La produzione di MFR e mIcro.STAT in breve

Indagini su famiglie e individui

Istituto Nazionale di Statistica (DIRM/DCME/MEA)
Luigi Virgili, Flavio Foschi

Sommario

1. Introduzione

La diffusione dei microdati d'indagine è sottoposta alle limitazioni previste dalla normativa vigente per proteggere la privacy dei rispondenti.

Le procedure statistiche a tutela della riservatezza mirano ad individuare un compromesso tra mantenimento del contenuto informativo e protezione da eventi di intrusione. Esse vengono aggiornate periodicamente secondo le migliori pratiche internazionali.

La tutela statistica della riservatezza comporta sempre una perdita di contenuto informativo.

2. Le variabili identificative

Le variabili identificative, cioè quelle che possono essere utilizzate per associare le informazioni rilasciate e i rispondenti, 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 circoscrivere la popolazione alla quale appartiene l'interessato; considerate congiuntamente favoriscono l'identificazione.

3. Il lavoro di tutela statistica della riservatezza

I diversi passi nei quali si articola l'applicazione delle regole di riservatezza sono esposti nei paragrafi successivi. Il percorso logico-operativo descritto riproduce sequenza e contenuto delle operazioni da compiere.

3.1. Fase 1 del lavoro: la struttura dei dati

Nell'esperienza Istat riguardante le indagini di tipo sociale, ricorrono usualmente tre tipi di data set:

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.

3.2. Fase 2 del lavoro: identificativi diretti e variabili chiave

1. Rimuovere gli identificativi diretti e le variabili cosiddette di lavoro/controllo ossia relative alle modalità di raccolta dati (Tipo orario dell'intervista, Nome dell'intervistatore, ecc.).
2. Individuare le variabili chiave.

Si definiscono variabili chiave[1] quelle che, anche per una sola modalità abbiano almeno una tra le seguenti caratteristiche:

[1] Nelle indagini di tipo sociale sugli individui alcune tra le principali variabili chiave da tenere sotto controllo sono:
- Comune, Provincia, Regione, Ripartizione di: residenza / domicilio / dimora / nascita / studio / lavoro,
- Densità abitativa per Km^2,
- Stato di nascita / cittadinanza / residenza / studio / lavoro,

- rarità dei valori nella popolazione oggetto d'indagine,
- visibilità del carattere, o di alcune sue modalità, da parte di un osservatore,
- tracciabilità in archivi esterni.

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, la cittadinanza lituana di uno dei residenti in un piccolo Comune italiano è senza dubbio una caratteristica rara. La rarità spesso dipende dalla popolazione di riferimento: ovviamente la cittadinanza lituana non sarebbe una caratteristica rara in Lituania.

La caratteristica di visibilità da parte dell'osservatore è ovvia per alcune variabili come genere, età, altezza, ma può anche riguardare alcune modalità di una data variabile, come accade ad esempio per il gruppo etnico di appartenenza e l'occupazione (si pensi alla divisa indossata da agenti di polizia, vigili del fuoco, operatori ecologici, ecc., nello svolgimento delle rispettive attività).

La tracciabilità concerne la possibilità di riscontrare determinate caratteristiche dell'unità statistica facendo uso di elenchi, registri o informazioni di pubblico dominio. Si pensi ad esempio alle variabili anagrafiche.

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

3.3. Fase 3 del lavoro: combinazione delle classi di età con le altre variabili chiave

Per individuare i casi unici più importanti (esempio, vedova 16enne), 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:

- 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 ciascuna delle variabili chiave rimanenti (una alla volta).

3.4. Fase 4 del lavoro: protezione delle combinazioni di due variabili

REGOLA: per le combinazioni (o celle) considerate nella fase 3, caratterizzate da frequenza assoluta inferiore a f (con $f \geq 2$), 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).

3.5. Fase 5 del lavoro: combinazioni di variabili chiave

Individuate r variabili chiave qualitative e/o quantitative discrete, sotto l'ipotesi che l'*intruder* ne conosca al più t ($t < r$), è possibile formare $\binom{r}{t}$ combinazioni.

- Classi di età,
- Genere,
- Stato civile,
- Titolo di studio,
- Corso di laurea/diploma specifico,
- Ateneo di conseguimento del titolo,
- Professione.

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 data set 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* e p come indicato nel successivo paragrafo 4, occorre controllare se

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

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

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

3.6. Fase 6 del lavoro: 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[2]: le modalità delle variabili chiave vengono accorpate avendo cura che le classi che ne derivano siano

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

– 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).

3.7. Fase 7 del lavoro: protezione delle variabili sensibili e giudiziarie

Le variabili sensibili e giudiziarie (di seguito, per brevità, sensibili) sono indicate nell'art. 4 del D.Lgs 30/06/03, n. 196 "Codice in materia di protezione dei dati personali"[3].

> L'art. 4 del D.Lgs 30/06/03, n. 196 ai punti d) ed e) definisce:
>
> d) "dati sensibili", i 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;
>
> e) "dati giudiziari", i 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 variabili sensibili conviene ricorrere alla

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 può essere così descritto:

1. se *n* è l'ampiezza campionaria, si selezionano casualmente *m* record in modo che $m/n \in [0.15, 0.45]$,

[3] http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/1311248.

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 le variabili sensibili. La protezione proposta per le variabili sensibili può essere estesa anche a variabili non menzionate nel D.Lgs 30/06/03 n. 196, quando gli esperti dell'indagine lo ritengano opportuno.

3.8. Fase 8 del lavoro: variabili quantitative

In presenza di

- variabili quantitative continue,
- variabili quantitative discrete non considerate nelle fasi 5 e 6 (come 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.

3.8.1. 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 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.

3.8.2. 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[4].

3.8.3. 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.

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

4. Specifiche per MFR e mIcro.STAT

Le indicazioni nei paragrafi precedenti sono valide per la produzione dei file MFR e mIcro.STAT, secondo le specifiche appresso riportate:

	MFR	mIcro.STAT
Data set di Input	Dati d'indagine	MFR
Parametri k e p (si veda la sez. 3.5)	$k_{MFR} \in \{2, 3\}$, $p \in [0, 0.1]$	$k_{mIcro} > k_{MFR}$, $p \in [0, 0.01]$
Fasi da implementare	1 – 8 (paragrafi da 3.1 a 3.8)	5 , 6 (paragrafi 3.5, 3.6)

Affinché i file MFR e mIcro.STAT siano adeguatamente protetti, è indispensabile che il mIcro.STAT 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.

5. Indagini articolate in diversi moduli

Nel caso di indagini che prevedono moduli a varie cadenze temporali, è consigliabile ragionare sulle misure di protezione considerando contestualmente le variabili presenti in tutti i moduli; in questo modo è possibile tenere conto organicamente delle relazioni tra tutte le variabili rilevanti per la tutela della riservatezza.

Qualora sia necessario diffondere i dati in tempi successivi, le misure di protezione applicate ai moduli già rilasciati non possono essere cambiate e diventano un vincolo per la protezione dei moduli (temporalmente) successivi. La ragione risiede nell'assoluta necessità di evitare rilasci multipli che possano vanificare le misure di protezione. A titolo di esempio, per comprendere cosa accadrebbe, si supponga che in occasione del rilascio del primo modulo le età dei rispondenti siano state codificate in classi decennali, mentre nel successivo rilascio – che comprende il secondo modulo – le età vengano ricodificate in classi quinquennali; il maggior livello di protezione indotto dalle classi decennali evidentemente non varrebbe per quanti abbiano accesso anche ai dati del secondo rilascio.

Il problema non sussiste quando nessuna tra le variabili che vengono via via aggiunte è riservata oppure utilizzabile come chiave per la re-identificazione; ad esempio se un modulo successivo rilevasse la variabile "utilizza autobus, filobus, tram all'interno del suo comune?", la nuova informazione potrebbe essere aggiunta senza la necessità di ulteriori controlli.

6. Documentazione

È opportuno corredare i file anonimizzati con i relativi metadati. Di seguito si ricordano quelli rilasciati dall'ISTAT per MFR e mIcro.STAT:

1. il tracciato record,
2. la nota metodologica dell'indagine,
3. il documento di descrizione del file,
4. l'abstract dell'indagine in italiano ed inglese,
5. il questionario dell'indagine.

È indispensabile, per ragioni di tutela della riservatezza, che i parametri utilizzati per la protezione dei data set non siano resi noti.

Riferimenti

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. e de Wolf, P.-P. (2012). Statistical Disclosure Control. Wiley.

Istat (2004). Metodologie e tecniche di tutela della riservatezza nel rilascio di informazione statistica (Metodi e norme, n. 20, 2004), http://www.istat.it/dati/catalogo/20040706_00/.

Willenborg, L. e de Waal, T. (1996). Statistical Disclosure Control in Practice. Lecture Notes in Statistics, 111, New York: Springer-Verlag.

Willenborg, L. e de Waal, T. (2000). Elements of statistical disclosure control. Lecture Notes in Statistics, 115, New York: Springer-Verlag.

Allegato A

Variables	Direct identifiers	Not to be released	Rareness	Visibility	Traceability	Sensitive variables (*DLgs n. 196/2003)*
Survey code		**x**				
Interviewer code		**x**				
Fiscal code	**x**					
Household sequential number						
Individual sequential number						
Region				**x**	**x**	
Sub-region				**x**	**x**	
Gender				**x**	**x**	
Age			**x**	**x**	**x**	
Weight			**x**	**x**		
Height			**x**	**x**		
Marital status					**x**	
Education level					**x**	
Working condition				**x**	**x**	
Number of household members			**x**	**x**	**x**	
Type of household				**x**	**x**	
Have you been hospitalized in the last 3 months						**x**
Number of days of hospitalization					**x**	**x**