; l'adozione di tale funzione garantisce che i pesi

5 Le 7 classi d'età considerate sono 0-14, 15-24, 25-44, 45-54, 55-64, 65-74, 75+.

6 Le 9 classi d'età considerate sono 0-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85+.

7 Il modello logit è definito dalla variabile dipendente binaria "esito di risposta", e dalle covariate provenienti dal set di dati del Master sample: tipologia familiare, cittadinanza familiare, ripartizione territoriale per tipologia comunale. Il modello utilizzato è stato scelto in seguito ad una fase di studio condotta attraverso una procedura di stepwise, che ha omesso dal modello le covariate numero di componenti della famiglia e titolo di studio più alto nella famiglia, in quanto non statisticamente significative.

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).

La presenza di valori estremi dei pesi finali è stata controllata tenendo conto della regola riportata nel manuale metodologico di EHIS wave3 (pag. 150). A tal fine è stata calcolata la quantità Q_{hij}:

$$Q_{hij} = \frac{w_{hij}}{\overline{w}_{hij}} \frac{\overline{d}_{hij}}{d_{hij}},$$

in cui d_{hij} è il peso iniziale attribuito a tutti i componenti della famiglia *j* del comune *i* dello strato *h*, $\overline{d}_{hij}$ e $\overline{w}_{hij}$ sono i pesi medi rispettivamente dei pesi iniziali e dei pesi finali.

Tale quantità è stata utilizzata per definire l'intervallo di accettazione dei valori dei pesi finali sulla base della relazione:

$$\frac{1}{C} \leq Q_{hij} \leq C,$$

dove *C* è una costante che assume valore 3.

I pesi finali, il cui *range* va da 31,87 a 5649,722, non presentano valori estremi, in termini di variazione tra peso iniziale e peso finale.

Per l'indagine in oggetto, il peso finale è stato determinato con la procedura di calibrazione sviluppata in Re-Genesees (Zardetto, 2015) [8]. Questo software integra una funzione (aggregate.stage=2) che permette di dare stesso peso finale agli individui appartenenti alla stessa famiglia.

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[9]. 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 perché è possibile dimostrare che tutti gli stimatori di ponderazione vincolata convergono asintoticamente, all'aumentare della numerosità campionaria, allo stimatore di regressione generalizzata.

[8] Zardetto D. (2015). ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys. Journal of Official Statistics. Volume 31, Issue 2, Pages 177–203, ISSN (Online) 2001-7367, June 2015

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

5. Valutazione del livello di precisione delle stime

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 $\hat{Var}(\hat{Y}_d)$ 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:

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

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 in precedenza, le stime prodotte dall'indagine sono state ottenute mediante uno stimatore di ponderazione vincolata definito in base a una funzione di distanza di tipo logaritmico troncato. Poiché, lo stimatore adottato non è funzione lineare dei dati campionari, per la stima della varianza $\hat{Var}(\hat{Y}_d)$ 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 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:

$$\hat{Y}_d \cong \hat{Z}_d = \sum_{h=1}^{H_d} \hat{Z}_h\ , \quad \text{essendo} \quad \hat{Z}_h = \sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} z_{hij} w_{hij} \qquad (5)$$

dove z_{hij} è la variabile linearizzata espressa come $z_{hij} = y_{hij} - \mathbf{x}_{\mathbf{hij}}^{'}\beta$, essendo $\mathbf{x}_{\mathbf{hij}} = \left({}_1x_{hij},...,{}_kx_{hij},....{}_Kx_{hij}\right)'$ il vettore contenente i valori delle K (K=18) variabili ausiliarie, osservati per la generica famiglia h_{ij} 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

$$\hat{Var}\left(\hat{Y}_d\right) \cong \hat{Var}\left(\hat{Z}_d\right) = \sum_{h=1}^{H_d} \hat{Var}\left(\hat{Z}_h\right). \qquad (6)$$

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, $\hat{Var}\left(\hat{Z}_h\right)$, della stima $\hat{Z}_h$ è differente a seconda che lo strato sia AR oppure NAR. Possiamo, quindi scomporre come segue

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

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:

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

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, qualora sia stato rilevato un solo comune campione nello 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:

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

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, $\hat{V}ar\left(\hat{Y}_d\right)$, 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:

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

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.

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 a 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.

Nei prospetti 2a e 3a 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 regione.

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 agli individui dell'Italia Nord occidentale, l'errore relativo corrispondente si ottiene introducendo nella (12) i valori dei parametri a e b riportati nella prima riga del prospetto 2a.

I prospetti 2b e 3b, presentati in aggiunta con riferimento agli individui, consentono di rendere più agevole il calcolo degli errori campionari. Essi contengono gli errori di campionamento relativo, per ciascun dominio territoriale di interesse, calcolati mediante la formula (12), corrispondenti alle stime di frequenze assolute.

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 il livello di stima (riportato in colonna) che più si avvicina alla stima di interesse e nel considerare come errore relativo il valore che si trova sulla riga 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 in 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 2a – Valori dei coefficienti a, b e dell'indice di determinazione R^2 (%) delle funzioni utilizzate per le interpolazioni degli errori campionari delle stime a livello nazionale e per ripartizione geografica

RIPARTIZIONE GEOGRAFICA	a	b	R^2(%)
Nord-Ovest	8,290	-1,076	97,9
Nord-Est	8,492	-1,109	97,3
Centro	7,995	-1,061	97,3
Sud	7,275	-1,006	97,4
Isole	7,466	-1,024	96,5
ITALIA	8,268	-1,072	98,1

Prospetto 2b - Valori interpolati degli errori campionari delle stime per ripartizione geografica

RIPARTIZIONE GEOGRAFICA	Valori della stima – frequenza assoluta									
	25000	50000	75000	100000	250000	500000	750000	1000000	2500000	5000000
Nord-Ovest	27,21	18,74	15,07	12,91	7,89	5,43	4,37	3,74	2,29	1,57
Nord-Est	25,45	17,33	13,84	11,80	7,10	4,83	3,86	3,29	1,98	1,35
Centro	25,27	17,50	14,11	12,11	7,45	5,16	4,16	3,57	2,20	1,52
Sud	23,30	16,44	13,41	11,60	7,32	5,16	4,21	3,64	2,30	1,62
Isole	23,45	16,44	13,36	11,53	7,21	5,06	4,11	3,55	2,22	1,56
ITALIA	27,40	18,90	15,21	13,03	7,98	5,50	4,43	3,79	2,32	1,60

Prospetto 3a – Valori dei coefficienti a, b e dell'indice di determinazione R^2 (%) delle funzioni utilizzate per le interpolazioni degli errori campionari delle stime per regione

REGIONE	a	b	R^2(%)
Piemonte	8,00	-1,08	97,4
Valle d'Aosta/Vallée d'Aoste	5,22	-1,09	95,5
Lombardia	8,51	-1,08	97,8
Trentino Alto Adige			
Bolzano- Bozen	5,98	-1,01	95,7
Trento	6,01	-1,02	95,1
Veneto	8,46	-1,11	97,7
Friuli Venezia Giulia	6,96	-1,05	96,7
Liguria	7,06	-1,05	96,8
Emilia Romagna	8,36	-1,09	96,6
Toscana	7,51	-1,04	97,4
Umbria	7,03	-1,06	95,2
Marche	6,65	-1,00	96,0
Lazio	8,20	-1,06	96,9
Abruzzo	7,30	-1,07	96,1
Molise	5,42	-0,98	93,8
Campania	7,22	-0,98	96,6
Puglia	7,40	-1,02	97,2
Basilicata	6,53	-1,06	96,3
Calabria	6,72	-0,99	96,7
Sicilia	7,72	-1,03	96,5
Sardegna	6,88	-1,02	97,1

Prospetto 3b - Valori interpolati degli errori campionari delle stime per regione

REGIONE	Valori della stima – frequenza assoluta									
	10000	25000	50000	75000	100000	250000	500000	750000	1000000	2500000
Piemonte	37,06	22,55	15,49	12,43	10,64	6,47	4,44	3,57	3,05	1,86
Valle D'Aosta/Vallée d'Aoste	8,82	5,34	3,66	2,93	2,50	1,52	1,04	0,83	0,71	0,43
Lombardia	48,70	29,68	20,41	16,39	14,03	8,55	5,88	4,72	4,04	2,46
Trentino Alto Adige										
Bolzano - Bozen	19,28	12,15	8,57	6,99	6,05	3,81	2,69	2,19	1,90	1,20
Trento	18,27	11,44	8,03	6,53	5,64	3,53	2,48	2,01	1,74	1,09
Veneto	41,93	25,25	17,20	13,74	11,72	7,06	4,81	3,84	3,28	1,97
Friuli Venezia Giulia	25,43	15,69	10,90	8,80	7,56	4,67	3,24	2,62	2,25	1,39
Liguria	27,20	16,81	11,69	9,45	8,12	5,02	3,49	2,82	2,43	1,50
Emilia Romagna	42,71	25,89	17,73	14,21	12,14	7,36	5,04	4,04	3,45	2,09
Toscana	34,88	21,61	15,05	12,18	10,48	6,49	4,52	3,66	3,15	1,95
Umbria	25,49	15,68	10,86	8,76	7,52	4,63	3,21	2,59	2,22	1,37
Marche	27,33	17,26	12,19	9,94	8,61	5,43	3,84	3,13	2,71	1,71
Lazio	46,05	28,36	19,65	15,86	13,62	8,39	5,81	4,69	4,03	2,48
Abruzzo	27,86	17,06	11,77	9,47	8,12	4,97	3,43	2,76	2,37	1,45
Molise	16,70	10,67	7,61	6,24	5,42	3,47	2,47	2,03	1,76	1,13
Campania	41,15	26,30	18,75	15,38	13,37	8,55	6,09	5,00	4,34	2,78
Puglia	37,74	23,70	16,67	13,57	11,73	7,37	5,18	4,22	3,65	2,29
Basilicata	20,03	12,34	8,55	6,90	5,93	3,65	2,53	2,04	1,75	1,08
Calabria	30,27	19,24	13,65	11,17	9,69	6,16	4,37	3,58	3,10	1,97
Sicilia	40,89	25,48	17,82	14,45	12,46	7,76	5,43	4,40	3,80	2,37
Sardegna	28,23	17,68	12,41	10,09	8,71	5,46	3,83	3,11	2,69	1,68

AVVERTENZE

Le ripartizioni geografiche costituiscono una suddivisione geografica del territorio e sono così articolate:

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

Nord-est: comprende Trentino-Alto Adige (Bolzano-Bozen, Trento), Veneto, Friuli-Venezia Giulia, Emilia-Romagna

*Centro***:** comprende Toscana, Umbria, Marche, Lazio

*Sud***:** comprende Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria

Isole: comprendono Sicilia, Sardegna