



**File ad uso pubblico
micro.STAT**

Aspetti della vita quotidiana
2013

Descrizione del file

Febbraio 2015

INDICE

1. Introduzione	3
2. Descrizione delle variabili	4
3. Metodologia statistica per la tutela della riservatezza.....	6
4. Riferimenti bibliografici.....	14
5. Curatori.....	14

1. Introduzione

I file di microdati ad uso pubblico (mlcro.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 (MFR¹).

Grazie ad un'apposita metodologia statistica che garantisce la tutela della riservatezza dei rispondenti, questi possono essere liberamente scaricati direttamente dal sito Istat.

Il file ad uso pubblico mlcro.STAT relativo all'indagine multiscopo *Aspetti della vita quotidiana*, periodo di riferimento 2013, ha le seguenti caratteristiche:

- a) è prodotto dal corrispondente file della ricerca attraverso tecniche di sottocampionamento;
- b) la struttura, il livello di dettaglio ed il trattamento delle variabili sono ereditate dal corrispondente file per la ricerca, MFR;
- c) la precisione e l'accuratezza delle stime che si possono calcolare utilizzando il mlcro.STAT sono inferiori alle corrispondenti stime calcolabili sia a partire dai microdati originali sia a partire dal corrispondente file per la ricerca. Quindi è possibile che alcuni dati ottenuti elaborando il mlcro.STAT non coincidano con quanto già pubblicato dall'Istat.

Nella Sezione 2 sono fornite alcune informazioni sintetiche sulle modifiche subite dalle variabili nella predisposizione del mlcro.STAT. La metodologia per la tutela della riservatezza è brevemente descritta nella Sezione 3.

¹ MFR è l'acronimo per il file per la ricerca (Microdata File for Research). La documentazione metodologica di tali file è consultabile sul sito Istat (<http://www.istat.it/it/archivio/file-per-la-ricerca>).

2. Descrizione delle variabili

In questo paragrafo sono descritti in breve i cambiamenti indotti dal metodo di protezione dei dati adottato per l'indagine multiscopo *Aspetti della vita quotidiana*, periodo di riferimento 2013. Tali cambiamenti sono stati necessari per proteggere la riservatezza dei rispondenti. La protezione dei dati individuali è stata raggiunta modificando il contenuto informativo di alcune variabili.

Il file mlcro.STAT è fornito in formato delimitato da tabulatore.

Le unità statistiche del mlcro.STAT sono le famiglie residenti in Italia e i membri che le compongono.

Il separatore decimale è “.”.

Il file mlcro.STAT contiene 20275 records e 742 variabili.

Tabella 1. Lo status² delle variabili nel file mlcro.STAT³ modificate per motivi di tutela della riservatezza.

Nome Campo	Nome Variabile	Disponibilità	Modifica SDC
prov	provincia di rilevazione	SI	Soppressa
com	comune di rilevazione	SI	Soppressa
profam	progressivo famiglia univoco a livello indagine	SI	Sostituita con un numero fittizio
codril	codice della rilevazione	SI	Soppressa
eta	Età in anni compiuti	SI	Modificata a 18 modalità
stciv	Stato civile	SI	Modificata a 4 modalità
istr	Titolo di studio	SI	Modificata a 6 modalità
cond	Condizione professionale	SI	Modalità 6, 8 insieme
posiz	posizione nella professione	SI	Modificata a 4 modalità
tipfa2	Tipologia familiare 2	SI	Modificata a 33 modalità
peso	validata solo per le persone di 18 anni e più: peso in kg	SI	Modificata
stat	validata solo per le persone di 18 anni e più: statura in cm	SI	Modificata
bicbirra	consumo abituale al giorno di birra, in bicchieri	SI	Aggregazione per valori maggiori di 5

² Si fa riferimento esclusivamente alle misure di riduzione del rischio di violazione della riservatezza.

³ Il file mlcro.STAT conserva la struttura ed il livello di dettaglio del corrispondente file per la ricerca, si veda <http://www.istat.it/it/archivio/file-per-la-ricerca>.

bicvino	consumo abituale al giorno di vino, in bicchieri	SI	Aggregazione per valori maggiori di 5
bicfuori	complessivamente in una settimana quanti bicchieri di vino o alcolici consuma abitualmente fuori dai pasti?	SI	Aggregazione per valori maggiori di 20
bicaltro	numero di bicchieri al giorno di aperitivi alcolici, amari o superalcolici	SI	Aggregazione per valori maggiori di 5
nbicalc	consumo di bevande alcoliche negli ultimi 12 mesi, numero bicchieri	SI	Aggregazione per valori maggiori di 20
modco	modalità di compilazione del questionario autocompilato	SI	Soppressa
difco	completezza e difficoltà di compilazione del questionario autocompilato	SI	Soppressa
colf	la famiglia si avvale di: collaboratore/collaboratrice domestico/a	SI	Modalità 02, 23, 03 insieme
babys	la famiglia si avvale di: baby-sitter	SI	Modalità 02, 23, 03 insieme
assanz	la famiglia si avvale di: persona che assiste un anziano o un disabile	SI	Modalità 02, 23 e 03 insieme
nmotor	numero motorini, scooter	SI	Aggregazione per valori maggiori di 3
nmoto	numero di motociclette, moto	SI	Aggregazione per valori maggiori di 3
nauto	numero automobili	SI	Aggregazione per valori maggiori di 3
notfam	le notizie sulla scheda familiare sono state fornite da uno o da più componenti?	SI	Soppressa

3. Metodologia statistica per la tutela della riservatezza

3.1. Scenari di violazione della riservatezza

La valutazione statistica del rischio di violazione della riservatezza avviene attraverso diverse ipotesi su quali siano le intenzioni di un *intruder* e su quali informazioni abbia a disposizione.

Le unità statistiche a rischio di violazione della riservatezza sono state identificate ipotizzando i seguenti scenari di violazione della riservatezza:

- il primo definito dalle informazioni registrate in un archivio esterno pubblico e nominativo disponibile all'*intruder*, contenente informazioni associabili a quelle contenute nel file rilasciato, che permetterebbe all'intruso di identificare la persona di riferimento della famiglia (*pr*), la famiglia nel suo insieme e i componenti della famiglia.
- il secondo definito da un'identificazione spontanea causata dalla natura o dal dettaglio di registrazione di alcune variabili (valori anomali) presenti nell'archivio indagine.

Considerando il fatto che le unità statistiche nel file non sono indipendenti, ma sussistono relazioni di dipendenza fra le unità statistiche, ossia sussiste una relazione gerarchica fra un individuo e la famiglia a cui appartiene, si è ritenuto opportuno valutare il rischio di violazione della riservatezza utilizzando un approccio individuale, proposto da Benedetti e Franconi (1998), basato sul rischio di violazione definito in termini probabilistici per la stima del rischio individuale per tutti gli individui contenuti nel file da rilasciare.

3.2. Modifiche del contenuto informativo delle variabili

3.2.1. Soppressione delle variabili

Le variabili soppresse dal file mlcro.STAT per motivi di tutela della riservatezza sono elencate di seguito:

Tabella 2. Variabili soppresse per motivi di tutela della riservatezza.

Nome Campo	Nome Variabile
prov	provincia di rilevazione
com	comune di rilevazione
codril	codice della rilevazione
modco	modalità di compilazione del questionario

	autocompilato
difco	completezza e difficoltà di compilazione del questionario autocompilato
notfam	le notizie sulla scheda familiare sono state fornite da uno o da più componenti?

La soppressione di una variabile equivale alla completa cancellazione del suo contenuto nel file mlcro.STAT.

3.2.2. Ricodifica globale

La ricodifica globale è uno dei più semplici e noti metodi di protezione dei dati individuali.

Si può applicare solamente alle variabili categoriali. Essa consiste nell'accorpamento di alcune modalità della variabile categoriale. Ovviamente, le nuove modalità non hanno più il significato precedente.

Per il rilascio del file ad uso pubblico, mlcro.STAT, *Aspetti della vita quotidiana*, periodo di riferimento 2013, sono state ricodificate le seguenti variabili:

1. *eta* – La variabile età è stata rilasciata utilizzando la seguente classificazione:

Etichetta	Descrizione
1	0 – 2 anni
2	3 – 5 anni
3	6 – 10 anni
4	11 – 13 anni
5	14 – 17 anni
6	18 – 19 anni
7	20 – 24 anni
8	25 - 29 anni
9	30 – 34 anni
10	35 – 39 anni
11	40 – 44 anni
12	45 – 49 anni
13	50 – 54 anni
14	55 – 59 anni
15	60 – 64 anni
16	65 – 74 anni
17	75 – 84 anni
18	85 e più anni

2. *stciv* – La variabile stato civile è stata ricodificata con le seguenti modalità:

Etichetta	Descrizione
1	celibe/nubile
2	coniugato/a coabitante con il coniuge
3	coniugato/a non coabitante con il coniuge (separato/a di fatto) o separato/a legalmente o divorziato/a
6	vedovo/a

3. *tipfa2* – La variabile tipologia familiare è stata ricodificata con le seguenti modalità:

Etichetta	Descrizione
1	persona sola
2	genitore con figli non celibi o nubili
3	insieme di parenti
4	parenti ed altri
5	persone non parenti
6	coppia coniugata senza figli, senza isolati
7	coppia non coniugata senza figli senza isolati
8	coppia coniugata con figli senza isolati
9	coppia non coniugata con figli senza isolati
10	monogenitore maschio celibe senza isolati
11	monogenitore maschio separato di fatto senza isolati o monogenitore maschio separato legalmente senza isolati o monogenitore maschio divorziato senza isolati
14	monogenitore maschio vedovo senza isolati
15	monogenitore femmina nubile senza isolati
16	monogenitore femmina separata di fatto senza isolati o monogenitore femmina separata legalmente senza isolati o monogenitore femmina divorziata senza isolati
19	monogenitore femmina vedova senza isolati
20	coppia coniugata senza figli con isolati
21	coppia non coniugata senza figli con isolati
22	coppia coniugata con figli con isolati
23	coppia non coniugata con figli con isolati
24	monogenitore maschio celibe con isolati
25	monogenitore maschio separato di fatto con isolati o monogenitore maschio separato legalmente con isolati o monogenitore maschio divorziato con isolati
28	monogenitore maschio vedovo con isolati
29	monogenitore femmina nubile con isolati
30	monogenitore femmina separata di fatto con isolati o monogenitore femmina separata legalmente con isolati o monogenitore femmina divorziata con isolati
33	monogenitore femmina vedova con isolati
34	a due generazioni senza isolati

35	di tipo fraterno senza isolati
36	di altro tipo senza isolati
37	a due generazioni con isolati
38	di tipo fraterno con isolati
39	di altro tipo con isolati
40	nuclei senza isolati
41	con isolati

4. *posiz* – La variabile posizione lavorativa è stata ricodificata con le seguenti modalità:

Etichetta	Descrizione
1	dirigente o autonomo come imprenditore o libero professionista
2	direttivo, quadro o impiegato
3	capo operaio, operaio subalterno e assimilati o apprendista o lavorante a domicilio per conto d'impresa
4	lavoratore in proprio o socio cooperativa Produzione Beni e/o prestazioni di servizio o coadiuvante o collaborazione coordinata e continuativa (con o senza progetto) o prestazione d'opera occasionale

5. *istr* – La variabile titolo di studio più alto conseguito è stata ricodificata con le seguenti modalità:

Etichette	Descrizione
1	Dottorato di ricerca
2	Laurea (di II livello), Master di II livello e Scuola di specializzazione post-laurea o Diploma accademico di Alta Formazione Artistica, Musicale e Coreutica (A.F.A.M) di II livello o Laurea di 3 anni di I livello, compreso Master di I livello o Diploma accademico di Alta Formazione Artistica, Musicale e Coreutica (A.F.A.M) di I livello o Diploma universitario di 2- 3 anni
7	Diploma di maturità/Diploma di istruzione secondaria superiore di 4-5 anni o Diploma di qualifica professionale di 2-3 anni
9	Licenza media/Diploma di istruzione secondaria di I grado
10	Licenza elementare
11	Nessun titolo (sa leggere e scrivere) o Nessun titolo (non sa leggere e/o scrivere)

6. *cond* – La variabile condizione professionale è stata ricodificata con le seguenti modalità:

Etichette	Descrizione
1	occupato
2	in cerca di nuova occupazione
3	in cerca di prima occupazione
4	casalinga
5	studente
7	ritirato dal lavoro
8	altra condizione o inabile al lavoro

7. *colf* – La variabile la famiglia si avvale di collaboratore/trice domestico/a,
baby- la variabile la famiglia si avvale di: baby-sitter,
assanz – la variabile la famiglia si avvale di persona che assiste un anziano o
disabile,
sono state rilasciate utilizzando la seguente classificazione:

Etichette	Descrizione
1	no
2	sì, italiano o sia italiano che straniero o sì, straniero

8. *stat* – La variabile statura in cm e *peso* – la variabile peso in kg, sono state rilasciate
utilizzando la seguente classificazione:

- valori puntuali in corrispondenza della modalità 2, *Normopeso*, della variabile *BMI* (indice di massa corporea per le persone di 18 anni e più);
- valori accorpati nella modalità 777 in corrispondenza della modalità 1, *Sottopeso*, della variabile *BMI* ;
- valori accorpati nella modalità 888 in corrispondenza della modalità 3, *Sovrappeso*, della variabile *BMI* ;
- valori accorpati nella modalità 999 in corrispondenza della modalità 4, *Obeso*, della variabile *BMI* .

9. *nrica* – La variabile quante volte è stato ricoverato negli ultimi tre mesi?

i valori superiori a 4 sono stati ricodificati in 5, con significato “5 e oltre”.

10. *ggrica* – La variabile quante volte è stato ricoverato negli ultimi tre mesi? è stato
rilasciato utilizzando la seguente classificazione:

Etichetta	Descrizione
puntuali	Per i valori minore e uguali a 15
16	16 – 20
21	21 – 30
31	31 – 40
41	41 – 50
51	51 e oltre

Variabili relative ai mezzi di trasporto personali

Per le seguenti variabili:

- *nmotor*, numero motorini, scooter;
- *nmoto*, numero di motociclette, moto;
- *nauto*, numero automobili;

è stato applicato il metodo di top-coding, in particolare i valori uguali e superiori a 4 sono stati ricodificati in 4, con significato “4 e oltre”.

Variabili relative al consumo di bevande alcoliche

Invece, per le seguenti variabili:

- *bicbirra*, consumo abituale al giorno di birra, in bicchieri;
- *bicvino*, consumo abituale al giorno di vino, in bicchieri;
- *bicaltro*, numero di bicchieri al giorno di aperitivi alcolici, amari o superalcolici;

i valori superiori a 5 sono stati ricodificati in 6, con significato “6 e oltre”.

Per le variabili:

- *nbicalc*, consumo di bevande alcoliche negli ultimi 12 mesi, numero bicchieri;
- *bicfuori*, complessivamente in una settimana quanti bicchieri di vino o alcolici consuma abitualmente fuori dai pasti?

i valori superiori a 20 sono stati ricodificati in 21, con significato “21 e oltre”.

3.2.4 Campionamento effettuato per realizzare il mlcro.STAT

Il file mlcro.STAT è stato prodotto a partire dal file MFR eseguendo operazioni di sotto-campionamento, al fine di ridurre il rischio di violazione della riservatezza.

In particolare, il mlcro.STAT è ottenuto combinando due strumenti metodologici:

1. l'allocazione multivariata-multidominio di Bethel (Falorsi, Ballin, De Vitiis, Scepi, 1998), per determinare l'allocazione del campione;
2. il campionamento bilanciato utilizzando il metodo Cube (Deville e Tillé, 2004, 2005), per estrarre il campione.

Le variabili di allocazione sono state le variabili:

- $_1Y$: il totale di popolazione;
- $_2Y$: *ncomp*, numero dei componenti la famiglia attuale;
- $_3Y$: sesso.

La variabile di dominio è stata:

- d : $reg \cap dom$.

Il campionamento, pertanto, è stato orientato alla minimizzazione del rischio e all'approssimativa riproduzione, nel tipo di dominio $d \equiv reg \cap dom$, degli insiemi di totali $_1Y$, $_2Y$ e $_3Y$.

Il coefficiente di variazione è stato imposto al 4%.

I coefficienti di riporto all'universo (variabile *coeMicro*) sono stati aggiornati rispetto a quelli presenti nel relativo file della ricerca, MFR, per le analisi dei dati.

A seguito del campionamento per qualche variabile sono assenti alcune modalità presenti, invece, nel corrispondente file per la ricerca.

Sono stati definiti gli errori medi assoluti relativi delle stime del totale delle variabili $_1Y$, $_2Y$ e $_3Y$ ottenute utilizzando lo stimatore di Horvitz Thompson.

Le stime ottenibili dal file mlcro.STAT sono confrontate con le stime ottenibili dal corrispondente file di microdati di ricerca (MFR). Di conseguenza, gli errori medi assoluti relativi sono calcolati utilizzando la formula:

$$\frac{1}{J_d} \sum_{j_d} \left| \frac{\sum_{i \in MFR} y_i w_i - \sum_{j \in mlcro.STAT} y_j w_j}{\sum_{i \in MFR} y_i w_i} \right|$$

dove:

- y rappresentano i valori della variabile d'interesse;

- w rappresentano i coefficienti di riporto all'universo registrati nel file MFR e nel mlcro.STAT;
- d rappresenta il tipo di dominio;
- $j_d=1, \dots, J_d$ rappresenta il dominio di tipo d ;
- $\sum_{i \in MFR} y_i w_i$ rappresenta il totale pesato del file per la ricerca nel dominio j_d , della y variabile;
- $\sum_{j \in mlcro.STAT} y_j w_j$ rappresenta il totale pesato del file mlcro.STAT nel dominio j_d , della y variabile.

Ciò precisato, per i totali negli insiemi $_1Y$, $_2Y$ e $_3Y$, gli errori medi assoluti relativi sono risultati inferiori a 0.02.

4. Riferimenti bibliografici

1. 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/.
2. Willenborg, L. e de Waal, T. (2001). Elements of statistical disclosure control. Lecture Notes in Statistics, 115, New York: Springer-Verlag.
3. Bethel J. (1989), Sample Allocation in Multivariate Surveys, Survey Methodology, 15, pp 47-57.
4. Deville J., Tillé Y. (2004), Efficient Balanced Sampling: The Cube Method, Biometrika, Vol. 91, No. 4 (Dec., 2004) (pp. 893-912).
5. Foschi F. , Casciano C, Franconi L. , Ichim D., Designing Multiple Releases from the Small and Medium Enterprises Survey, J. Domingo-Ferrer and I. Tinnirello (Eds.): Privacy in Statistical Databases, PSD 2012, LNCS 7556, pp. 200–213, 2012, Springer-Verlag Berlin Heidelberg 2012.
6. Ichim D., Franconi L., Strategies to Achieve SDC Harmonisation at European Level: Multiple Countries, Multiple Files, Multiple Surveys J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 284–296, 2010 Springer-Verlag Berlin Heidelberg 2010.

Numerosi riferimenti bibliografici sul tema della tutela statistica della riservatezza sono reperibili sul sito <http://neon.vb.cbs.nl/casc/> nella documentazione prodotta nell'ambito dei diversi progetti europei che si sono occupati del tema: CASC project (2000-2003); CENEX project (2006); ESSnet project (2008-2009), ESSnet on common tools and harmonised methodology for SDC in the ESS, 2010-2012.

5. Curatori

Il documento è stato redatto da Laura Corallo.