Istat
Istituto Nazionale di Statistica

Censimento dei Laureati

Anno 2007

Descrizione del file

03/02/2012

INDICE

1. Introduzione

La necessità di tutela della riservatezza discende da vincoli di legge e da obblighi assunti verso i rispondenti. Seguendo *l'Handbook on Statistical Disclosure Control* (AA.VV., 2010), l'intrusione in informazioni non pubbliche si concretizza quando tramite i dati sia possibile acquisire notizie in merito a specifiche unità statistiche e può essere di due tipi: identificazione del rispondente o associazione ad una persona (o organismo) di dati noti dalle rilevazioni. Limitando l'attenzione al rilascio di microdati, l'intrusione avviene quando il singolo record è correttamente assegnato ad un record contenuto in un file esterno a disposizione dell'intrusore. Dunque la valutazione del rischio non può prescindere dalla considerazione di uno scenario idoneo a definire la quantità di informazioni di cui egli disponga. Le unità statistiche oggetto dell'analisi sono i laureati e, condizionatamente alla stratificazione di interesse, sono considerate a rischio quelle che, per una data combinazione di caratteri, risultano uniche nella popolazione. Mentre il secondo paragrafo è dedicato ad una breve descrizione delle variabili, il terzo dapprima discute il tema della valutazione del rischio, secondo gli scenari di intrusione ipotizzati e ritenuti ragionevoli, e successivamente illustra la strategia seguita per la protezione della confidenzialità. Il quarto paragrafo offre una sintesi dei risultati ottenuti.

2. Descrizione delle variabili

Le variabili raccolte nel censimento sono esposte nella tabella 1:

Tabella 1: variabili rilevate nel censimento dei laureati (fonte: Servizio Istruzione, Formazione e Lavoro)

variabile	descrizione
Progr_LL07	Progressivo della lista di partenza
Progr_IL11	Progressivo dell'indagine 2011 sull'inserimento professionale dei laureati
Ateneo	Ateneo di conseguimento della laurea
Sede_Facoltà	Sede della facoltà in cui è stata conseguita la laurea
Facoltà	Denominazione della facoltà
Tipologia_corso_LL07	Tipo di corso di laurea in cui è stato conseguito il titolo (5 modalità)
Tipologia_corso_IL11	Tipo di corso di laurea in cui è stato conseguito il titolo (raggruppamento a 3 modalità)
Gruppo_disciplinare	Gruppo disciplinare di afferenza del corso
Area_disciplinare	Area disciplinare di afferenza del gruppo
Classe_laurea	Classe del corso di laurea
Corso_LL07	Denominazione del corso di laurea
Classe_Corso_IL11	Aggregazione di classi/corsi per l'indagine 2011 sui laureati

Tabella 1: continua

variabile	descrizione
Sede_didattica_Regione	Regione dove si sono tenute le lezioni del corso di laurea
Sede_didattica_Provincia	Provincia dove si sono tenute le lezioni del corso di laurea
Sede_didattica_SLL01	Sistema locale del lavoro (2001) di appartenenenza del comune dove si sono tenute le lezioni del corso di laurea
Sede_didattica_Comune	Comune dove si sono tenute le lezioni del corso di laurea
Sesso	Sesso
Cittadinanza	Cittadinanza del laureato
Paese_cittadinanza	Paese di cittadinanza del laureato (se non italiano)
Residenza_Regione	Regione di residenza del laureato risultante presso l'ateneo
Residenza_Provincia	Provincia di residenza del laureato risultante presso l'ateneo
Residenza_SLL01	Sistema locale del lavoro (2001) di appartenenza del comune di residenza
Residenza_Comune	Comune di residenza del laureato risultante presso l'ateneo
Confr_Sede_didattica_Residenza_Locale	Confronto a livello locale tra sede didattica di studio e residenza del laureato
Confr_Sede_didattica_Residenza_Provinciale	Confronto a livello provinciale tra sede didattica di studio e residenza del laureato
Campione_IL11	Laureati estratti nel campione per l'indagine 2011
Esito_IL11	Esiti dei tentativi di contatto sui laureati estratti nel campione per l'indagine 2011
Dettaglio_esito_IL11	Dettaglio degli esiti di contatto

Le prime due e le ultime tre contengono alcune informazioni relative alla costruzione del campione afferente *l'Indagine sull'Inserimento Professionale dei Laureati*. Sono possibili alcune osservazioni:

- il codice *Ateneo* è un sottoinsieme di *Sede_Facoltà*;
- *Tipologia_corso_LL07* esprime un dettaglio di *Tipologia_corso_IL11*;
- le mutabili *Classe_Corso_* sono logicamente legate tra loro e con le rispettive *Tipologia_corso_* ;
- i primi due digit di *Sede_didattica_Provincia* coincidono con i corrispondenti di *Ateneo*, *Sede_Facoltà*, *Sede_didattica_Comune*;
- *Area_disciplinare* è una forma di sintesi di *Gruppo_disciplinare*;
- *Cittadinanza* è la controparte dicotomica di *Paese_cittadinanza*;
- le mutabili *Sede_didattica_* sono geograficamente annidate così come le omologhe *Residenza_*.

Approfondimenti ulteriori in ordine alle articolazioni delle variabili eccedono gli scopi del presente lavoro e per essi si rinvia alla documentazione resa disponibile dalla *Direzione Centrale delle Statistiche Socio-Economiche*. È invece opportuno ricordare che il

campione afferente *l'Indagine sull'Inserimento Professionale dei Laureati* è stato estratto in modo da essere rappresentativo degli strati *Cittadinanza* (Italiana o straniera), combinazioni di *Ateneo*, *Tipologia_corso_IL11* (ricondotta a due modalità) e *Area_disciplinare*, combinazioni di *Classe_corso_IL11* e *Sesso*. Queste considerazioni strutturali suggeriscono due possibili contesti di riferimento; il primo attiene al contributo di singole variabili alle identificazioni spontanee o accidentali; nel secondo, si assume venga perseguito intenzionalmente l'abbinamento tra i dati di indagine e una lista di microdati disponibile secondo un prefissato livello di dettaglio. La riduzione del rischio di violazione della riservatezza è stata perseguita combinando soppressioni di variabili e campionamento dei record.

3. Metodologia statistica per la tutela della riservatezza

3.1 Apprezzamento del rischio

In ordine al primo ambito di riferimento, concernente le cosiddette intrusioni accidentali, l'identificazione origina "spontaneamente" dall'osservazione di un ristretto numero di variabili, alcune delle quali aventi un livello di dettaglio sufficiente ad isolare poche unità statistiche in base alla sola distribuzione marginale di frequenze. Benché le principali mutabili espressive di informazioni facilmente accessibili non siano sufficienti ad ingenerare casi unici, il numero di modalità di alcune ha richiesto ulteriori verifiche in ordine agli incroci con un modesto numero delle rimanenti. Con riguardo al secondo ambito, le unità statistiche a rischio di violazione della riservatezza sono state identificate ipotizzando uno scenario di abbinamento tra dati di indagine e una lista di microdati disponibile ad un terzo. La selezione dello scenario è stata orientata da considerazioni di opportunità in ordine alla possibilità di tenere conto, secondo un livello di dettaglio ragionevole, delle tipologie di informazione offerte dai dati di censimento circa cittadinanza, sesso, ateneo, facoltà e residenza.

3.2 Protezione dei dati

In termini molto generali, si assuma che i dati di popolazione siano organizzati in N record ed M variabili ${}_1Z,\ldots,{}_MZ$; i vincoli sull'utilità dei dati siano espressi in termini di totali di popolazione per le variabili $\{{}_1Y,\ldots,{}_pY\}\subset\{{}_1Z,\ldots,{}_MZ\}$ all'interno dei domini di stima $\{D_1,\ldots,D_E\}\subset\mathcal{P}({}_1Z,\ldots,{}_MZ)$, con $\mathcal{P}$ insieme di potenza dell'argomento. La concreta individuazione dei sottoinsiemi $\{{}_1Y,\ldots,{}_PY\}$ e $\{D_1,\ldots,D_E\}$ deve rispondere a criteri individuati dai responsabili della rilevazione. Il campionamento stratificato persegue il miglioramento

della precisione degli stimatori dei parametri rispetto al campionamento delle unità elementari dall'intera popolazione. L'allocazione multivariata multidominio di Bethel (Falorsi, Ballin, De Vitiis, Scepi, 1998) è volta alla minimizzazione del costo associato all'inclusione di unità nel campione, condizionatamente al vincolo rappresentato – per ciascuna variabile d'interesse – dal contenimento della varianza dello stimatore del parametro entro le soglie fissate in riferimento ai domini. Definendo

h	il generico strato
n_h	il numero di unità del campione allocate nello strato h,
N_h	il numero di unità della popolazione facenti capo allo strato h,
c_h	il costo unitario nello strato h,
S_h	la deviazione standard della popolazione nello strato h (ad esempio stimata sulla base di precedenti rilevazioni),
$p=1,\ldots, P$	l'indice della variabile d'interesse,
$d=1,\ldots, D$	il tipo di dominio,
$j_d=1,\ldots, J_d$	il dominio di tipo d,
H_{jd}	il numero di strati contenenti il dominio j_d,
V	la varianza dell'argomento (nel caso in esame, il totale della variabile d'interesse: $\tilde{Y}$),
V^*	il massimo valore ammissibile per V,

le numerosità ottime sono individuate dal seguente problema di programmazione convessa:

$$n_h = \arg\min \sum_{h=1}^{H} c_h n_h \text{ sub: } V({}_p\tilde{Y}_{j_d}) \equiv \sum_{h=1}^{H_{jd}} \frac{N_h^2}{n_h} {}_pS_h^2 - \sum_{h=1}^{H_{jd}} N_h {}_pS_h^2 \leq {}_pV_{j_d}^* \qquad (1)$$

L'adattamento dell'algoritmo di Bethel a problemi di tutela statistica della riservatezza (per brevità, *SDC* secondo l'acronimo inglese di *Statistical Disclosure Control*) è possibile combinando due accorgimenti (Casciano, Ichim, Corallo, 2011):

- considerando il costo c_h funzione del numero di unità a rischio in h,
- includendo nella definizione di ciascuno strato h la mutabile dicotomica che induce la partizione in unità a rischio e non.

Così facendo, l'algoritmo viene orientato a selezionare le unità da includere nel campione attingendo di preferenza dagli strati meno costosi in termini di rischio, via via incrementando la dimensione campionaria fino al soddisfacimento dei vincoli. Questi ultimi ed i costi di strato rappresentano in definitiva i fattori che definiscono il trade-off tra utilità e protezione dei dati, condizionatamente allo scenario di intrusione ipotizzato. La tabella 2

evidenzia le principali scelte operate ai fini dell'applicazione, al censimento dei laureati, della metodologia descritta.

Tabella 2: mutabili utilizzate nel campionamento SDC

Mutabili di strato:	*Ateneo*; *Tipologia_corso_IL11*; *Gruppo_disciplinare*; *Cittadinanza*; *Sesso*; *Rischio.*
Mutabili di dominio:	*Ateneo*; *Tipologia_corso_IL11*; *Gruppo_disciplinare*; *Cittadinanza*; *Sesso*.
Mutabili di allocazione:	*Confr_Sede_didattica_Residenza_Locale.*

In maggior dettaglio, sono stati considerati tutti gli incroci delle mutabili di strato mentre ciascuna di quelle di dominio è stata utilizzata come tipo di dominio: nella simbologia precedentemente introdotta, D=5 e j_d è l'indice dei domini all'interno del d^{mo} (d=1,…, D) tipo di dominio. *Rischio* è una mutabile dicotomica che indica la condizione di rischio mentre *Confr_Sede_didattica_Residenza_Locale* – ritenuta d'interesse per le analisi svolte dagli utilizzatori dei dati – è articolata in quattro categorie. Affinché gli esiti del campionamento volto alla tutela della riservatezza non risultino vanificati, risulta inoltre necessaria la soppressione di alcune variabili.

Tabella 3: variabili soppresse

Progr_LL07
Progr_IL11
Sede_Facoltà
Facoltà
Tipologia_corso_LL07
Classe_laurea
Corso_LL07
Sede_didattica_SLL01
Sede_didattica_Comune
Paese_cittadinanza
Residenza_SLL01
Residenza_Comune
Campione_IL11
Esito_IL11
Dettaglio_esito_IL11

Facendo riferimento alla tabella 3, si osservi che fatta eccezione per le prime due e le ultime tre variabili soppresse, parte rilevante del contenuto informativo delle rimanenti è mantenuta dalle mutabili di strato, allocazione e dominio: esse ne rappresentano sostanzialmente una ricodifica, secondo un diverso livello di dettaglio, bilanciando le esigenze di utilità dei dati con le necessità di tutela della confidenzialità.

4. Analisi del contenuto informativo

Alcune statistiche sono utili a riassumere gli esiti delle azioni intraprese. Ponendo

N	la numerosità della popolazione
${}_pY$	la p^{ma} variabile d'interesse
$p=1,\ldots,P$	l'indice della variabile d'interesse,
$d=1,\ldots,D$	il tipo di dominio,
$j_d=1,\ldots,J_d$	il dominio di tipo d,
$r_i \in \{0, 1\}$	l'indicatore dicotomico di rischio per l'i^{ma} unità statistica,
$\delta_i \in \{0, 1\}$	l'indicatore dicotomico di inclusione nel campione per l'i^{ma} unità statistica,
π_i	la probabilità di inclusione per l'i^{ma} unità statistica ricavata dalla soluzione di (1)

si possono definire il *Tasso di Protezione* e *l'Errore Relativo Assoluto*:

$$TP \equiv 1 - \sum_{i=1}^{N} r_i \delta_i \Big/ \sum_{i=1}^{N} r_i$$

$$ERA_{{}_pY_{j_d}} \equiv \left| 1 - \sum_{i \in j_d} \frac{{}_pY_i}{\pi_i} \delta_i \Big/ \sum_{i \in j_d} {}_pY_i \right|$$

Ai fini interpretativi, occorre sottolineare che in presenza di caratteri qualitativi si debbono intendere quali variabili gli indicatori di presenza/assenza riferiti a ciascuna modalità. Facendo uso del software R (R Development Core Team, 2010) e del package SamplingStrata (Barcaroli, 2011), sono stati ottenuti i risultati sintetizzati nella tabella 4.

Tabella 4 esiti del campionamento SDC

numerosità del campione	214438			
TP conseguito	0.741			
95mo percentile degli *ERA* (*) calcolati su tutti i domini j_d	0.0431	0.0517	0.0177	0.1210

(*): le quantità sono riferite alle quattro modalità della mutabile di allocazione

5. Riferimenti bibliografici

AA.VV. *Handbook on Statistical Disclosure Control*. ESSnet on Statistical Disclosure Control, 2010. http://neon.vb.cbs.nl/casc/handbook.htm. 18/08/2011.

Barcaroli G. (2011). *SamplingStrata: Optimal stratification of sampling frames for multipurpose sampling surveys.* R package version 0.9-1. URL http://CRAN.R-project.org/package=SamplingStrata.

Casciano C., Ichim D., Corallo L. (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, P.D., Ballin M., De Vitiis C., Scepi G. (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.

6. Contatti

Per la tutela della riservatezza: **rilascio.microdati@istat.it** (DIQR/DCIQ/PSS)

Curatore

Flavio Foschi
DIQR/DCIQ/PSS/1
Piazza Indipendenza 4, 00185 Roma, foschi@istat.it