La produzione dei micro file per la ricerca (MFR) e dei micro file a uso pubblico (PUF) in INAPP

Nota tecnica

Indice

Premessa[1]

Il Sistema statistico nazionale (SISTAN) è la rete di soggetti pubblici e privati che fornisce l'informazione statistica ufficiale in Italia, istituito con il Decreto legislativo n.322 del 1989.
L'INAPP, in qualità di ente di informazione statistica del SISTAN, è tenuto a seguire gli indirizzi e le direttive emanate dal COMSTAT (Comitato di indirizzo e coordinamento dell'informazione statistica).
In tale quadro, la direttiva n.11/2018 del COMSTAT adotta le "Linee guida per l'accesso a fini scientifici ai dati elementari del Sistema statistico nazionale" (in attuazione dell'art. 5-ter del decreto legislativo n.33/2013) che fissano le condizioni in base alle quali gli enti e gli uffici del SISTAN possono consentire ai ricercatori l'accesso per fini scientifici ai dati elementari, di cui sono titolari, privi di riferimenti che permettano l'identificazione diretta delle unità statistiche.
Da suddette linee guida deriva che l'accesso a dati elementari, privi di identificativi diretti, prodotti da Inapp è consentito, per fini scientifici, a ricercatori[2] appartenenti a Enti di ricerca riconosciuti dal COMSTAT (sulla base di criteri definiti all'interno delle medesime linee guida) o facenti parte dell'elenco degli Enti di ricerca riconosciuti da Eurostat (Regolamento (UE) n. 557/2013). L'accesso avviene tramite la fornitura di ***file per la ricerca*** (MFR) cui sono stati applicati metodi di controllo per la tutela della riservatezza.
I file MFR vengono comunicati ai ricercatori esclusivamente per il raggiungimento dei fini specificati nel Progetto di ricerca. La trasmissione dei MFR, in capo al Servizio statistico dell'Inapp, deve avvenire con modalità idonee ad assicurare la sicurezza dei collegamenti e l'autenticità degli interlocutori.

Obiettivo del presente documento è di descrivere le procedure da adottare per la produzione dei MFR. In dettaglio, partendo dalla definizione delle variabili identificative che possono essere utilizzate per associare le informazioni rilasciate e i rispondenti, vengono descritte le diverse fasi nelle quali si articola le regole per garantire la riservatezza da applicare per la produzione dei file MFR.
Le procedure statistiche a tutela della riservatezza da adottare per la produzione dei MFR descritte nel presente documento sono mutuate dalle procedure definite e impiegate dall'Istat per la produzione dei propri file MFR (Istat 2013).

Considerate le finalità di ricerca scientifica per le quali sono prodotti i MFR, i metodi di protezione utilizzati devono assicurare il mantenimento di un elevato contenuto informativo.

[1] La presente nota tecnica, a cura del Servizio statistico dell'Inapp, è stata redatta da Valentina Gualtieri.

[2] Sono considerati ricercatori ammessi a ricoprire la funzione di Ricercatore responsabile del progetto e ad accedere ai dati: i) i professori universitari (ordinari, associati, aggregati, a contratto), ricercatori o figure assimilabili (ad esempio tecnologi), assegnisti di ricerca di enti di ricerca riconosciuti; ii) responsabili degli enti/strutture di ricerca riconosciuti; iii) dipendenti di enti/strutture di ricerca riconosciuti che svolgono attività di ricerca; iv) soci di società scientifiche. Sono inoltre ricercatori ammessi a partecipare ad una Proposta di ricerca e ad accedere ai dati, le seguenti figure: v) dottorandi; vi) altri soggetti, con collaborazione formalizzata di ricerca con l'Ente riconosciuto.

I criteri di protezione statistica sono stabiliti a seguito di una valutazione d'impatto sulla protezione dei dati finalizzata a determinare i rischi per i diritti e le libertà delle unità statistiche, tenuto conto dell'eventuale coesistenza di rilasci di altri file di dati elementari che contengono dati riferiti alla stessa unità statistica - anche se già trattati ai fini della riservatezza - o di altre fonti liberamente accessibili, considerato che dal confronto tra più file dati elementari potrebbero ottenersi informazioni sui rispondenti, tali da invalidare le misure di protezione adottate. A seguito della valutazione d'impatto sulla protezione dei dati, gli interventi di tutela statistica della riservatezza debbono essere commisurati alla:

- probabilità dell'evento di *re-identificazione del rispondente* in rapporto al livello di dettaglio delle variabili (di seguito indicate come *identificativi indiretti*) le quali, considerate congiuntamente, permettono di circoscrivere la popolazione alla quale appartiene il rispondente, come ad esempio: l'età, il genere, il comune di residenza, l'occupazione, ecc.;
- conseguenza dell'evento intrusivo, tenendo conto delle caratteristiche del rispondente disponibili nei dati elementari e potenzialmente soggette ad intrusione (*attribute disclosure*).

In considerazione delle particolari garanzie che l'ordinamento riconosce a categorie particolari di dati (dati idonei a rivelare l'origine razziale o etnica, le opinioni politiche, le convinzioni religiose o filosofiche, lo stato di salute, la vita sessuale o l'orientamento sessuale, dati genetici e biometrici, dati relativi alle condanne penali, ai reati e alle connesse misure di sicurezza), qualora i file MFR, pur in presenza di un rischio solo residuale di re-identificazione dei rispondenti, contengano variabili idonee a rivelare tali informazioni, devono essere adottate apposite tecniche per assicurare l'anonimità di tali variabili.

I metodi di controllo per la tutela della riservatezza vengono adottati dall'INAPP tenendo conto delle procedure definite dall'Istat sulla base degli sviluppi metodologici sull'argomento a livello nazionale e internazionale e impiegate dall'Istat stesso per la produzione dei propri file MFR.

L'INAPP, mettendo a disposizione i file MFR, deve documentare le misure di protezione adottate. Le suddette documentazioni debbono essere rese disponibili, dietro richiesta, alla Commissione per la garanzia della qualità dell'informazione statistica. Le misure di protezione adottate devono essere rese note ai ricercatori dall'INAPP, ad esclusione dei parametri utilizzati per bilanciare riservatezza e utilità dei dati e delle informazioni che possano indebolire la protezione statistica dei dati (ad esempio le combinazioni di *identificativi indiretti* che sono state considerate, quali e quanti record di quali variabili sono stati sottoposti a misure di protezione).

Come previsto dalla normativa in materia di segreto statistico, inoltre, Inapp produce collezioni campionarie di dati elementari, **resi anonimi e privi di ogni riferimento che ne permetta il collegamento con singole persone fisiche e giuridiche** denominati **PUF (Public User File).**

I PUF-Inapp sono sviluppati per alcune particolari indagini a partire dai corrispondenti File per la ricerca, cui vengono applicate ulteriori tecniche di protezione della riservatezza che, necessariamente, implicano una riduzione del contenuto informativo.

Nel presente documento sono illustrate anche le procedure da adottare nella costruzione dei file PUF-Inapp.

In riferimento alla necessità di dover documentare gli step e le misure di protezione adottate per ciascun file dati creato, sia in forma di MFR che di PUF, nell'Allegato è riportato il format che viene usato, ovvero compilato, durante il processo di costruzione di ciascun file.

1. Le procedure statistiche a tutela della riservatezza per la produzione dei MFR[3]

Le procedure statistiche a tutela della riservatezza mirano ad individuare un compromesso tra mantenimento del contenuto informativo e protezione da eventi di intrusione. **La tutela statistica della riservatezza comporta sempre una perdita di contenuto informativo.**

1.2 Le variabili identificative

Le variabili identificative sono le variabili contenute in un file di dati elementari che possono essere utilizzate per associare le informazioni rilasciate e i rispondenti. Le variabili identificative si distinguono in:

- identificative dirette, come ad esempio nome e cognome, codice fiscale, recapito telefonico, ecc., che permettono di riconoscere il rispondente;
- identificative indirette o variabili chiave, che consentono di delimitare la popolazione alla quale appartiene l'interessato e che, se analizzate congiuntamente, possono favoriscono l'identificazione di un interessato.

1.2 Le fasi per l'applicazione delle regole di riservatezza

Le diverse fasi nelle quali si articola l'applicazione delle regole di riservatezza sono:

1. la struttura dei dati
2. identificativi diretti e variabili chiave
3. combinazione delle classi di età con le altre variabili chiave
4. protezione delle combinazioni di due variabili
5. combinazioni di variabili chiave
6. protezione delle combinazioni di più variabili
7. protezione delle variabili sensibili e giudiziarie
8. variabili quantitative

Di seguito si descrive ciascuna fase, ossia il percorso operativo e le operazioni da compiere in ciascun passo al fine dell'applicazione delle regole di riservatezza.

1. La struttura dei dati

Nelle indagini di tipo sociale, ricorrono usualmente tre tipi di file dati elementari:

[3] Le procedure descritte nel presente paragrafo sono una rielaborazione di quanto presente nel documento Istat "La produzione di MFR e mIcro.STAT in breve" (Istat 2013).

a) le informazioni riguardano soltanto singoli individui (ossia i dati non presentano una struttura gerarchica del tipo individui raggruppati in famiglie, alunni raggruppati in scuole, ecc.);
b) le informazioni su ciascun componente del gruppo (famiglia, scuola, ecc.) sono disposte su record distinti (dati con struttura gerarchica);
c) le informazioni su tutti i componenti del gruppo (famiglia, scuola, ecc.) sono elencate in un unico record (dati con struttura gerarchica).

Nel caso della struttura dati (c), è utile ricondursi alla struttura dati (b). Dopo aver completato il lavoro di protezione, si potrà ripristinare la struttura (c) originaria.

La gran parte dei file dati elementari prodotti a seguito delle rilevazioni condotte all'interno di indagini Inapp sono da ricondursi al caso a) individuato nella procedente classificazione.

2. Identificativi diretti e variabili chiave

- Individuare e rimuovere gli identificativi diretti e le variabili cosiddette "di lavoro" ossia relative alle modalità di raccolta dati (p. es., orario dell'intervista, identificativo dell'intervistatore, ecc.);
- Individuare le variabili chiave.

Si definiscono variabili chiave quelle che, anche per una sola modalità abbiano almeno una tra le seguenti caratteristiche:

- Rarità dei valori nella popolazione oggetto d'indagine

Quando una modalità ricorre raramente nella popolazione oggetto d'indagine, il rispondente che ne è caratterizzato appartiene a un gruppo ristretto di individui. Ad esempio, l'età molto elevata di uno dei residenti in un piccolo Comune italiano è senza dubbio una caratteristica rara. La rarità spesso dipende dalla popolazione di riferimento.

- Visibilità del carattere, o di alcune sue modalità, da parte di un osservatore;

La visibilità da parte dell'osservatore in alcuni casi e per alcune variabili, come ad esempio il genere, risulta ovvia, ma può anche riguardare alcune modalità di una data variabile, come accade ad esempio per il gruppo etnico di appartenenza e la professione (si pensi alle professioni che necessitano di una divisa).

- Tracciabilità in archivi esterni

La tracciabilità si riferisca alla possibilità di riscontrare determinate caratteristiche del rispondente sfruttando elenchi, registri o informazioni di pubblico dominio. Si pensi ad esempio alle variabili anagrafiche.

Nelle indagini di tipo sociale sugli individui alcune tra le principali variabili chiave da tenere sotto controllo sono:

- Comune, Provincia, Regione di residenza
- Comune, Provincia, Regione di domicilio
- Comune, Provincia, Regione di nascita
- Comune, Provincia, Regione dove si studia
- Comune, Provincia, Regione dove si lavora
- Stato estero di nascita
- Cittadinanza
- Età
- Genere
- Stato civile
- Titolo di studio
- Professione

Lo schema seguente facilita l'annotazione del tipo di variabile ai fini dell'applicazione delle regole di riservatezza.

Variabili	Identificativi diretti	Da non diffondere	Rarità	Visibilità	Tracciabilità	Categorie particolari (ex. variabili sensibili)
Codice indagine		x				
Codice Intervistato		x				
Codice fiscale	x					
ID Famiglia						
ID Individuo						
Regione				x	x	
Sub-regione (provincia, comune)				x	x	
Genere				x	x	
Età			x	x	x	
Peso			x	x		
Altezza			x	x		
Stato civile					x	
Livello di istruzione					x	
Condizione occupazionale				x	x	
Numero di componenti la famiglia			x	x	x	
Tipologia familiare				x	x	
Ricovero in ospedale negli ultimi 3 mesi						x
Numero di giorni di ricovero					x	x

3. Combinazione delle classi di età con le altre variabili chiave

Per individuare i casi unici più importanti, si calcolano le frequenze campionarie delle combinazioni costruite considerando insieme classi di età e ciascuna delle altre variabili chiave

(qualitative e/o quantitative discrete) riferite a singoli individui, prese una alla volta, come ad esempio:

- Classi di età *x* Stato civile;
- Classi di età *x* Cittadinanza;
- Classi di età *x* Titolo di studio;
- Classi di età *x* Professione;
- Classi di età *x* ciascuna delle variabili chiave rimanenti (una alla volta).

4. Protezione delle combinazioni di due variabili

REGOLA: per le combinazioni (o celle) considerate nella fase 3, caratterizzate da frequenza assoluta inferiore a f (con f>=2), si accorpano le classi di età e/o si pongono missing le modalità della "seconda" variabile (così facendo, a fronte di pochi casi unici, se ne mantiene il contenuto informativo).

Affinché le misure di protezione siano utili, occorre considerare i legami logici tra variabili: la soppressione di un valore è inefficace se quel valore può essere dedotto (con qualche approssimazione) considerando le rimanenti variabili (ad esempio, stato occupazionale e professione svolta).

5. Combinazioni di variabili chiave

Individuate *r* variabili chiave qualitative e/o quantitative discrete, sotto l'ipotesi che l'*intruder* (l'intruso, ovvero l'ipotetico soggetto che intende svolgere attività di re-identificazione dell'unità statistica) ne conosca al più *t* (*t* < *r*), è possibile formare $\binom{r}{t}$ combinazioni.

Ad esempio, con *r*=7 (Residenza, Genere, Classi di età, Stato civile, Cittadinanza, Titolo di studio, Professione svolta) e *t*=4 si ottengono $\binom{r}{t} = \binom{7}{4} = 35$ combinazioni.

Qualora si ritenga plausibile l'ipotesi che l'*intruder* in ogni combinazione mantenga fisse *j* delle *t* variabili, il numero di combinazioni si riduce a $\binom{r-j}{t-j}$. Ad esempio, con

- *r*=7 (Residenza, Genere, Classi di età, Stato civile, Cittadinanza, Titolo di studio, Professione svolta);
- *t*=4;
- *j*=3 (Residenza, Genere, Classi di età).

le combinazioni risultanti sono $\binom{r-j}{t-j} = \binom{4}{1} = 4$, ossia:

- Residenza x Genere x Classi di età x Stato civile;
- Residenza x Genere x Classi di età x Cittadinanza;
- Residenza x Genere x Classi di età x Titolo di studio;

- Residenza x Genere x Classi di età x Professione svolta.

Allo stesso modo, quando sono presenti variabili che caratterizzano i gruppi (famiglie, scuole, ecc.), è necessario calcolare le frequenze delle corrispondenti combinazioni.

Continuando nell'esempio precedente, se i gruppi fossero rappresentati dai nuclei famigliari, potrebbero essere formate combinazioni come:

- Residenza x Tipologia famigliare x Numero di componenti x Tipo di abitazione;
- Residenza x Tipologia famigliare x Numero di componenti x Numero di vani dell'abitazione.

Se nel file dati sono presenti variabili che rappresentano codifiche diverse di uno stesso "oggetto d'interesse" (ad esempio istruzione, attività economica, professione ecc.), occorre distinguere due casi:

- codifiche annidate (ad esempio NACE a 2 e a 3 digit); si deve considerare quella di maggiore dettaglio;
- codifiche non annidate; le variabili corrispondenti debbono essere inserite simultaneamente, ad esempio: Residenza x Genere x Classi di età x Professione svolta (codifica 1) x Professione svolta (codifica 2).

REGOLA: per ciascuna combinazione, fissati i parametri $k \in \{2,3\}$ e $p \in [0,0.1]$ occorre controllare se:

$$\frac{\text{n. individui di gruppi distinti che definiscono celle con frequenza} < k}{\text{n. totale di individui nel file dati}} < p \qquad \text{(a)}$$

e, quando sono presenti informazioni sui diversi componenti dei gruppi:

$$\frac{\text{n. gruppi con almeno un individuo che appartiene a celle con frequenza} < k}{\text{n. totale di gruppi nel file dati}} < p \qquad \text{(b)}$$

6. Protezione delle combinazioni di più variabili

Quando le precedenti condizioni (a) e (b) sono entrambe verificate si passa alla successiva fase 7. In caso contrario occorre modificare le modalità delle variabili chiave utilizzando:

➢ RICODIFICA GLOBALE[4]: le modalità delle variabili chiave vengono accorpate avendo cura che le classi che ne derivano siano:

[4] Quando sono contemporaneamente presenti più variabili che condividono le stesse categorie ed hanno un elevato numero di modalità (ad esempio luogo di residenza / domicilio / dimora / nascita / studio / lavoro), è consigliabile

- standard, ossia conformi a quelle adottate nelle pubblicazioni ufficiali e, nel caso di indagini armonizzate, siano riconducibili a quelle utilizzate da Eurostat. Le ricodifiche debbono essere statisticamente significative e permettere sempre il passaggio dalla classificazione più dettagliata a quella meno dettagliata (codifiche annidate).

È auspicabile che le nuove codifiche siano mantenute nel tempo sia per motivi di riservatezza, sia per permettere la confrontabilità tra dati rilevati in occasioni successive.

- coerenti, nel senso che le variabili legate concettualmente debbono essere aggregate in modo simile; ad esempio, se si aggrega l'età formando la classe 0-14, occorre anche evitare di fornire il dettaglio della frequenza scolastica per gli alunni di materne, elementari e secondarie di primo grado (tale informazione vanificherebbe la ricodifica 0-14 anni);
- di dettaglio non superiore ai domini di stima pianificati.

La ricodifica globale, essendo per definizione relativa a tutti i record (anche quelli che non presenterebbero criticità sotto il profilo della tutela della riservatezza), può comportare una perdita di dettaglio informativo consistente. Essa va attuata quando vi sono molti casi con frequenza inferiore alla soglia *k* fissata. Altrimenti conviene ricorrere alla

➢ SOPPRESSIONE LOCALE: in corrispondenza dei soli record che violano le condizioni (a) e/o (b), la modalità di una variabile chiave viene posta *missing*. Questo permette di evitare la ricodifica globale di tutti i record; ad esempio, se la regola (a) risulta violata perché un solo individuo manifesta la modalità "Nuova Zelanda" per la variabile chiave Stato di nascita, il solo *missing* che cancella la modalità "Nuova Zelanda" evita di dover sostituire la variabile Stato di nascita con la variabile meno dettagliata Area geografica di nascita.

Anche le soppressioni locali debbono essere coerenti, nel senso che la soppressione praticata su una certa variabile deve essere estesa a tutte quelle che – per via di legami logici - possono permettere di dedurne il valore; ad esempio, se si sopprime l'età, anche la variabile sulla frequenza scolastica deve essere posta *missing*.

➢ RICODIFICA LOCALE: si ottiene combinando ricodifica globale e soppressione locale. Una variabile viene riportata con due codifiche distinte e annidate (ad esempio, la Residenza espressa secondo Regione e Ripartizione). La soppressione locale viene applicata alle modalità più dettagliate (nell'esempio quelle della Regione).

7. Protezione di categorie particolari di dati

assegnare ad una variabile (per fissare le idee, luogo di residenza) il ruolo di riferimento per le rimanenti, utilizzando per tutte le altre ricodifiche come:

1 = Nello stesso Comune di residenza
2 = In altro Comune della stessa Provincia
3 = In altra Provincia della stessa Regione
4 = In altre Regioni
5 = Estero

Le “categorie particolari di dati” (precedente denominati “dati sensibili o giudiziari”) sono definite negli articoli 9 e 10 del GDPR (Regolamento UE 2016/679) e nel D.Lgs n. 196/2003 (Codice in materia di protezione dei dati personali) e s.m.i.

Nel dettaglio le categorie particolari di dati sono:

- dati personali idonei a rivelare l'origine razziale ed etnica, le convinzioni religiose, filosofiche o di altro genere, le opinioni politiche, l'adesione a partiti, sindacati, associazioni od organizzazioni a carattere religioso, filosofico, politico o sindacale, nonché i dati personali idonei a rivelare lo stato di salute e la vita sessuale;
- dati personali idonei a rivelare provvedimenti di cui all'articolo 3, comma 1, lettere da a) a o) e da r) a u), del d.P.R. 14 novembre 2002, n. 313, in materia di casellario giudiziale, di anagrafe delle sanzioni amministrative dipendenti da reato e dei relativi carichi pendenti, o la qualità di imputato o di indagato ai sensi degli articoli 60 e 61 del codice di procedura penale.

Il livello di protezione dei dati deve essere proporzionale al danno (potenziale) arrecato dalla violazione. In presenza di “categorie particolari di dati” conviene ricorrere a tecniche di casualizzazione.

Previa definizione degli strati di principale interesse per le stime (ad esempio Residenza x Genere x Classi di età), il modo più semplice di procedere per la casualizzazione può essere così descritto:

1. se n è l’ampiezza campionaria, si selezionano casualmente m record in modo che $m/n \in [0.15, 0.45]$;
2. se nello strato h ricadono almeno due degli m record selezionati, limitatamente alle variabili d’interesse si effettua lo scambio casuale delle modalità. È consigliabile che lo scambio tra i record riguardi simultaneamente tutte le variabili d’interesse in modo da conservarne la coerenza (ad esempio presenza di “malattia cronica” e patologia “artrosi”).

In questo modo sono mantenute le stime di strato riguardanti questa particolare tipologia di variabili.

La protezione proposta per le “categorie particolari di dati” può essere estesa anche a variabili non menzionate nel GDPR o nel D.Lgs n. 196/2003 e s.m.i, quando gli esperti dell’indagine lo ritengano opportuno.

8. Variabili quantitative

In presenza di:

- variabili quantitative continue;
- variabili quantitative discrete non considerate nelle fasi 5 e 6 (ad esempio il Numero di carte di credito possedute, il Numero di autovetture, ecc.);

strumenti molto utili alla tutela statistica della riservatezza sono:

- top coding e/o bottom coding;
- micro-aggregazione;
- arrotondamento.

8.1 Top e bottom coding

L'eventuale diradamento delle osservazioni in corrispondenza di valori molto piccoli o molto elevati suggerisce di raggruppare le unità statistiche collocate lungo le code della distribuzione: con il top coding vengono codificati i valori più elevati, mentre il bottom coding concerne le intensità più piccole. Ad esempio, i valori reddituali oltre i 3000 € possono essere raggruppati nella classe "oltre 3000 €" o, in alternativa, riportati tutti al valore di 3000 €.

In alcuni casi, in particolare per le variabili famigliari (come numero di componenti della famiglia o numero di figli), il top coding è efficace solo a condizione che non sia possibile risalire al numero di componenti o di figli contando i record o i campi individuo che hanno identico codice famigliare. Spesso una soluzione è rappresentata dall'eliminazione di tutti i record relativi agli individui oltre il sesto componente, oppure dalla soppressione di tutti i record famigliari.

8.2 Micro-aggregazione

Nel caso di variabili con un campo di variazione molto grande rispetto all'unità di misura adottata (ad esempio, le retribuzioni espresse in euro), soprattutto quando i dati d'indagine possano essere confrontati con archivi esterni, è consigliabile effettuare - oltre alla codifica dei valori estremali - la micro-aggregazione delle intensità rilevate: individuata una partizione (minimo 3 unità) molto fine dei dati, i valori originali vengono sostituiti con il valore medio di ogni gruppo di osservazioni in modo da mantenere il contenuto informativo dei dati originali[5].

8.3 Arrotondamento

In alcuni casi può risultare opportuno il ricorso all'arrotondamento dei valori rilevati. Ad esempio, con riferimento alla variabile Orario dell'incidente stradale, invece di fornire l'informazione completa è possibile rilasciare il dato arrotondato alla mezz'ora o all'ora, senza particolari conseguenze per le analisi da parte dei fruitori. Tuttavia, questa soluzione va utilizzata con molta attenzione, soprattutto in presenza di ordini di grandezza molto diversi: nel caso dei redditi da lavoro dipendente, il margine di incertezza indotto dall'arrotondamento ai 10 euro potrebbe essere sufficiente per retribuzioni di 1000 o 1500 euro ma non per retribuzioni di 10000 euro.

2. Le procedure statistiche a tutela della riservatezza per la produzione dei PUF (su individui e famiglie)

I PUF-Inapp sono file contenenti dati elementari resi anonimi e privi di ogni riferimento che ne permetta il collegamento con singole persone fisiche e giuridiche.

[5] In alternativa alla micro-aggregazione si può utilizzare la codifica dei valori rilevati mediante raggruppamento in classi secondo standard nazionali o internazionali.

I PUF-Inapp sono sviluppati per alcune particolari indagini a partire dai corrispondenti File per la ricerca (MFR), cui vengono applicate ulteriori tecniche di protezione della riservatezza che, necessariamente, implicano una riduzione del contenuto informativo.
Affinché i file MFR e i file PUF siano adeguatamente protetti, è **indispensabile che il PUF venga ricavato a partire dal MFR** e non dai dati (originali) d'indagine: in questo modo viene garantito che le codifiche adottate per i due file siano annidate.

Rispetto a quanto illustrato nelle pagine precedenti nelle fasi di costruzione dei MFR, per la costruzione dei file PUF occorre operare nel seguente modo.

- Si prende in input il MFR corrispondente
- Si ripercorrono le fasi da implementare (i.e. dalla fase 1 alla fase 8) prestando particolare attenzione alle **fasi 5 e 6**
  Nella fase 5 devono essere fissati: $k_{puf}>k_{mfr}$ e $p \in [0,0.01]$

Bibliografia

Istat (2013), La produzione di MFR e mIcro.STAT in breve. Indagini su famiglie e individui. Istituto Nazionale di Statistica (DIRM/DCME/MEA) https://www.istat.it/it/files/2013/12/Linee-guida-MFR-e-mIcro.STAT_.pdf

Direttiva n.11/2018 del COMSTAT
https://www.sistan.it/fileadmin/Repository/Home/NORME_E_PROCEDURE/ORGANIZZAZIONE_E_FUNZIONAMENTO/UFFICI_DI_STATISTICA/Direttiva%2011%20del%207%20novembre%202018.pdf

Pagina sito web Istat su microdati

https://www.istat.it/it/dati-analisi-e-prodotti/microdati

Allegato

MISURE DI PROTEZIONE ADOTTATE PER LA COSTRUZIONE DEL MFR

Denominazione Indagine:	
Denominazione banca dati di input:	
Denominazione banca dati di output (FILE MFR):	
Anno di riferimento banca dati:	
Unità di rilevazione:	

FASE 1: Struttura dei dati	
Le informazioni riguardano soltanto singoli individui (i dati non presentano una struttura gerarchica)	
Le informazioni su ciascun componente del gruppo (famiglia, scuola, ecc.) sono disposte su record distinti (dati con struttura gerarchica)	
Le informazioni su tutti i componenti del gruppo (famiglia, scuola, ecc.) sono elencate in un unico record (dati con struttura gerarchica	

FASE 2: Identificativi diretti e variabili chiave

Le principali variabili chiave da tenere sotto controllo sono: Comune, Provincia, Regione, Ripartizione (di residenza, domicilio, dimora, nascita, studio, lavoro); Cittadinanza; Stato (di nascita, residenza, studio, lavoro); Anno di Nascita; Età; Genere; Stato civile; Titolo di studio; Corso di laurea/diploma specifico; Ateneo di conseguimento del titolo; Professione.

Elenco variabili con identificativi diretti presenti nella banca dati di input:

Descrizione	Nome variabile

Elenco variabili chiave presenti nella banca dati di input:

Descrizione	Nome variabile

Elenco variabili da non diffondere (da eliminare per la banca dati di output):

Descrizione	Nome variabile

Definizione classi di età da adottare nelle fasi successive:

classi di età quinquennali	
classi di età decennali	
classi di età quindicinali	
classi di età "altro"	specificare:

FASE 3: Combinazione delle classi di età con le altre variabili chiave

Elenco delle combinazioni prese ad esame:

Classi di età *x* Provincia	Esempi da modificare rispetto alle specificità della banca dati e alle effettive combinazioni che si analizzano (coerenti con quanto elencato nella Fase 2)
Classi di età *x* Genere	
Classi di età *x* Stato civile	
Classi di età *x* Cittadinanza	
Classi di età *x* Titolo di studio	
Classi di età *x* Professione svolta	
Classi di età *x*	

Tabelle con frequenze fase 3

Inserire le tabelle di frequenza delle combinazioni esaminate

FASE 4: Protezione delle combinazioni di due variabili (eventuali accorpamenti)

REGOLA: per le combinazioni (o celle) considerate nella fase 3, caratterizzate da frequenza assoluta inferiore a f (con f>=2), si accorpano le classi di età e/o si pongono missing le modalità della "seconda" variabile (così facendo, a fronte di pochi casi unici, se ne mantiene il contenuto informativo).

f*=	inserire l'f* definito

Elenco delle combinazioni che presentano valori di f inferiore al valore predefinito

15-24 *x* Valle d'Aosta	Esempi da modificare rispetto a quanto ottenuto nella fase 3
15-24 x Licenza Elementare	

Descrizione dettagliata delle azioni intraprese

Inserire per ciascuna combinazione con f effettiva inferiore al valore f* predefinito l'azione intrapresa

FASE 5: Combinazioni di variabili chiave

r (numero variabili chiave)=	inserire r
t (numero variabili chiave note all'intruder)=	inserire t (t<r)
j (numero variabili chiave note all'intruder tenute fisse)=	inserire j (j<t)
combinazioni totali	(r-j t-j)
parametro k {2,3}=	inserire k

parametro p [0,0.1]=	inserire p

Elenco delle combinazioni prese ad esame:

Residenza x Genere x Classi di età x Stato civile	Esempi da modificare rispetto alle specificità della banca dati e alle effettive combinazioni che si analizzano (coerenti con i parametri riportati)
Residenza x Genere x Classi di età x Cittadinanza	
Residenza x Genere x Classi di età x Titolo di studio	
Residenza x Genere x Classi di età x Professione svolta	

Tabelle con frequenze fase 5

Inserire le tabelle di frequenza delle combinazioni esaminate evidenziando, solo per le celle che hanno f<k, quelle con (f/n)>=p

FASE 6: Protezione delle combinazioni di più variabili

Da effettuare sono nel caso in cui nella fase 5 sono presenti celle che oltre ad avere f<k hanno anche (f/n)>=p

Descrizione dettagliata delle azioni intraprese

Specificare se è stata fatta una ricodifica globale, una soppressione locale e/o una soppressione locale. Come è stato operato e su quali variabili

Tabelle con frequenze fase 6

Inserire le tabelle di frequenza a seguito delle modifiche apportate

FASE 7: Protezione di categorie particolari di dati

Da effettuare sono nel caso in cui della banca dati sono presenti variabili idonee a rivelare: l'origine razziale ed etnica, le convinzioni religiose, filosofiche o di altro genere, le opinioni politiche, l'adesione a partiti, sindacati, associazioni od organizzazioni a carattere religioso, filosofico, politico o sindacale, lo stato di salute e la vita sessuale, provvedimenti in materia di casellario giudiziale, di anagrafe delle sanzioni amministrative dipendenti da reato e dei relativi carichi pendenti, o la qualità di imputato o di indagato ai sensi del codice di procedura penale.

Elenco variabili riferite a categorie particolari di dati

Descrizione	Nome variabile

Tabelle con frequenze fase 7-1

Inserire le tabelle di frequenza delle variabili riferite a categorie particolari di dati

Descrizione dettagliata delle azioni intraprese

Specificare come è stata effettuata la casualizzazione

Tabelle con frequenze fase 7-2

Inserire le tabelle di frequenza delle variabili riferite a categorie particolari di dati a seguito della casualizzazione

FASE 8: Variabili quantitative

Da effettuare sono nel caso in cui della banca dati sono presenti variabili quantitative continue o discrete non considerate nelle fasi precedenti.

Elenco variabili continue

Descrizione	Nome variabile

Tabelle con min, max, media, mediana, q10, q25, q75, q90, sd

Inserire le tabelle per ciascuna variabile quantitativa

Descrizione dettagliata delle azioni intraprese

Specificare come ciascuna variabile quantitativa è stata lavorata (Top e bottom coding, micro-aggregazione, arrotondamento

Tabelle con min, max, media, mediana, q10, q25, q75, q90, sd

Inserire le tabelle per ciascuna variabile quantitativa a seguito dell'azione