ISTAT

ISTITUTO NAZIONALE DI STATISTICA

INDAGINE LONGITUDINALE SUGLI SBOCCHI PROFESSIONALI DEI LAUREATI
Anni 1989-1991

DOCUMENTAZIONE TECNICA E DESCRIZIONE
DEL FILE STANDARD

PREMESSA

Il Decreto Legislativo n° 322 del 6/9/1989 regola la diffusione delle informazioni statistiche prodotte nell'ambito del Sistema Statistico Nazionale al fine di garantire la riservatezza dei rispondenti.

In particolare, per la diffusione di dati elementari, l'articolo 10, comma 2, dispone quanto segue: "Sono distribuite altresì, ove disponibili, su richiesta motivata e previa autorizzazione del Presidente dell'ISTAT, collezioni campionarie di dati elementari, resi anonimi e privi di ogni riferimento che ne permetta il collegamento con singole persone fisiche e giuridiche".

Nell'osservanza di tale legge l'ISTAT ha adottato misure e tecniche che rendono impossibile, o altamente improbabile, il collegamento dei dati rilasciati con l'unità statistica a cui si riferiscono. Per tale motivo sono state apportate alcune modifiche sui files originali delle indagini, nell'intento di garantire la massima protezione ai dati contenendo al minimo l'eventuale perdita di informazione.

Le metodologie applicate si concretizzano nell'accorpamento e/o riclassificazione di modalità di variabili e nell'oscuramento di variabili. In quest'ultimo caso, nei campi del tracciato record è riportata la dicitura "RISERVATO ISTAT".

Con l'occasione si ricorda al richiedente che si impegna a:

- utilizzare i dati soltanto per gli scopi dichiarati;
- non fornire a terzi i dati elementari, consentendone l'accesso, sotto la propria responsabilità, soltanto alle persone direttamente coinvolte nel lavoro per il quale essi sono stati richiesti;
- citare la fonte ISTAT nell'eventuale divulgazione di elaborazioni dei dati;
- inviare alla Biblioteca dell'ISTAT due copie delle pubblicazioni eventualmente prodotte con l'utilizzo dei dati ottenuti.

Obiettivi e tecnica di indagine

Nell'anno 1991 l'Istituto Nazionale di Statistica ha effettuato, per la prima volta, un'indagine di tipo longitudinale sugli sbocchi professionali dei laureati, che segue quella di tipo "trasversale" realizzata nel 1989 [1] con la quale sono state descritte le modalità di accesso dei laureati al mercato del lavoro.

Obiettivo dell'indagine longitudinale è quello di fornire un'analisi dinamica del processo di inserimento dei laureati nel mondo del lavoro, descrivendo i cambiamenti nella condizione occupazionale nel periodo 1989-1991 e analizzando la tendenza alla mobilità all'interno del mercato del lavoro. A tale scopo si è adottata una particolare tecnica di indagine, il panel, che consiste nell'intervistare in tempi successivi lo stesso campione di individui. In tal modo, raccogliendo informazioni sulle variazioni intervenute nei singoli individui in momenti diversi, è possibile stimare i cambiamenti di stato o di flusso (ad esempio i passaggi dalla condizione di disoccupato a quella di occupato, e viceversa) evitando, allo stesso tempo, gli errori dovuti alla memoria che si verificano nelle indagini longitudinali che utilizzano quesiti retrospettivi.

Nella presente indagine, i laureati nell'anno 1986 che avevano risposto alla prima indagine del 1989, sono stati nuovamente intervistati (in una "seconda ondata") due anni dopo, nel 1991. Nella seconda rilevazione è stato usato un diverso questionario, con il quale sono state raccolte informazioni sui cambiamenti intervenuti, nel corso dei due anni, rispetto alla situazione lavorativa che era stata rilevata nel 1989.

Entrambe le indagini, sono state effettuate tramite questionario autocompilato, inviato per posta.

Con la pubblicazione dei risultati dell'indagine longitudinale l'Istat completa il quadro informativo relativo al processo di transizione dall'università al mondo del lavoro dei laureati dell'anno solare 1986, iniziato con la pubblicazione dei risultati dell'indagine del 1989.

Le principali caratteristiche metodologiche del disegno di campionamento della rilevazione sono state descritte nel volume "Indagine 1989 sugli sbocchi professionali dei laureati" [2] al quale si rimanda per maggiori dettagli. Si ritiene qui opportuno ricordare che il campo di osservazione della ricerca è costituito dagli oltre 72.000 studenti che hanno conseguito un diploma di laurea nel corso dell'anno solare 1986 in tutte le sedi universitarie italiane e comprende tutti i corsi di laurea esistenti in tale anno. Non sono stati inclusi nella rilevazione, invece, gli

---

[1] "Indagine 1989 sugli sbocchi professionali dei laureati" ISTAT, Collana di informazione 1990, n. 17.
[2] "Indagine 1989 sugli sbocchi professionali dei laureati" ISTAT, Collana di informazione 1990, n. 17.

studenti che hanno conseguito un diploma non equiparabile alla laurea, come ad esempio il diploma in statistica, educazione fisica, vigilanza scuole elementari, ecc. Le liste dei laureati, fornite dalle segreterie delle varie università, erano organizzate per corso di laurea e sede universitaria. Tali caratteristiche sono state utilizzate come variabili di stratificazione; il campione, pertanto, risulta ad uno stadio stratificato. Le informazioni relative al sesso dei laureati sono state utilizzate per il procedimento di post-stratificazione, in quanto disponibili soltanto a livello aggregato.

Il campione utilizzato per effettuare l'indagine longitudinale comprende 9712 unità, ovvero tutti i laureati che avevano risposto all'indagine precedente (corrispondenti al 72,2% del campione programmato nel 1989). Di questi, hanno risposto alla seconda intervista 7090 laureati (il 73%).

Per il trattamento delle mancate risposte totali si è scelto di utilizzare una metodologia, ancora sperimentale, di ponderazione dei dati elementari [3] che utilizza, oltre ai pesi diretti derivati dal disegno di campionamento, anche dei fattori correttivi di tali pesi, ottenuti vincolando le stime della seconda "ondata" a quelle di alcune variabili stimate nella prima rilevazione del 1989 e ritenute particolarmente significative rispetto agli obiettivi conoscitivi dell'indagine. Le variabili considerate sono state quelle che nel 1989 descrivevano la condizione occupazionale e la ripartizione geografica di residenza del laureato.

E' stato così possibile riponderare i coefficienti diretti di riporto all'universo degli individui rispondenti alla seconda "ondata" in modo tale da ricostruire, tranne minime differenze dovute agli arrotondamenti, la medesima struttura della popolazione stimata, rispetto alle variabili prescelte, nella precedente indagine del 1989.

Tuttavia, a causa della non uniforme distribuzione delle mancate risposte totali, che ha determinato elevati tassi di non rispondenti in alcuni strati, è stato necessario riaggregare i dati per grandi ripartizioni geografiche (Nord-Centro e Mezzogiorno) e presentare una classificazione dei corsi meno analitica, in alcuni casi, di quella utilizzata per l'indagine del 1989.

---

[3] Tale metodologia è descritta nella nota interna ISTAT "Un metodo di stima generalizzato per le indagini sulle imprese e sulle famiglie" (di P.D. Falorsi e S. Falorsi)

Metodologia utilizzata per il controllo del rischio di violazione

Nella diffusione di files di collezioni campionarie di dati in forma elementare, privati di codici identificativi diretti, è presente un rischio di violazione determinato dalla possibilità di associare, con l'ausilio di un file esterno contenente i codici identificativi diretti, alcuni records presenti nel file.

Vediamo allora più in dettaglio come può avvenire e quali fattori possono rendere possibile l'identificazione di rispondenti.

Nell'archivio sono presenti informazioni (sotto forma di variabili qualitative o quantitative) di diverso tipo:

1. Variabili pubbliche. Sono quelle contenute in registri accessibili al pubblico (luogo di nascita, residenza, sesso, età, stato civile, etc.). Fra queste è importante distinguere le variabili territoriali (comune, provincia, regione di residenza, etc.)
2. Variabili riservate o sensibili. Sono quelle variabili non contenute in registri accessibili al pubblico e relative ad aspetti che per motivi di natura diversa devono ritenersi confidenziali (comportamento sessuale, condizioni di salute, reddito, etc.).
3. Altre variabili. Sono quelle variabili non contenute in registri accessibili al pubblico che non riguardano aspetti confidenziali.

Le variabili pubbliche presenti nell'archivio per le quali è possibile ipotizzare la presenza in archivi esterni che contengono anche codici identificativi diretti sono chiamate **variabili chiave** perché potenzialmente presenti in archivi esterni ed utilizzabili come chiave per la identificazione di unità.

Il **primo importante fattore di rischio** è la percentuale di casi unici nella popolazione rispetto alle variabili chiave. 'Caso unico' è detta una unità che da sola presenta determinate caratteristiche rispetto alle variabili chiave. Se ad esempio si considerano quattro variabili chiave sesso, età, stato civile e comune di residenza e si verifica, in un determinato comune, l'esistenza di una sola vedova di 20 anni, questa costituisce un caso unico rispetto alle quattro variabili chiave considerate. L'incidenza della percentuale di casi unici nella popolazione sul rischio è del tutto evidente se si considera che le unità che sono casi unici sia nell'archivio che in quello esterno possono essere poste in collegamento. Le risposte presenti nell'archivio possono essere in tal modo attribuite all'unità corrispondente dell'archivio esterno nel quale sono contenuti gli identificatori diretti.

Per calcolare la percentuale di casi unici della popolazione occorre distinguere due situazioni: i) si opera su indagini totali (censimenti o altre indagini che riguardano tutta la popolazione oggetto di indagine), ii) si opera su indagini campionarie. E' importante evidenziare che ciò che comunque interessa è la percentuale di casi unici della popolazione e non quella del campione (infatti una parte consistente dei casi unici del campione deriva da casi che nella popolazione, e quindi nell'archivio esterno sono doppi, tripli, etc. e per i quali non è immediata l'identificazione). Nel caso di indagini totali la percentuale di casi unici potrà essere frutto di un calcolo dai dati disponibili, nel caso di indagini campionarie

occorrerà utilizzare procedure di stima che dai dati del campione consentono di risalire alla popolazione.

Per questo motivo sono stati introdotti modelli probabilistici fra i quali il più noto è un modello proposto da Bethlehem (Bethlehem et al., 1990) che si basa sulla combinazione di due distribuzioni teoriche molto note: la distribuzione Poisson e la distribuzione Gamma. Tuttavia le capacità previsive del metodo, se si escludono situazioni particolari, risultano insoddisfacenti (Biggeri, Zannella, 1991). Nel seguito mostreremo un modello di previsione proposto dallo scrivente (Crescenzi 1992(a), 1992(b)), che consente stime molto precise della percentuale di casi unici anche partendo da dati campionari.

Il **secondo fattore** che influenza il rischio di violazione è il tasso di campionamento della collezione campionaria (consideriamo per semplicità solo schemi di tipo autoponderante). Se l'indagine è campionaria la collezione da rilasciare può essere composta da tutte le unità rilevate e quindi il tasso coincide con il tasso di campionamento dell'indagine, oppure da un sub-campione. Nel caso di un'indagine totale si dispone di tutte le unità della popolazione e la collezione campionaria sarà costituita da un campione da questa estratto. In ogni caso occorre tener conto degli aspetti connessi all'efficienza del disegno campionario e dell'attendibilità delle stime.

L'influenza del tasso di campionamento sul rischio è evidente se si considera che, data la percentuale di casi unici della popolazione, sarà resa disponibile una frazione di questi tanto più piccola, quanto più è basso il tasso di campionamento.

Il **terzo fattore** non dipende dalle caratteristiche della collezione campionaria che viene rilasciata, ma dalla quantità e dalla concentrazione delle informazioni esterne. A grandi linee si possono individuare tre modi in cui si può caratterizzare la conoscenza esterna. Un primo modo è la conoscenza di tipo individuale; ogni individuo conosce e dispone di informazioni su amici, parenti, vicini di casa, etc. In genere sono conoscenze molto limitate e frammentarie e perciò poco incidenti sul rischio. Vi sono poi conoscenze di tipo diffuso ma parziale, spesso su supporto informatico, che riguardano qualche decina di migliaia di individui (archivi di ditte sui dipendenti, affiliati di associazioni, dati rilevati da società di rilevazione private, etc.). Infine vi sono conoscenze di tipo globale che riguardano un gran numero, se non la totalità di individui (anagrafi, archivi di ministeri o enti pubblici, etc.). Ovviamente si possono presentare situazioni intermedie fra queste. Tuttavia, quanto più le informazioni sono consistenti e concentrate, tanto più saranno rilevanti gli effetti sul rischio, per questo motivo è essenziale delinearne quanto più possibile le caratteristiche.

Per vari motivi è possibile che le variabili chiave siano codificate in modo diverso nell'archivio da diffondere ed in quello esterno (utilizzazione di definizioni e classificazioni differenti, presenza di un intervallo temporale fra la situazione registrata in un file e quella registrata nell'altro, presenza di errori di codifica nei due archivi). La probabilità che le variabili chiave siano codificate in modo identico nei due file costituisce perciò il **quarto fattore** che influenza il rischio di violazione, tanto più sarà bassa tale probabilità, tanto più si ridurrà il rischio di violazione.

Occorre infine tener presente che altri elementi influenzano il rischio di violazione: possono essere infatti prese varie precauzioni dal punto di vista normativo-contrattuale nel rapporto fra Istituto ed utente come la propensione dell'utente ad effettuare un tentativo di identificazione e la stabilità nel tempo della codifica.

Se introduciamo la seguente simbologia:

f_U = frequenza relativa dei casi unici nella popolazione;

f_R = tasso di campionamento della collezione campionaria che viene rilasciata;

f_A = frequenza relativa delle unità presenti nell'archivio esterno di cui può disporre l'utilizzatore;

f_T = propensione dell'utente ad effettuare un tentativo di identificazione;

f_I = probabilità che le variabili chiave siano codificate identicamente nel file rilasciato e nell'archivio esterno;

f_S = probabilità che la codifica sia stabile nel tempo;

N = numero di unità della popolazione.

Il numero medio atteso di identificazioni può essere scritto come:

$$\mu = N f_U f_R f_A f_I f_T f_S$$

Per la probabilità di esatta codifica e registrazione delle variabili chiave nell'archivio esterno (archivio dell' Università di Palermo), si è ipotizzata la completa correttezza dei dati, ponendo quindi la suddetta probabilità pari ad 1.

Per quanto riguarda la probabilità di esatta codifica e registrazione delle variabili chiave nell'archivio rilasciato, si sono utilizzate le informazioni presenti nel Prospetto 10 a pag. 27 del fascicolo dell'indagine sugli sbocchi professionali dei laureati del 1989, e precisamente il numero di errori riscontrati dopo la registrazione su supporto informatico riferibili a ciascuna sezione del questionario.

Le variabili chiave (tranne "Provincia della sede universitaria" e "Corso di laurea") sono tutte nelle sezioni 4 e 6.

Nel Prospetto 10 la sezione 6 è stata aggregata alla sezione 5. Le due sezioni comprendono rispettivamente 8 e 10 variabili. Si è calcolato il numero medio di errori per variabile nelle sezioni 5 e 6 e si è ottenuto un risultato pari a 38 (0,39% sul totale dei records).

La sezione 4 comprende 29 variabili. Il numero medio di errori per le variabili di questo gruppo è risultato pari a 24 (0,24% sul totale dei records).

Nell'ipotesi che le frequenze degli errori riscontrati siano una stima sia pur approssimata della probabilità degli errori relativi a queste variabili, il complemento ad 1 può rappresentare una stima delle probabilità di esatta codifica e registrazione.

I valori ottenuti sono:

VARIABILE	QUESITO

Cittadinanza	Sez. 6	0,9961
Voto di laurea	Q. 20 - Sez. 4	0,9976
Anno di nascita	Sez. 6	0,9961
Diploma di scuola superiore	Q. 27 - Sez. 4	0,9976
Sesso	Sez. 6	0,9961
Prov. res. al mom. dell'iscr. all'Univ.	Sez. 6	0,9961

Per le due variabili "Provincia della sede universitaria" e "Corso di laurea" si sono supposte le modifiche percentuali pari a 0,1% con conseguente valore delle probabilità di esatta codifica e registrazione pari a 0,999.

Il prodotto fra le varie probabilità (supponendo queste indipendenti fra loro) ci da il valore della probabilità di esatta codifica e registrazione delle variabili chiave dell'archivio rilasciato:

Probabilità di esatta cod. e reg. = $0,999 \times 0,999 \times 0,9961 \times 0,9976 \times 0,9961 \times 0,9976 \times 0,9961 \times 0,9961 = 0,97$.

Avvertenze

1. I coefficienti di espansione si ottengono dividendo per 10.000 i valori che compaiono nelle colonne 207-215 (per l'indagine svolta nel 1989) e 256-264 (per l'indagine svolta nel 1991) del file. Il primo insieme di coefficienti è stato calcolato senza tener conto della longitudinalità dell'indagine, questi devono essere quindi utilizzati se si vuole analizzare variabili relative alla sola indagine svolta nel 1989, in tutti gli altri casi risulta necessario utilizzare il secondo insieme di coefficienti.

2. Nel tracciato record le colonne 14-192 sono relative al questionario del 1989, le colonne 216-245 sono relative al questionario del 1991, le rimanenti colonne riportano informazioni comuni o sono poste RISERVATO ISTAT.

3. Il file è costituito da 9123 record (9712 dell'indagine del 1989). I questionari pervenuti sono stati 9760 di cui 48 sono risultati inutilizzabili. Inoltre 11 record (gli ultimi del file) contraddistinti dal numero d'ordine "0000" (col. 9-13), sono stati aggiunti per tener conto delle mancate risposte.

4. Nel caso di caduta campionaria nella seconda indagine (1991), per le variabili relative a questa parte del record vi sono tutti *blank*. I *blank* corrispondenti a singoli campi, derivano da mancate risposte (tra il 1989 e il 1991).