ISTAT

Indagine sulle discriminazioni in base al genere, all'orientamento sessuale, all'appartenenza etnica

2011

Codice SN131

ADPSS-SOCIODATA
Archivio Dati e Programmi
per le Scienze Sociali

www.sociologiadip.unimib.it/sociodata
E-mail: adpss.sociologia@unimib.it
Tel.: 02 64487513
Fax: 02 64487561

Università degli Studi di Milano-Bicocca
Dipartimento di Sociologia e Ricerca Sociale

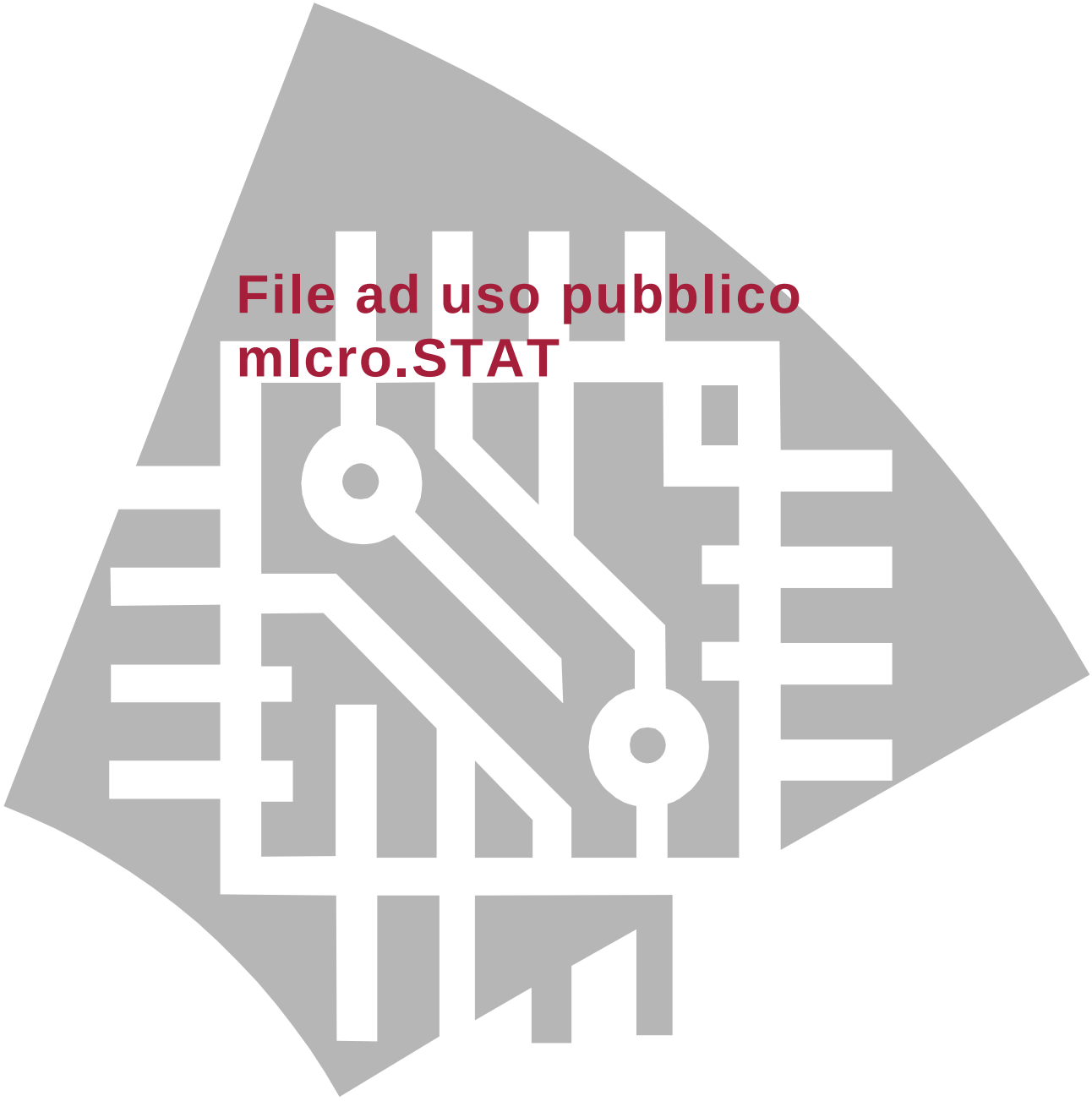
INDICE
TABLE OF CONTENTS

Note metodologiche
Methodological Notes

ASPETTI METODOLOGICI DELL'INDAGINE

SURVEY'S METHODOLOGICAL ISSUES

Istat
Istituto Nazionale di Statistica



File ad uso pubblico
mIcro.STAT

Discriminazioni in base al genere, all'orientamento sessuale e all'appartenenza etnica

Anno 2011

Aspetti metodologici dell'indagine

Ottobre 2014

INDICE

1. Introduzione

L'Indagine sulle Discriminazioni in base al genere, all'orientamento sessuale e all'appartenenza etnica è stata realizzata per la prima volta nel 2011, a seguito di una Convenzione stipulata con il Dipartimento delle Pari Opportunità, con l'obiettivo di colmare il gap informativo sulla diffusione e le forme che i fenomeni discriminatori assumono nel nostro Paese, con particolare riferimento a tre specifiche dimensioni: il genere, l'orientamento sessuale e l'appartenenza etnica. L'indagine è nata con un duplice obiettivo: da un lato, quello di rilevare la diffusione di stereotipi e atteggiamenti discriminatori nei confronti delle categorie oggetto di interesse; dall'altro, stimare il numero di persone che hanno subito esperienze discriminatorie. L'indagine consente, dunque, sia di rilevare le opinioni dei cittadini rispetto ai ruoli di genere, all'immigrazione, all'omosessualità, sia di stimare il numero e le caratteristiche delle vittime di atti discriminatori, con particolare riferimento alle discriminazioni subite nel contesto scolastico e in quello lavorativo (distinto in ricerca di lavoro e attività lavorativa). L'indagine rientra tra quelle comprese nel Programma statistico nazionale, che raccoglie l'insieme delle rilevazioni statistiche necessarie al Paese.

2. La popolazione di riferimento

La popolazione di interesse dell'indagine in oggetto, ossia l'insieme delle unità statistiche intorno alle quali si intende investigare, è costituita dagli individui residenti in famiglia in Italia; 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.

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 quattro ripartizioni geografiche (Italia nord-occidentale, Italia nord-orientale, Italia centrale, Italia meridionale e insulare);
- la tipologia comunale ottenuta suddividendo i comuni italiani in cinque classi formate in base a caratteristiche socio-economiche e demografiche:

A) comuni appartenenti all'area metropolitana suddivisi in:

A_1, comuni centro dell'area metropolitana: Torino, Milano, Venezia, Genova, Bologna, Firenze, Roma, Napoli, Bari, Palermo, Catania, Cagliari;

A_2, comuni che gravitano intorno ai comuni centro dell'area metropolitana;

B) comuni non appartenenti all'area metropolitana suddivisi in:

B_1 comuni aventi fino a 10.000 abitanti;

B_2 comuni con 10.001-50.000 abitanti;

B_3 comuni con oltre 50.000 abitanti.

3. Il disegno campionario

Il disegno campionario è stato progettato per garantire la precisione delle principali stime a livello dei domini di stima pianificati, costituiti dalle quattro ripartizioni geografiche e dalle cinque tipologie comunali. Solamente ai fini della stratificazione dei comuni è stata considerata anche la regione geografica per garantire una maggiore distribuzione del campione sul territorio. Pertanto i comuni sono stati stratificati all'interno di domini definiti dall'incrocio della regione geografica con le cinque aree A_1, A_2, B_1, B_2, e B_3 e 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. 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.

Da ogni famiglia inclusa nel campione viene selezionato casualmente un individuo a partire dalla lista degli individui di età compresa tra 18 e 74 anni, appartenenti alla famiglia di fatto.

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.

3.1 Definizione della numerosità 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 nazionale, una soluzione approssimativamente ottimale sarebbe quella di determinare la numerosità nazionale e ripartirla tra i domini sub-nazionali in modo proporzionale alla loro dimensione demografica; viceversa, avendo la finalità di produrre stime con uguale attendibilità a livello sub-nazionale, una soluzione approssimativamente ottimale sarebbe quella di selezionare un campione uguale in tutti i domini sub-nazionali. Quest'ultima soluzione, però, risulterebbe poco efficiente per le stime a livello nazionale.

In base alle considerazioni precedenti si è operato utilizzando il software MAUSS che, estendendo l'approccio proposto da Bethel (1989)[1], implementa una metodologia di allocazione multi-variata e multi-dominio (De Vitiis et al., 1998)2; tale software consente di ottenere un'allocazione del campione che rispetta dei prefissati vincoli di errore campionario a livello dei domini di stima definiti e per un certo numero di variabili di interesse.

Poiché per l'indagine in oggetto non erano disponibili informazioni sulla variabilità dei fenomeni di interesse, si è proceduto utilizzando come stime di interesse delle prevalenze tipiche ritenute opportune e fissando i vincoli sugli errori in modo differenziato sui domini territoriali di stima. In particolare si è fissato un errore relativo massimo pari al 15% per una prevalenza dell'1% a livello Italia e un errore massimo pari all'8% per una prevalenza del 10% a livello delle quattro ripartizioni geografiche e a livello di tipologia comunale.

In tal modo è risultato un campione nazionale di circa 8.000 individui

[1] Bethel J. (1989). Sample Allocation in Multivariate Surveys, *Survey Methodology*, 15, 47-57.

[2] De Vitiis C. Falorsi P.D., Ballin M., Scepi G., Principi e metodi del software generalizzato per la definizione del disegno di campionamento nelle indagini sulle imprese condotte dall'Istat", *Statistica applicata*, 10(2), 1998, pp. 235-257.

3.2 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 esame, i comuni vengono stratificati in base alla loro dimensione demografica e nel rispetto delle seguenti condizioni:

- autoponderazione del campione a livello di ciascun dominio definito dall'incrocio della ripartizione geografica e della tipologia comunale; in tal modo si garantisce che tutti gli individui appartenenti al medesimo dominio abbiano la stessa probabilità di inclusione nel campione;
- selezione di c=2 comuni campione nell'ambito di ciascuno strato definito sui comuni dell'insieme NAR;
- scelta di un numero minimo di famiglie da selezionare in ciascun comune campione; tale numero è stato posto pari a 8 per i comuni della tipologia B3 e a 9 per i restanti comuni;
- 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 aree A_1, A_2, B_1, B_2, e B_3 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:

$$_d\lambda = \frac{_d\overline{m}\ _d\delta}{_d f}$$

  in cui per il generico dominio *d* si è indicato con: $_d\overline{m}$ il numero minimo di famiglie da intervistare in ciascun comune campione; $_d\delta$ il numero medio di componenti per famiglia; $_d f$ la frazione di campionamento;
- suddivisione di tutti i comuni nei due sottoinsiemi AR e NAR: i comuni di dimensione superiore o uguale a $_d\lambda$ sono definiti come comuni AR e i rimanenti come NAR;
- suddivisione dei comuni dell'insieme NAR in strati aventi dimensione, in termini di

popolazione residente, approssimativamente costante e all'incirca pari a $c_d\lambda$.

Effettuata la stratificazione, i comuni AR sono inclusi con certezza nel campione; per quanto riguarda, invece, i comuni NAR, nell'ambito di ogni strato viene estratto un comune campione con probabilità proporzionale alla dimensione demografica, mediante la procedura di selezione sistematica proposta da Madow.[3]

La selezione delle famiglie campione in ogni comune selezionato viene effettuata dalla lista anagrafica senza reimmissione e con probabilità uguali. La selezione dell'individuo campione all'interno di ciascuna famiglia è stata effettuata con probabilità uguali dalla lista degli individui eleggibili.

Nel prospetto 1 viene riportata la distribuzione dei comuni e delle famiglie dell'universo e del campione teorico per le quattro ripartizioni geografiche e le cinque tipologie comunali.

Prospetto 1 – Distribuzione dei comuni e delle famiglie nell'universo e nel campione per ripartizione geografica e tipologia comunale

	Comuni		**Famiglie**		
	Universo	Campione	Universo	Campione teorico	Campione effettivo
Ripartizione Geografica					
Nord-Ovest	3.061	180	7.029.432	2.156	2045
Nord-Est	1.480	133	4.879.752	1.483	1420
Centro	996	112	4.856.608	1.671	1614
Sud e Isole	2.557	240	7.875.954	2.723	2646
Tipologia Comunale					
A_1	12	12	3.988.170	1.587	1537
A_2	483	158	2.876.282	1.586	1427
B_1	6.645	202	7.165.635	1.642	1636
B_2	841	183	6.254.321	1.620	1516
B_3	113	110	4.357.338	1.598	1509
Totale Italia	8.094	665	24.641.746	8.033	7.725

[3] Madow, W.G. "On the theory of systematic sampling II", *Annals of Mathematical Statistics,* 20, (1949): 333-354.

4. La rilevazione e il trattamento dei dati

4.1 La tecnica di rilevazione e i contenuti informativi

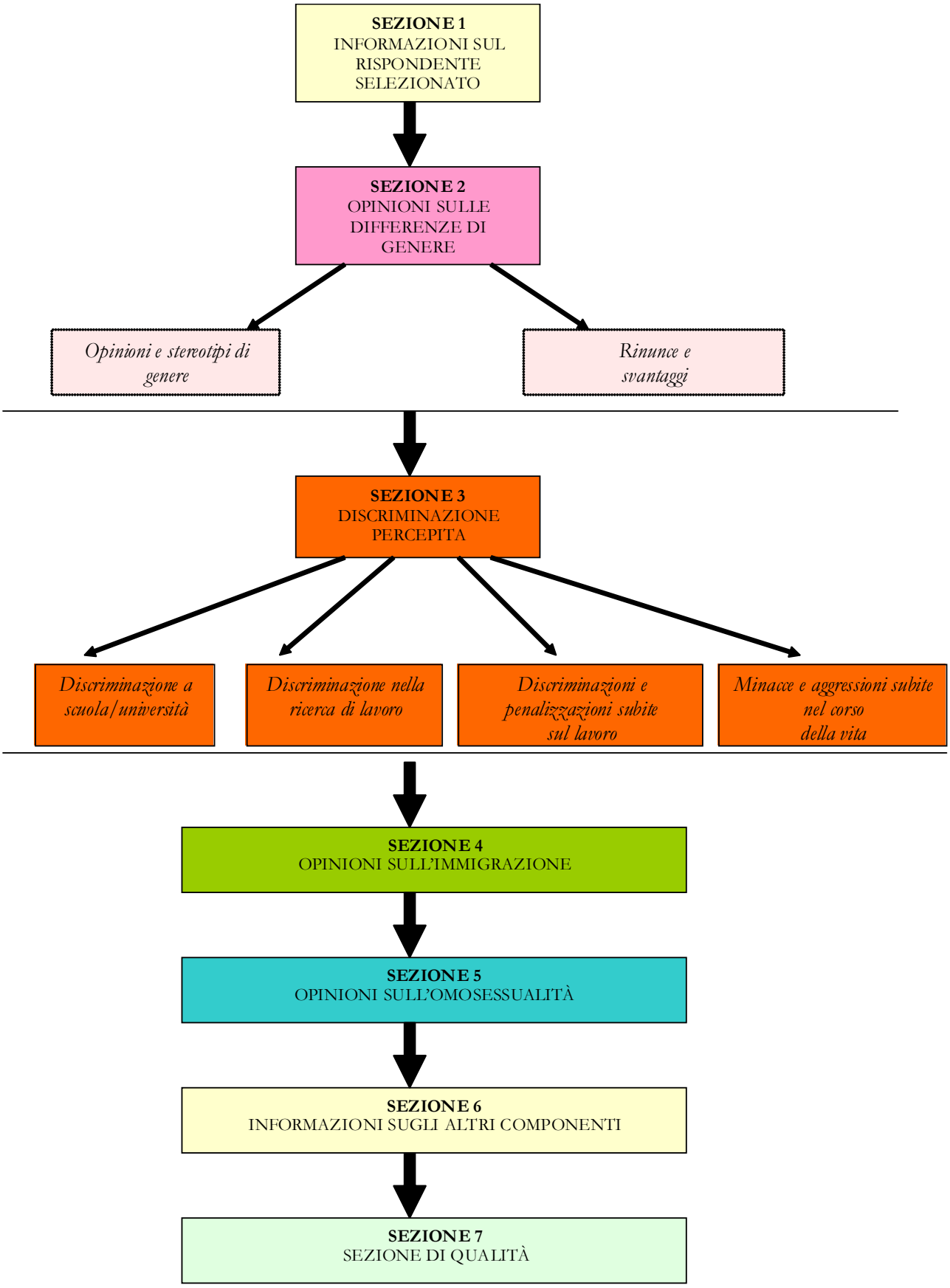
L'indagine è stata realizzata con tecnica mista CAPI – SAQ (Self Administered Questionnaire) tra giugno e dicembre 2011. La realizzazione delle interviste è stata affidata ad una rete di rilevazione privata. Le interviste sono state effettuate presso il domicilio della famiglia da intervistatori appositamente formati.

Per l'intervista CAPI, le informazioni sono state raccolte mediante un questionario elettronico, in cui registrare le risposte fornite dagli intervistati. Il questionario è stato progettato in modo da definire automaticamente i percorsi di compilazione e da prevedere controlli in corso di intervista, così da evitare incoerenze tra le risposte fornite e ridurre l'errore non campionario in fase di acquisizione dati.

Il questionario per intervista diretta è suddiviso in due parti. La prima è composta da una "Scheda contatto", che raccoglie le informazioni su tutti i contatti avvenuti con la famiglia, e da una Scheda Generale, contenente alcune informazioni di base sui componenti della famiglia, necessarie al software per procedere all'estrazione casuale della persona da intervistare. Nella seconda parte si susseguono sei sezioni tematiche, composte come segue (cfr. Figura 1):

- la sezione 1 e la sezione 6 raccolgono informazioni socio-demografiche relative rispettivamente al selezionato e agli altri componenti della famiglia;
- le sezioni 2, 4 e 5 raccolgono informazioni sulle opinioni, gli atteggiamenti, gli stereotipi, rispettivamente, sulle differenze di genere, l'immigrazione e l'omosessualità;
- la sezione 3 contiene quesiti sulle discriminazioni subite dall'intervistato/a nel contesto scolastico, durante la ricerca del lavoro e/o nello svolgimento di un'attività lavorativa;

Figura 1 - Diagramma di flusso del questionario



La metodologia di indagine prevedeva che, dopo aver somministrato l'intervista in modalità CAPI, l'intervistatore consegnasse all'intervistato un questionario cartaceo da autocompilare, allontanandosi in modo da consentire al rispondente di compilarlo in condizioni di massima riservatezza. Nel questionario cartaceo, attraverso una batteria di quesiti, è stato rilevato l'orientamento sessuale del rispondente e, nel caso di omosessuali e bisessuali, l'eventuale esperienza di *coming out* e le discriminazioni subite dall'intervistato/a in situazioni diverse da quelle già indagate nell'intervista CAPI (accesso ai servizi sanitari, accesso a locali pubblici, vicinato, etc.)

Una volta compilato, il questionario cartaceo veniva inserito, dallo stesso intervistato, in una busta da consegnare chiusa all'intervistatore. Quest'ultimo, in presenza dell'intervistato provvedeva a riporre la prima busta in una seconda busta, chiudendola con un'etichetta sigillante. Questa metodologia aveva come obiettivo quello di creare le condizioni ottimali per garantire al rispondente il rispetto della privacy e, quindi, metterlo a proprio agio nella compilazione dei quesiti mirati alla rilevazione dell'orientamento sessuale.

4.2 L'organizzazione del fieldwork

Considerata la natura sensibile dei temi trattati, particolare attenzione è stata posta a tutte le strategie da mettere in atto per favorire la disponibilità delle famiglie a essere intervistate. In particolare, alle famiglie campione è stata inoltrata una lettera di preavviso a firma del presidente dell'Istat, nella quale oltre a presentare la ricerca, si sottolineava l'importanza della loro partecipazione all'indagine e si fornivano rassicurazioni sulla protezione dei dati personali. Altro strumento messo in campo per migliorare la predisposizione delle famiglie alla partecipazione all'indagine è stata una linea telefonica gratuita (numero verde), alla quale le famiglie potevano rivolgersi per avere rassicurazioni in merito agli obiettivi dell'indagine, oltre che per ottenere informazioni sulla stessa.

Inoltre, nella progettazione dell'indagine, particolare cura è stata prestata alla formazione degli intervistatori coinvolti nel lavoro sul campo (oltre 150 intervistatori impegnati su tutto il territorio nazionale). Il percorso formativo ha mirato a condividere l'importanza e le finalità dell'indagine e a sviluppare competenze nella gestione del questionario elettronico e nella comunicazione con le famiglie campione. La formazione è consistita sia in lezioni teoriche, sia in esercitazioni e role-playings (ovvero giochi di ruolo in cui i partecipanti simulano condizioni simili a quelle reali).

Giornalmente gli esiti dei contatti con le famiglie e i file contenenti le interviste effettuate venivano trasmessi all'Istat. Questo flusso informativo giornaliero ha consentito di monitorare l'andamento del fieldwork, attraverso la produzione e l'aggiornamento continuo di un articolato sistema di indicatori che ha consentito di seguire il lavoro sul campo, ravvisando e risolvendo tempestivamente le difficoltà incontrate.

4.3 Il trattamento dei dati

Il processo di controllo e correzione è stato in larga parte semplificato dall'acquisizione della maggior parte delle informazioni tramite un questionario elettronico, predisposto in modo da prevedere la compilazione obbligatoria per la gran parte dei quesiti e controlli di coerenza tra le risposte fornite, contestualmente all'intervista. Inoltre, il data entry dei questionari cartacei è stato centralizzato presso la Società incaricata di effettuare le interviste ed è avvenuto tramite un software di acquisizione controllata dei dati, che consentiva di agganciare il modello autocompilato all'intervista CAPI corrispondente. Tali scelte hanno garantito un buon livello qualitativo del file grezzo, avendo consentito di ridurre il numero di mancate risposte parziali, di fuori range e di incoerenze da sanare in fase di trattamento dei dati.
Terminata la fase di acquisizione dei dati, sono state avviate le procedure di controllo e correzione degli stessi. L'individuazione degli errori è avvenuta tramite un articolato piano di controllo dei dati composto da procedure SAS di vario livello di complessità, mirate a ravvisare tutti i possibili errori presenti sul file dati. La priorità è stata data alle variabili socio-demografiche che rappresentano i pilastri su cui basare la successiva correzione degli eventuali errori riscontrati nelle variabile di indagine. In particolare, sono state applicate sia procedure di localizzazione e correzione deterministica degli errori sistematici, sia procedure di correzione e imputazione probabilistica degli errori stocastici, tramite procedure automatizzate basate sulla metodologia di Fellegi-Holt e implementate nel software generalizzato SCIA (Sistema di Controllo e Imputazione Automatica).

5. La metodologia di calcolo dei pesi campionari

L'indagine deve produrre le stime riferite al numero di individui che nella popolazione di riferimento (i 18-74enni, per questa indagine) 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 4.

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.

5.1 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 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;

π_j probabilità, assegnata dal disegno di campionamento, che l'unità j-esima sia inclusa nel campione S;

[4] 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.

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); dalla probabilità di estrazione della famiglia di appartenenza tra le famiglie eleggibili del comune; e della probabilità di essere selezionato come rispondente all'interno della famiglia estratta tra tutti gli individui tra i 18 e i 74 anni del nucleo familiare.

Per un generico individuo tra i 18 e i 74 anni, appartenente alla 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:

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

dove :

h denota l'indice di strato;

i è l'indice di comune;

j denota l'indice della famiglia;

n_{hij} indica il numero di componenti familiari tra i 18 e i 74 anni della famiglia *j* nel comune *i* dello strato *h;*

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.*

5.2 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. Nell'indagine in questione, l'utilizzo delle quartine ha fatto sì che il problema della mancata risposta totale si mantenesse a livelli molto bassi[5]: il campione finale è infatti pari a 7.725 unità rispetto alle 8.033 del disegno. Per ovviare comunque alla mancata partecipazione di alcune unità, sotto ipotesi che il comportamento dei rispondenti sia simile a quello dei non rispondenti all'interno dei comuni dello stesso strato, il correttore per mancata risposta assume la forma dell'inverso del tasso di risposta (δ_{hi}):

$$\frac{1}{\delta_{hi}} = \frac{m_{hi}}{m_{hi}^{r}} \qquad (5)$$

in cui m_{hi}^{r} rappresenta il numero di famiglie rispondenti nel comune *i* dello strato *h*[6]. In questa maniera, il coefficiente di riporto all'universo corretto per mancata risposta, da assegnare al campione rispondente, risulta essere:

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

5.3 La calibrazione a fonti esterne

Per il calcolo dei coefficienti di riporto all'universo finali si adottano gli stimatori di ponderazione vincolata (*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**) degli individui 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] Per ogni famiglia estratta, ne sono state estratte altre tre di profilo simile che potessero eventualmente sostituirla. Una famiglia estratta è quindi caduta se è caduta tutta la quartina.

[6] Nonostante il peso di riporto sia calcolato a livello individuale, la correzione per la mancata risposta è a livello familiare. Questo per due motivi: la famiglia è l'unità di rilevazione mentre l'individuo è l'unità di analisi; e perché il numero corretto di individui eleggibili all'interno di una famiglia eleggibile è determinato solamente in fase di acquisizione dati, ed è quindi disponibile per le sole famiglie rispondenti.

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

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. Le stime campionarie sono state vincolate ai seguenti totali noti:

- popolazione residente di 18-74 anni per ripartizione (Nord, Centro e Mezzogiorno), sesso e classi d'età (18-24, 25-34, 35-44, 45-54, 55-64, 65-74);
- popolazione residente di 18-74 anni per ripartizione (Nord, Centro e Mezzogiorno), sesso e tipologia del comune di residenza (Comuni centro di area metropolitana; comuni periferia di area metropolita; altri comuni sotto i 10.000 abitanti; altri comuni tra i 10.000 e i 50.000 abitanti; altri comuni sopra i 50.000 abitanti);
- popolazione residente di 18-74 anni per ripartizione (Nord, Centro e Mezzogiorno), sesso e cittadinanza (italiana; straniera).

5.4 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:

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

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

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:

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

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.

5.5 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:

$$\log\left(\hat{e}^2(\ddot{Y})\right) = a + b * \log(\ddot{Y})$$

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 2 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.

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:

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

Prospetto 2 - Valori dei coefficienti a, b e R2 delle funzioni utilizzate per l'interpolazione degli errori campionari delle stime

RIPARTIZIONE GEOGRAFICA	a	b	R^2
Italia	10.7525	-1.16909	90.3
Nord-Ovest	11.0984	-1.21438	89.1
Nord-Est	10.1691	-1.14800	86.9
Centro	11.0686	-1.21991	88.0
Mezzogiorno	9.6857	-1.10316	89.3

Infine, il prospetto 3 ha lo scopo di rendere più agevole e immediata la valutazione degli errori campionari. In testata sono elencati valori crescenti di stima (50.000, 100.000, 250.000,…) di frequenze assolute; 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 assoluta (o relativa), 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

RIPARTIZIONE GEOGRAFICA	STIME								
	50.000	100.000	250.000	500.000	750.000	1.000.000	2.500.000	5.000.000	10.000.000
Italia	38.7	25.8	15.1	10.1	8.0	6.7	3.9	2.6	1.8
Nord-Ovest	36.0	23.7	13.6	8.9	7.0	5.8	3.4	2.2	1.4
Nord-Est	32.4	21.8	12.9	8.6	6.9	5.8	3.4	2.3	1.5
Centro	34.5	22.6	12.9	8.5	6.6	5.5	3.2	2.1	1.4
Mezzogiorno	32.5	22.1	13.4	9.1	7.3	6.2	3.8	2.6	1.7

6. La diffusione dei risultati dell'indagine

I principali risultati dell'indagine sono stati diffusi attraverso le seguenti Statistiche Report: Stereotipi, rinunce e discriminazioni di genere, I migranti visti dai cittadini, La popolazione omosessuale nella società italiana, accessibili sul sito Istat.
Ricercatori e studiosi possono, inoltre, accedere al Laboratorio di Analisi dei Dati Elementari ADELE per effettuare di persona le proprie analisi statistiche sui microdati dell'indagine, nel rispetto delle norme sulla riservatezza dei dati personali.

7. Glossario

Discriminare significa trattare qualcuno in maniera meno favorevole di altri per alcune caratteristiche fisiche, mentali o altre caratteristiche personali che in sé non sono rilevanti ai fini dell'attività da svolgere o del contesto in cui ci si trova. Le caratteristiche personali possono riguardare il genere, l'età, le condizioni di salute, l'orientamento sessuale, etc.

8. Contatti

Unità operativa: DISA/B – "Progettazione e implementazione delle nuove reti di rilevazione e tecniche di indagine"
Maria Clelia Romano
e-mail: romano@istat.it
Alessandra Federici
e-mail: federici@istat.it

Curatori dei capitoli

Il documento è stato curato da Maria Clelia Romano e Alessandra Federici
Si devono a:
Claudia De Vitiis, i paragrafi 2 e 3
Andrea Cutillo, il paragrafo 5.

DESCRIZIONE
DEL FILE

FILE
DESCRIPTION

Istat
Istituto Nazionale di Statistica

Indagine sulle discriminazioni in base al genere, all'orientamento sessuale, all'appartenenza etnica

2011

Descrizione del file

Ottobre 2014

INDICE

1. Introduzione

I file di microdati ad uso pubblico (mIcro.STAT) sono collezioni campionarie di dati elementari relative ad alcune indagini svolte dall'Istat, per le quali siano già stati sviluppati i corrispondenti file di microdati per la ricerca (MFR).

Grazie ad una appropriata metodologia statistica volta a tutelare la riservatezza dei rispondenti, i mIcro.STAT possono essere scaricati liberamente e direttamente dal sito Istat.

Il file ad uso pubblico mIcro.STAT presenta le seguenti caratteristiche:

- è prodotto dal corrispondente file della ricerca attraverso tecniche di campionamento,
- la struttura, il livello di dettaglio ed il trattamento delle variabili sono "ereditate" dal corrispondente file per ricerca,
- a seguito del campionamento vengono calcolati i pesi di riporto all'universo da utilizzare per le analisi dei dati,
- precisione e accuratezza delle stime risultano inferiori rispetto a quelle ottenute utilizzando i microdati originali oppure il corrispondente file per la ricerca. Quindi è possibile che alcuni dati ottenuti elaborando il mIcro.STAT non coincidano con quanto già pubblicato dall'Istat.

Questo documento illustra le misure di protezione adottate al fine di ridurre il rischio di violazione della riservatezza.

La Sezione 2 fornisce alcune informazioni sintetiche sullo status di un insieme di variabili presenti nel file per la ricerca e nel mIcro.STAT, mentre la Sezione 3 contiene brevi cenni su ulteriori aspetti rilevanti per la riservatezza.

2. Descrizione delle variabili

Per il significato delle variabili e delle rispettive modalità si rinvia al questionario d'indagine. Si sottolinea che in alcuni casi il contenuto delle categorie può cambiare rispetto a quello originale, a parità di etichetta: nel prospetto seguente viene fornita indicazione delle variazioni che il mIcro.STAT "eredita" dal file per la ricerca.

Tabella 1. Trattamento delle variabili a fini di tutela della riservatezza nel file mIcro.STAT

VARIABILE	CODIFICA	DESCRIZIONE	RICODIFICA
RELPAR	1	Persona di riferimento (PR)	1
	2	Marito o moglie (coniuge) di PR	2
	3	Convivente di PR	2
	4	Genitore o coniuge/convivente del genitore di PR	4
	5	Suocero/a (Genitore o coniuge/convivente del genitore del coniuge o del convivente di PR)	4
	6	Figlio di PR nato dall'ultimo matrimonio (o convivenza)	6
	7	Figlio di PR o del coniuge o del convivente di PR nato da precedente matrimonio o convivenza	6
	8	Genero/Nuora (Coniuge del figlio di PR o del figlio del coniuge di PR)	6
	9	Convivente del figlio di PR (o del coniuge o convivente di PR)	6
	10	Nipote (figlio del figlio) di PR (o del coniuge o convivente di PR)	10
	11	Nipote (figlio del fratello/sorella) di PR (o del coniuge o convivente di PR)	10
	12	Fratello/sorella di PR	12
	13	Cognato/a: fratello/sorella del coniuge o convivente di PR	12
	14	Cognato/a: Coniuge del fratello/sorella di PR (o del coniuge o convivente di PR)	12
	15	Convivente del fratello/sorella di PR (o del coniuge o convivente di PR)	12
	16	Nonno/a	16
	17	Cugino/a	17
	18	Zio/a	17
	19	Altro parente di PR (o del coniuge o convivente di PR)	17
	20	Amico/a (altra persona convivente non legata da vincoli di parentela)	20
STCIV	1	Celibe/nubile	1
	2	Coniugato/a coabitante con il coniuge	2
	3	Separato/a di fatto (Coniugato/a non coabitante con il coniuge)	3
	4	Separato/a legalmente	3
	5	Divorziato/a	3
	6	Vedovo/a	6
ISTR	1	Nessun titolo e non sa leggere né scrivere	1
	2	Nessun titolo ma sa leggere e scrivere	1
	3	Licenza elementare/attestato di valutazione finale	3
	4	Licenza media (o avviamento professionale)/Diploma di istruzione secondaria di primo grado	4
	5	Diploma di qualifica professionale di scuola superiore di 2-3 anni che non permette l'iscrizione all'Università	5
	6	Diploma di maturità/Diploma di istruzione sec. sup. di 4-5 anni/Diploma intermedio di conservatorio o di danzatore	5
	7	Diploma di Accademia Belle Arti, Istituto Superiore Industrie Artistiche, ecc…	5
	8	Diploma universitario di due/tre anni, Scuola diretta a fini speciali, Scuola parauniversitaria	8
	9	Laurea di primo livello (corsi di 3 anni)	8
	10	Laurea specialistica (corsi di secondo livello di 2 anni)	10

	11	Laurea di 4 anni o più (vecchio ordinamento o laurea specialistica/magistrale a ciclo unico)	10
	12	Titolo di studio post-laurea (diploma di specializzazione, master universitario di 1° e di 2° livello)	12
	13	Titolo di dottore di ricerca	12
	14	Non sa/Non ricorda	14
POSIZ	1	Un lavoro alle dipendenze	1
	2	Un lavoro di collaborazione coordinata e continuativa (con o senza progetto)	2
	3	Un lavoro di prestazione d'opera occasionale	2
	4	Imprenditore	4
	5	Libero professionista	5
	6	Lavoratore in proprio	6
	7	Coadiuvante nell'azienda di un familiare	7
	8	Socio di cooperativa	7
CONDIZ	1	Occupati	1
	2	In cerca di nuova occupazione	2
	3	In cerca di prima occupazione	2
	4	Casalinghe	4
	5	Studenti	5
	6	Ritirati dal lavoro	6
	7	Cercano ma non attivamente e/o non disponibili	2
	8	Inabili al lavoro	8
	9	In altra condizione	8
	10	In altra condizione	8
NCOMP		Numero di componenti la famiglia	oltre 6 viene indicata la modalità 7

3. Metodologia statistica per la tutela della riservatezza

3.1. Apprezzamento del rischio

Sono variabili identificative indirette (o chiave) quelle che da sole, o in combinazione con altre, possono portare alla re-identificazione di uno o più record. In particolare, sono considerate l'identificazione da archivio esterno e l'identificazione spontanea sulla base di conoscenze a priori ("circolo delle conoscenze"). L'esistenza di un archivio esterno pubblico e nominativo, contenente informazioni associabili a quelle contenute nel file rilasciato, permetterebbe all'intruso di identificare l'unità statistica ed apprendere informazioni su di essa. Diversa è la fattispecie dell'identificazione riconducibile al "circolo delle conoscenze" ossia ascrivibile a informazioni sufficienti alla re-identificazione e note all'utente senza che venga posto in essere alcun tentativo di intrusione.

3.2. Protezione dei dati

Assimilando il rischio di violazione della riservatezza al costo di inclusione di una unità statistica nel campione (Casciano, Ichim, Corallo, 2011), il campionamento è stato

orientato alla minimizzazione del rischio e all'approssimativa riproduzione degli insiemi di totali riferiti alle combinazioni di variabili:

- ${}_1Y$ = RIP∩ETA∩L_DISCRIM
- ${}_2Y$ = RIP∩ETA∩C_DISCRIM
- ${}_3Y$ = RIP∩ETA∩S_DISCRIM
- ${}_4Y$ = RIP∩ETA∩D_DISCRIM
- ${}_5Y$ = RIP∩ETA∩I_DISCRIM
- ${}_6Y$ = RIP∩ETA∩O_DISCRIM
- ${}_7Y$ = RIP∩ETA∩L_DISCRIM
- ${}_8Y$ = RIP∩ETA∩C_DISCRIM
- ${}_9Y$ = RIP∩ETA∩S_DISCRIM
- ${}_{10}Y$ = RIP∩ETA∩D_DISCRIM
- ${}_{11}Y$ = RIP∩ETA∩I_DISCRIM
- ${}_{12}Y$ = RIP∩ETA∩O_DISCRIM

A seguito del campionamento sono stati aggiornati i pesi di riporto all'universo da utilizzare per le analisi dei dati.

4. Analisi del contenuto informativo

Sottolineando, ai fini interpretativi, che in presenza di caratteri qualitativi si debbono intendere quali variabili gli indicatori di presenza/assenza riferiti a ciascuna modalità, ponendo

$p=1,\ldots,P$	l'indice della variabile d'interesse,
d	il tipo di dominio,
$j_d=1,\ldots,J_d$	il dominio di tipo d,
${}_p\hat{Y}_{j_d}$	Il totale pesato del file per la ricerca, nel dominio j_d, della p^{ma} variabile,
${}_p\hat{\hat{Y}}_{j_d}$	Il totale pesato del mIcro.STAT, nel dominio j_d, della p^{ma} variabile,

si possono definire gli errori medi assoluti relativi :

$$emar_{p,d} \equiv \frac{1}{J_d}\sum_{j_d}\left|1-{}_p\hat{\hat{Y}}_{j_d}\Big/{}_p\hat{Y}_{j_d}\right|$$

A seguito della procedura di campionamento, per tutti i totali negli insiemi ${}_1Y$, …., ${}_{12}Y$, gli errori medi assoluti relativi sono risultati inferiori a 0.05.

5. Riferimenti bibliografici

Hundepool, Anco, e altri. 2007. Handbook on Statistical Disclosure Control. CENtre of EXcellence for Statistical Disclosure Control. http://neon.vb.cbs.nl/casc/handbook.htm. 04/09/2009.

Casciano, Maria Cristina, e Daniela Ichim, Laura Corallo. 2011. Sampling as a way to reduce risk and create a Public Use File maintaining weighted totals. Unece/Eurostat Work Session on Statistical Data Confidentiality. 26 - 28 Ottobre 2011, Tarragona (Spagna).

Falorsi, Piero Demetrio, e Marco Ballin, Claudia De Vitiis, Germana Scepi. 1998. Principi e metodi del software generalizzato per la definizione del disegno di campionamento nelle indagini sulle imprese condotte dall'Istat. Statistica applicata. 10(2). 235-257.

Curatore dei paragrafi

Il documento è stato redatto da Flavio Foschi.