CONTRIBUTI ISTAT

n. 6/2010

Tutela statistica della riservatezza: una proposta metodologica per la valutazione del rischio orientata all'indagine sulle forze di lavoro

F. Foschi e L. Franconi

Istat

SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

ISTAT – Servizio Editoria - Centro stampa

CONTRIBUTI ISTAT

n. 6/2010

Tutela statistica della riservatezza: una proposta metodologica per la valutazione del rischio orientata all'indagine sulle forze di lavoro

F. Foschi(*) e L. Franconi(*)

(*) ISTAT – Direzione Centrale per le tecnologie ed il supporto metodologico

Contributi e Documenti Istat

Istituto Nazionale di Statistica
Servizio Editoria – Centro stampa
Via Tuscolana, 1788 - 00173

Tutela statistica della riservatezza: una proposta metodologica per la valutazione del rischio orientata all'indagine sulle forze di lavoro

Flavio Foschi, Istat Direzione Centrale per le tecnologie ed il supporto metodologico
Luisa Franconi, Istat Direzione Centrale per le tecnologie ed il supporto metodologico

Sommario: La necessità di tutela della riservatezza discende da vincoli di legge e da obblighi assunti verso i rispondenti. L'intrusione si concretizza quando tramite i dati rilasciati sia possibile acquisire informazioni non pubbliche circa unità statistiche presenti nel collettivo di riferimento. Seguendo l'approccio di considerare a rischio quelle che uniche nel campione siano tali anche nella popolazione, vengono affrontati due possibili ambiti di intrusione; il primo attiene al contributo di singole variabili ad 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. In ordine a quest'ultimo viene illustrata una strategia di valutazione non vincolata all'espressione del legame funzionale tra la probabilità di aver osservato nel campione unità statistiche uniche anche nella popolazione e modalità manifestate dai caratteri rilevati. La valutazione del rischio viene implementata per l'indagine sulle forze di lavoro, considerando i dati trasversali dei sedici trimestri riferiti agli anni 2005-2008, nonché il primo trimestre 2009 limitatamente alle retribuzioni rilevate per i lavoratori dipendenti.

Parole chiave: intrusione, coefficiente di espansione, processo di Dirichlet

1. Premessa

Ai fini della tutela statistica della riservatezza saranno trattati alcuni aspetti inerenti la valutazione del rischio di intrusione in informazioni non pubbliche (non necessariamente costituite da dati sensibili) contenute nell'indagine sulle forze di lavoro. Poiché la descrizione dei dati ivi raccolti eccede gli scopi del presente lavoro, per essa si rinvia alla pubblicazione di riferimento (AA.VV., 2006). Obiettivo delle analisi svolte è la definizione del grado di dettaglio secondo il quale i microdati di indagine possano essere rilasciati per finalità di ricerca. 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., 2009), l'intrusione in informazioni non pubbliche si concretizza quando tramite i dati sia possibile acquisire notizie in merito a specifiche unità statistiche e possono essere di due tipi: identificazione del rispondente o associazione ad una persona (o organismo) di dati noti dalle indagini. 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 gli individui e, condizionatamente alla stratificazione di interesse, sono considerate a rischio (secondo una proposta consueta in letteratura, ad es. Skinner e Shlomo, 2008) quelle che, uniche nel campione, sono tali anche nella popolazione. Nella seconda sezione vengono brevemente discussi gli scenari delle analisi condotte sui dati trasversali dell'indagine, mentre la terza concerne alcuni aspetti di metodo e la quarta una sintesi dei principali risultati[1].

2. Ambiti di Riferimento per la Valutazione del Rischio di Intrusione

Sono stati oggetto di valutazione i sedici trimestri del periodo 2005-2008, nonché il primo trimestre 2009 limitatamente alle retribuzioni rilevate per i lavoratori dipendenti. Le variabili specificate nel questionario di indagine hanno suggerito 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.

2.1 Intrusione accidentale

L'eventuale identificazione origina "spontaneamente" dall'osservazione di un ristretto numero di variabili, alcune delle quali aventi un grado di risoluzione sufficiente ad isolare poche unità statistiche in base alla sola distribuzione marginale di frequenze. Una valutazione di fattispecie potenzialmente rischiose, ossia strati contenenti una sola unità rispondente, ha riguardato posizione professionale e settore di attività, nonché lo stato estero di nascita; inoltre, limitatamente ai dati disponibili al primo trimestre 2009, è stata esaminata la variabile retribuzioni. In riferimento alle prime, per ottenere una plausibile approssimazione degli ordini di grandezza coinvolti è stato conteggiato il numero di modalità che in un trimestre siano state rappresentate da una sola unità statistica, espungendo quelle presenti in tre trimestri consecutivi (per i motivi esposti nel paragrafo 3.1). In dettaglio:

- le retribuzioni in unità di euro, figura 1, rilevate soltanto per i lavoratori dipendenti, seguono una distribuzione con valori sensibilmente diversificati soltanto nelle code: il primo ed il terzo quartile ammontano approssimativamente a 940 e 1440 euro, le intensità ordinate in successione non decrescente che ricadono entro l'1% delle frequenze relative cumulate oscillano tra i 20 e i 260 euro, quelle oltre il 99% tra i 3060 e i 12000. A fini di protezione, mentre la sparsità delle intensità estremali ne impone la codifica in classi, i dati appartenenti alla porzione di distribuzione maggiormente densa di osservazioni potrebbero essere semplicemente arrotondati ai 10 euro senza comprometterne la significatività a fini di ricerca;

[1] Luisa Franconi (DPTS/DCMT/MSS/C) ha curato la sezione 2, Flavio Foschi (DPTS/DCMT/MSS/C) le rimanenti. Si ringrazia Antonio Rinaldo Discenza (DPTS/DCCV/FOL/B) per il supporto informativo in ordine all'indagine sulle forze di lavoro e la disponibilità dei dati esposti nella tavola 2.

- la posizione professionale, figura 2, se rilasciata secondo la codifica a 4 digit, potrebbe esporre una ristretta categoria di intervistati a rischi di identificazione. Si tratta di quanti appartengono al novero dei legislatori e dirigenti. Ad esempio, di tredici casi sospetti rilevati per il quarto trimestre 2005, uno appartiene alla categoria degli "ambasciatori, ministri plenipotenziari ed altri dirigenti della carriera diplomatica". Riscontri con stato estero di nascita, età e provincia di residenza, ridurrebbero notevolmente l'alea sull'identificazione dei rispondenti. La limitazione al terzo digit per le posizioni professionali indicate sarebbe sufficiente a mitigare il problema;

Figura 1: *Istogramma delle retribuzioni, primo trimestre 2009*

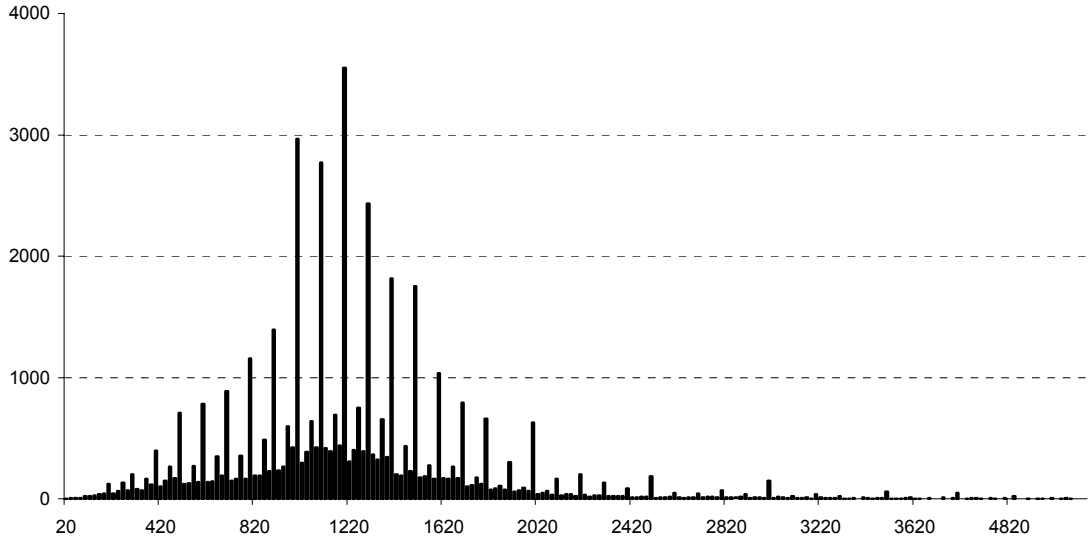
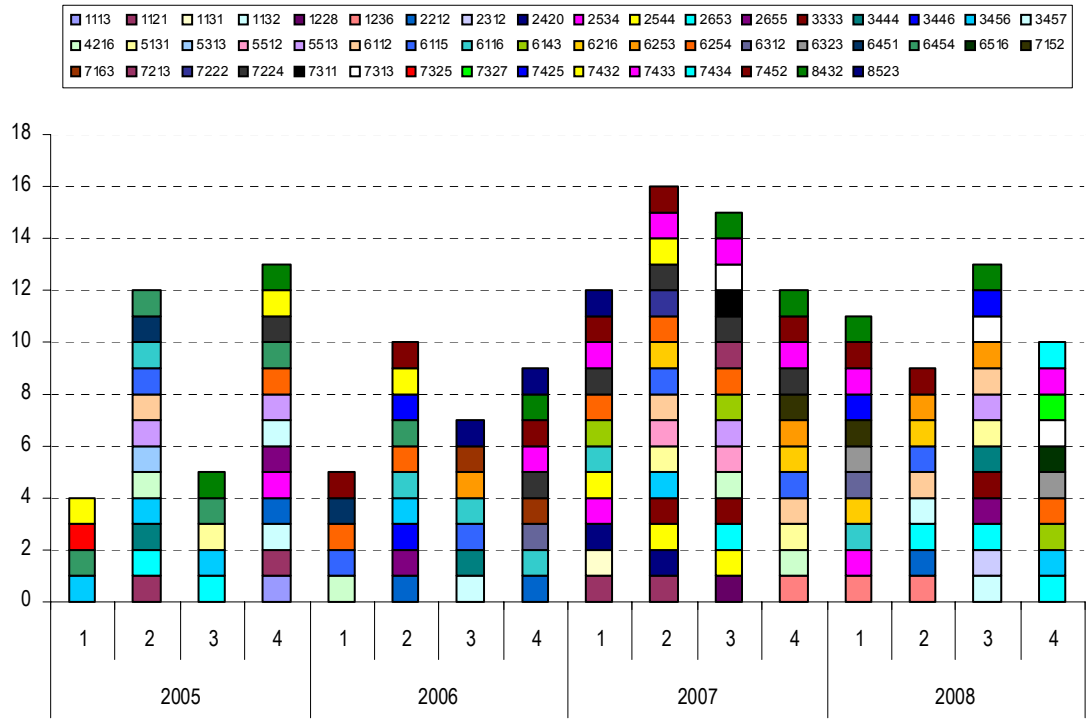


Figura 2: *Strati con una sola unità campionata per professione a 4 digit*



- gli stati esteri di nascita e cittadinanza, marginalmente considerati, individuano 24 casi sospetti nei trimestri terzo 2005 e quarto 2007 e, rispettivamente, 29 casi nel terzo trimestre 2005; in presenza di un dettaglio territoriale provinciale, l'identificazione sarebbe verosimilmente non immediata, fatte salve le sovrapposizioni con attività professionali "rilevanti" discusse nel punto precedente;

- la NACE a 4 digit viene assegnata dal servizio che cura l'indagine secondo le risposte fornite dall'intervistato e, almeno in linea di principio, l'informazione potrebbe non coincidere con quella indicata dalle aziende nei questionari di indagine e nelle documentazioni amministrative. Il picco di situazioni potenzialmente a rischio è stato osservato nel quarto trimestre del 2007 con quindici casi, in buona parte falsi positivi data la natura delle attività coinvolte, tipicamente riguardanti un'ampia pluralità di addetti. Sembra ragionevole ritenere che il contenuto informativo della mutabile sia meno influente di quello veicolato dalla posizione professionale.

2.2 Intrusione non accidentale

Il rischio di violazione della riservatezza viene stimato ipotizzando che informazioni esterne siano associate all'indagine usando quali chiavi di abbinamento le variabili e i livelli di dettaglio appresso specificati.

- Codice comune ovvero codice provincia (con 103 modalità fino al 2007 e 107 dal 2008),
- sesso;
- età; 65 classi (annuali, fatti salvi i raggruppamenti fino a 2 anni, da 3 a 5 anni, da 6 a 10 anni, da 11 a 14 anni, 75 anni e oltre);
- stato civile; 4 modalità (non coniugati, coniugati, separati e divorziati, vedovi);
- numero componenti il nucleo familiare;
- stato occupazione; 4 modalità (non occupato, dipendente, collaboratore, autonomo);
- cittadinanza, come ricodificata dal servizio che cura l'indagine; 3 modalità (cittadinanza italiana, straniera UE, straniera non UE).

Lo scenario così delineato in parte riflette l'informazione resa fruibile da Eurostat per le omologhe indagini di altri Paesi e rappresenta un'assunzione circa il bagaglio di conoscenze dell'intrusore. In particolare, l'accorpamento in unica classe di separati, legali o di fatto, e divorziati coincide con l'articolazione adottata da Eurostat; i raggruppamenti in classi per le età da 0 a 14 anni sono sovrapponibili con le durate dei corsi di istruzione primaria e secondaria inferiore dell'ordinamento italiano e, mentre non sottraggono informazioni sulle forze lavoro, possono offrire un contributo allo studio dei fenomeni legati alla scolarità.

3. La stima del rischio di intrusione non accidentale

L'approccio proposto da Skinner e Shlomo fonda la quantificazione del rischio associato a ciascun record sulla stima di un modello loglineare il più possibile parsimonioso, compatibilmente con l'ottenimento di un parametro di dispersione coerente con una verosimiglianza Poisson. L'idea di fondo risiede nella possibilità di discernere, sulla base delle informazioni disponibili, gli strati che con maggiore probabilità abbiano nella popolazione, così come nel campione, una sola unità statistica. Sfortunatamente, nel caso dell'indagine sulle forze di lavoro il livello di dettaglio delle modalità non permette di esplorare modelli che abbiano più di un certo numero di interazioni del secondo ordine, insufficienti ad evitare seri problemi di sovradispersione.

Benché Skinner e Shlomo li considerino meno gravi rispetto a quelli di sottodispersione (l'underfitting porterebbe mediamente a valutazioni pessimistiche del rischio scongiurando – ove ciò avvenga – il rilascio di dati che permettano una facile identificazione del rispondente), sembra opportuno individuare delle alternative che, condizionatamente alle variabili di scenario, non incorporino gli effetti della mispecificazione. Tale premessa suggerisce di concentrare l'attenzione sul modello saturo. Secondo questa prospettiva, l'alea in ordine agli strati rischiosi si restringe al coefficiente di espansione. Quest'ultimo è ottenuto correggendo le probabilità di inclusione delle unità statistiche (famiglie) per le mancate risposte ed effettuandone la calibrazione rispetto a domini non annidati tra i quali sono rilevanti:

- regione, sesso, 14 classi di età,
- provincia, sesso, 5 classi di età,
- grandi comuni, sesso, 5 classi di età,
- stranieri residenti per regione, sesso, 5 classi di età.

Dunque, un ragionevole punto di partenza può essere ravvisato nel tentativo di rappresentare l'incertezza connaturata alle stime dei coefficienti. È importante sottolineare come l'approssimativa identità, nei domini di calibrazione, tra totali noti della popolazione e totali campionari moltiplicati per i pesi implichi la tendenziale assenza di sovrastima o sottostima sistematica delle frequenze concernenti gli strati non pianificati ivi annidati.

3.1 Metodologia proposta

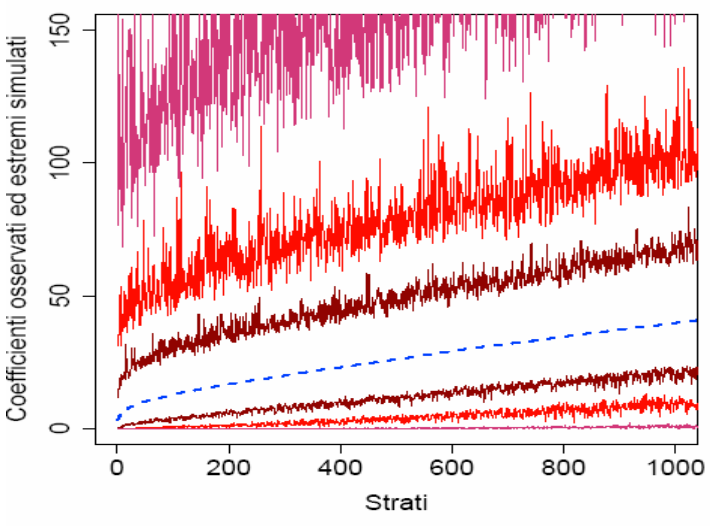
Definendo $\hat{Y}_s$ il numero di unità dello strato s^{mo} (s=1,…,N) ottenuto come prodotto tra frequenza campionaria y_s e coefficiente di espansione π_s, la soluzione proposta si fonda sulla simulazione delle $p_s \equiv p(\pi_s)$ secondo un processo di Dirichlet del quale vengano opportunamente inflazionate le varianze marginali. Una distribuzione di Dirichlet ammette delle marginali per le quali valgono le relazioni:

$$E(p_s) = \frac{\pi_s}{\sum \pi_j} \tag{1}$$

$$\sigma^2(p_s) = \frac{\pi_s \sum_{j \neq s} \pi_j}{\left(\pi_s + \sum_{j \neq s} \pi_j\right)^2 \left(\pi_s + \sum_{j \neq s} \pi_j + 1\right)}$$

Il prodotto delle π_s (s=1, …, N) per una costante k positiva e inferiore a 1 (alla quale può essere attribuito il significato di precisione) lascia immutati i valori attesi ma amplifica le varianze. Ogni "traiettoria" del processo è costituita dalle p_s che moltiplicate per $\Sigma_j \pi_j$ esprimono altrettante realizzazioni dei coefficienti. I valori attesi coincidono per costruzione con i π_s ma a causa della variabilità attorno alla media del processo è possibile dare conto dell'incertezza.

Figura 3: *Coefficienti come realizzazioni di processo*



La figura 3 mostra i valori assunti dai π_s nella successione ordinata dei primi mille strati (linea blu tratteggiata, coincidente con la funzione valor medio del processo), gli estremi inferiori e superiori delle realizzazioni $\tilde{\pi}_s \equiv p_s \sum \pi_j$ riferiti a precisioni k pari a 0.8, 0.2 e 0.05 (traiettorie tanto meno prossime alla funzione valor medio quanto minore è il valore di k). Definendo rischiosi gli strati rappresentati nel campione da una sola unità statistica e che abbiano realizzato, fissata la precisione k, in mille replicazioni, frequenze simulate $\hat{Y}_s$ inferiori a due almeno nell'uno per cento dei casi, è disponibile una misura di rischio riferita a ciascun singolo record. È opportuno notare che:

- essa ha immediata interpretazione dal punto di vista probabilistico,
- applicando la procedura indicata da Skinner e Shlomo, alla quantificazione degli score "individuali" dovrebbe seguire l'individuazione di un punto di cesura tale da separare ciò che si deve ritenere a rischio da quanto presumibilmente non lo è.

La precisione k è un parametro di disturbo eliminabile mediante marginalizzazione. Ammettendone per semplicità la discretizzazione, l'adozione di una legge uniforme si traduce nella scelta, relativamente conservativa, di mediare i risultati ottenuti a fronte dei diversi valori di precisione. Tale condotta può

essere temperata dall'uso di un filtro deterministico consistente nel giudicare non rischiosi gli strati che pur rilevati nel campione con una sola unità statistica siano presenti in tre trimestri consecutivi; poiché "l'indagine sulle forze di lavoro segue uno schema di rotazione trimestrale in cui le famiglie vengono intervistate per due trimestri consecutivi, escluse per due trimestri e successivamente re-intervistate per altri due trimestri", uno strato presente in tre trimestri consecutivi è verosimilmente ascrivibile alla presenza, nella popolazione, di almeno due unità statistiche che ne condividano le modalità. Il rischio globale r_g può essere ricavato come frazione dei record che siano risultati a rischio; in simboli, posti i l'indice dei record, r_i il rischio associato all'i^{mo} record, $\delta(\ )$ il delta di Dirac, n il numero di record inclusi nel dataset,

$$r_g = \frac{1}{n}\sum_i \delta(r_i > 0)$$

4 Sintesi dei risultati

Il modo di procedere descritto ha portato all'individuazione di strati classificati rischiosi almeno nell'1% delle replicazioni, dei quali può essere interessante esporre le distribuzioni marginali per provincia riferite globalmente ai sedici trimestri, nonché la media dei coefficienti di espansione risultanti dall'indagine. I dati della tavola 1 sono ordinati secondo la graduatoria decrescente dei casi.

Tavola 1: *Casi stimati a rischio per provincia*

Prov.	Casi	Media dei coefficienti	Prov.	Casi	Media dei coefficienti
102	3027	11.25	095	43	14.17
007	3002	10.41	104	33	12.56
022	2062	12.34	106	31	14.36
070	1741	11.76	053	25	15.05
057	1450	11.67	043	23	14.97
094	980	12.43	075	22	14.15
051	842	12.58	078	21	14.28
021	786	13.12	082	21	15.81
076	669	12.45	099	19	15.10
096	620	13.42	032	18	13.31
077	559	12.75	045	18	12.83
020	525	13.43	003	17	15.13
097	508	13.04	019	17	15.18
086	400	12.89	093	16	15.14
034	371	13.54	029	13	14.50
103	214	13.51	014	12	15.08
052	203	13.61	107	12	12.06
105	203	12.38	010	9	13.87
005	201	13.98	033	9	14.61
087	160	14.18	035	9	7.24
025	154	14.63	100	9	14.32
054	154	13.20	009	8	15.98
002	123	14.01	023	8	14.64
091	100	14.51	085	7	16.73
101	88	14.06	083	6	16.29
079	79	14.04	004	4	16.63
062	74	14.86	044	4	15.86
031	69	14.42	066	2	16.52
067	65	13.34	068	2	16.62
081	44	15.00	090	2	16.75

Si può notare come essa ricalchi sostanzialmente la graduatoria dei coefficienti ex-post ordinati in successione non decrescente, calcolati per provincia come media sui trimestri T del rapporto tra stima del totale e ampiezza del campione:

$$\hat{\bar{\pi}}_s = \frac{1}{T}\sum_t \frac{\hat{Y}_{s,t}}{y_{s,t}}$$

Fissato lo strato di riferimento (nella fattispecie la provincia), il fatto che essi risultino di modesta entità dipende dal fatto che in piccoli domini le unità campionate y_s rappresentano una quota relativamente grande della popolazione complessiva stimata.

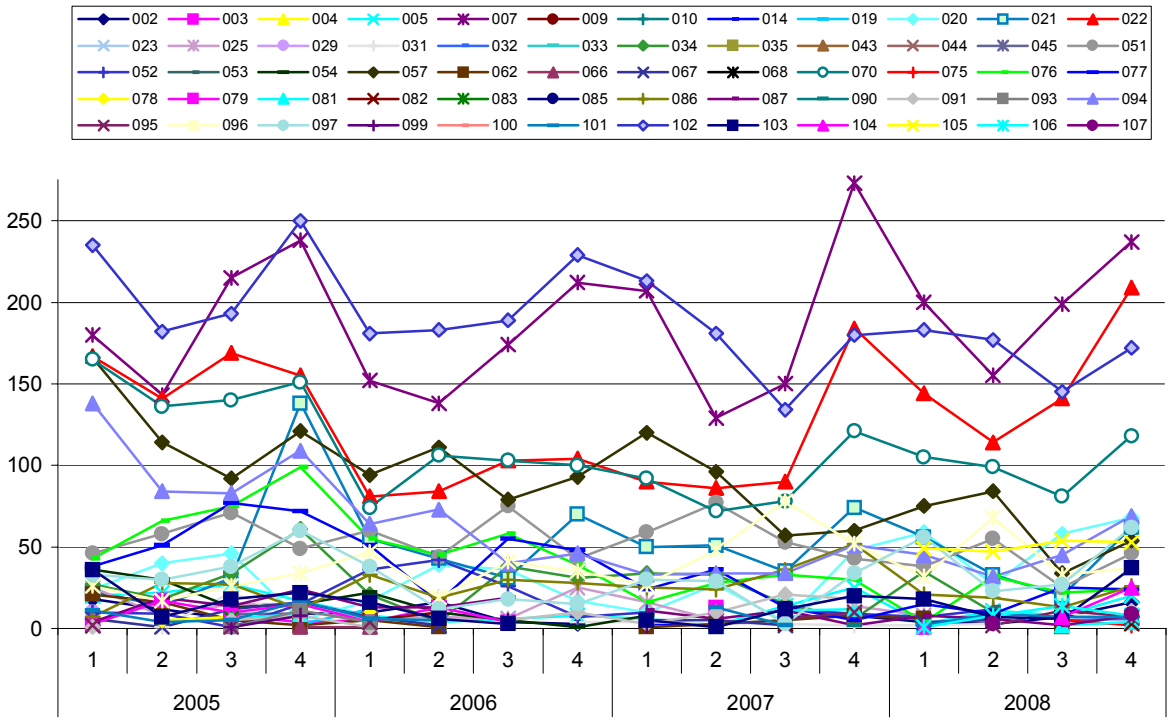
Tavola 2: *Pesi medi stimati*

Provincia	Peso	Provincia	Peso	Provincia	Peso	Provincia	Peso
001 - TO	527	028 - PD	777	055 - TR	383	082 - PA	323
002 - VC	143	029 – R0	377	056 - VT	649	083 - ME	309
003 – NO	234	030 - UD	466	057 - RI	80	084 - AG	683
004 - CN	461	031 - GO	162	058 - RM	854	085 - CL	409
005 - AT	131	032 - TS	357	059 - LT	487	086 - EN	121
006 - AL	497	033 - PC	245	060 - FR	355	087 - CT	233
007 - AO	44	034 - PR	144	061 - CE	355	088 - RG	918
008 - IM	411	035 - RE	364	062 - BN	213	089 - SR	408
009 - SV	371	036 - MO	599	063 - NA	506	090 - SS	465
010 - GE	320	037 - BO	908	064 - AV	442	091 - NU	210
011 - SP	657	038 - FE	696	065 - SA	540	092 - CA	413
012 - VA	784	039 - RA	439	066 - AQ	259	093 - PN	237
013 - CO	403	040 - FC	452	067 - TE	229	094 - IS	74
014 – S0	171	041 - PU	454	068 - PE	522	095 - OR	218
015 - MI	734	042 - AN	400	069 - CH	416	096 - BI	101
016 - BG	617	043 - MC	284	070 - CB	72	097 - LC	124
017 - BS	546	044 - AP	452	071 - FG	338	098 - LO	582
018 - PV	342	045 - MS	429	072 - BAi	472	099 - RN	286
019 - CR	215	046 - LU	896	073 - TA	595	100 - PO	492
020 - MN	141	047 - PT	418	074 - BR	351	101 - KR	166
021 - BZ	130	048 - FI	783	075 - LE	253	102 - VV	51
022 - TN	74	049 - LI	606	076 - PZ	114	103 - VB	125
023 - VR	467	050 - PI	418	077 - MT	105	104 - OT	220
024 - VI	521	051 - AR	136	078 - CS	311	105 - OG	87
025 - BL	157	052 - SI	185	079 - CZ	188	106 - VS	154
026 - TV	815	053 - GR	249	080 - RC	299	107 - CI	213
027 - VE	590	054 - PG	243	081 - TP	264		

I dati della tavola 2, forniti da DPTS/DCCV/FOL/B, permettono di apprezzare la circostanza. La relazione tra le graduatorie dei domini rischiosi e dei coefficienti ex-post medi è fisiologica a causa della (1) e appare conforme alle attese in quanto il sovracampionamento dei domini meno estesi dovrebbe aumentare la probabilità di osservare strati rappresentati, nella popolazione, da poche unità statistiche. Essa può essere interpretata come condizione necessaria (ancorché non sufficiente) per la coerenza del metodo applicato. L'andamento nel tempo dei casi stimati a rischio per ciascuna provincia è sintetizzato nella figura 4. Concentrando l'attenzione sulla parte del diagramma corrispondente alle ordinate più elevate, per Aosta, Vibo Valentia, Trento e Campobasso in diversi anni sono visibili picchi in corrispondenza del quarto trimestre. Le tendenze debolmente crescenti di Trento e Aosta a partire dal 2006 (spezzate di colore viola e rosso) sembrano ascrivibili alla circostanza che la diminuzione nel tempo del tasso di risposta (e con essa l'aumento dei pesi di espansione) comune alla maggioranza delle province è meno ampia per le due in discorso; in termini relativi, esse sono caratterizzate da una copertura crescente e corrispondenti pesi decrescenti. In riferimento alle rimanenti province che hanno

manifestato situazioni potenzialmente rischiose, la variabilità è più contenuta. Con le cautele legate alla breve estensione temporale dell'analisi, sembra presente inerzialità negli ordini di grandezza di ciascuna provincia. Tale circostanza rispecchia il legame della rischiosità con la frazione di unità campionate nei domini più sottili e lascia presumere che l'esercizio di valutazione del rischio possa essere limitato a periodicità più ampie rispetto al singolo trimestre.

Figura 4: *Casi a rischio per provincia, anno e trimestre*



Nella tavola 3 sono esposte le quantificazioni del rischio globale relativo a ciascun dataset dell'indagine trasversale sulle forze di lavoro, con articolazione territoriale limitata alla provincia:

Tavola 3: *Rischio globale associato ai dataset per anno e trimestre*

Anno	Trimestre	Record a rischio	Totale record	Rischio globale
2005	1	1529	182011	0.0084
2005	2	1297	177723	0.0073
2005	3	1528	170267	0.0090
2005	4	1791	174371	0.0103
2006	1	1224	171821	0.0071
2006	2	1081	173363	0.0062
2006	3	1153	169393	0.0068
2006	4	1168	169726	0.0069
2007	1	1104	174362	0.0063
2007	2	1016	171496	0.0059
2007	3	936	165699	0.0056
2007	4	1350	166189	0.0081
2008	1	1190	174883	0.0068
2008	2	1073	169775	0.0063
2008	3	1030	162729	0.0063
2008	4	1443	164552	0.0088

In relazione ai sedici trimestri esaminati, distinguendo i grandi comuni (individuati in base ai criteri fissati per il censimento 2001) dai restanti, si può notare come molti rappresentino una quota ragguardevole dei record per provincia:

Tavola 4: *Peso dei grandi Comuni nelle rispettive province*

Comune	(a): record provincia	(b): record comune	(b) / (a)
Torino	67067	23900	0.36
Genova	43368	29966	0.69
Milano	84003	26792	0.32
Verona	29484	8882	0.30
Venezia	22317	7119	0.32
Bologna	16527	5992	0.36
Firenze	19477	6529	0.34
Roma	72275	45717	0.63
Napoli	96726	29166	0.30
Bari	53571	10696	0.20
Palermo	61094	33280	0.54
Messina	33388	13008	0.39
Catania	74166	23675	0.32

Il rischio globale associato alle articolazioni territoriali complemento dei grandi comuni all'interno di ciascuna provincia è sintetizzato nella tavola 5:

Tavola 5: *Rischio globale per anno e trimestre: complemento dei grandi Comuni*

Anno	Trimestre	Record a rischio	Totale record	Rischio globale
2005	1	6076	165690	0.0367
2005	2	5206	162167	0.0321
2005	3	4943	155562	0.0318
2005	4	6173	159000	0.0388
2006	1	4649	156701	0.0297
2006	2	4196	158240	0.0265
2006	3	4224	155005	0.0273
2006	4	4923	155217	0.0317
2007	1	4650	159526	0.0291
2007	2	3822	156585	0.0244
2007	3	3104	151369	0.0205
2007	4	4141	151009	0.0274
2008	1	4096	159037	0.0258
2008	2	3796	154402	0.0246
2008	3	3144	148321	0.0212
2008	4	3777	149482	0.0253

5. Considerazioni finali

La valutazione del rischio di intrusione per i dati trasversali dell'indagine sulle forze di lavoro è stata effettuata nell'intento di definire il grado di dettaglio secondo il quale essi possano essere rilasciati a fini di ricerca. Sono stati considerati due possibili ambiti di riferimento. Il primo concerne il contributo di singole variabili alle identificazioni spontanee o accidentali, possibili osservando poche distribuzioni

marginali di frequenze; il secondo riguarda l'intenzionale tentativo di abbinamento tra dati di indagine e una lista di microdati disponibile all'intrusore. Adottando una proposta consueta nella letteratura sull'argomento, la valutazione del rischio mira ad individuare le unità statistiche che, uniche nel campione, siano tali anche nella popolazione. L'elevato livello di articolazione dei caratteri rilevati rende problematico l'uso di modelli loglineari volti ad esplicitare il legame tra rischio e modalità manifestate dalle unità rispondenti. La sovradispersione conseguente all'underfitting dovrebbe implicare valutazioni del rischio pessimistiche, ma sembra interessante capire se sia possibile evitare che le decisioni orientate dalla valutazione del rischio dipendano da modelli mispecificati. L'alternativa proposta consiste nell'apprezzamento del grado di incertezza implicito nelle stime dei coefficienti di espansione ed è giustificabile alla luce della tendenziale assenza di sovrastima o sottostima sistematica delle frequenze concernenti gli strati non pianificati contenuti all'interno dei domini di calibrazione. Simulazioni basate sui processi di Dirichlet hanno permesso una valutazione numerica del rischio associato a ciascun record. Le graduatorie dei domini rischiosi e dei coefficienti ex-post medi (calcolati per ciascuna provincia come media sui trimestri del rapporto tra stima del totale e ampiezza del campione) appaiono sovrapponibili, coerentemente con la circostanza che il sovracampionamento dei domini meno estesi aumenta la probabilità di osservare strati rappresentati, nella popolazione, da poche unità statistiche. I risultati esposti sono condizionati al set informativo specificato nel paragrafo 2.2. Dunque, l'effettivo numero di record a rischio è sottostimato o sovrastimato se le informazioni di cui il terzo dispone sono maggiori o minori di quelle ritenute plausibili ed usate per la valutazione. Se si riflette sulla circostanza che, a fronte di ogni possibile grado di dettaglio per le modalità dei singoli caratteri, sarebbero – almeno in linea di principio – concepibili scenari in numero pari alla cardinalità dell'insieme di potenza costruito sulle variabili d'indagine, si può comprendere come le analisi in discorso abbiano valenza orientativa per situazioni reali prossime a quelle ipotizzate.

Bibliografia

AA.VV. *EU Labour Force Survey database User Guide*. European Commission, Eurostat, Directorate F, Unit F-2, 2008. http://circa.europa.eu/irc/dsis/employment/info/data/eu_lfs/lfs_main/_LFSuserguide/EULFS_Database_UserGuide_2008.pdf. 04/09/2009.

AA.VV. *Handbook on Statistical Disclosure Control.* ESSnet on Statistical Disclosure Control, 2009. http://neon.vb.cbs.nl/casc/handbook.htm. 04/09/2009.

AA.VV. *La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione*. Roma: Istat, 2006 (Metodi e Norme n. 32).

Skinner C. e Natalie Shlomo. "Assessing Identification Risk in Survey Microdata Using Log-Linear Models". *Journal of the American Statistical Association* 103, 483 (2008): 989-1001.

Contributi ISTAT(*)

1/2006 – Ennio Fortunato – *L'analisi integrata delle esigenze informative dell'utenza Istat: Il contributo del Sistema di Indicatori Territoriali*

2/2006 – Francesco Altarocca – *I design pattern nella progettazione di software per il supporto alla statistica ufficiale*

3/2006 – Roberta Palmieri – *Le migranti straniere: una lettura di genere dei dati dell'osservatorio interistituzionale sull'immigrazione in provincia di Macerata*

4/2006 – Raffaella Amato, Silvia Bruzzone, Valentina Delmonte e Lidia Fagiolo – *Le statistiche sociali dell'ISTAT e il fenomeno degli incidenti stradali: un'esperienza di record linkage*

5/2006 – Alessandro La Rocca – *Fuzzy clustering: la logica, i metodi*

6/2006 – Raffaella Cascioli – *Integrazione dei dati micro dalla Rilevazione delle Forze di Lavoro e dagli archivi amministrativi INPS: risultati di una sperimentazione sui dati campione di 4 province*

7/2006 – Gianluca Brogi, Salvatore Cusimano, Giuseppina del Vicario, Giuseppe Garofalo e Orietta Patacchia – *La realizzazione di Asia Agricoltura tramite l'utilizzo di dati amministrativi: il contenuto delle fonti e i risultati del processo di integrazione*

8/2006 – Simonetta Cozzi – *La distribuzione commerciale in Italia: caratteristiche strutturali e tendenze evolutive*

9/2006 – Giovanni Seri – *A graphical framework to evaluate risk assessment and information loss at individual level*

10/2006 – Diego Bellisai, Annalisa Lucarelli, Maria Anna Pennucci e Fabio Rapiti – *Feasibility studies for the coverage of public institutions in sections N and O*

11/2006 – Diego Bellisai, Annalisa Lucarelli, Maria Anna Pennucci e Fabio Rapiti – *Quarterly labour cost index in public education*

12/2006 – Silvia Montagna, Patrizia Collesi, Florinda Damiani, Danila Fulgenzio, Maria Francesca Loporcaro e Giorgia Simeoni – *Nuove esperienze di rilevazione della Customer Satisfaction*

13/2006 – Lucia Coppola e Giovanni Seri – *Confidentiality aspects of household panel surveys: the case study of Italian sample from EU-SILC*

14/2006 – Lidia Brondi – *L'utilizzazione delle surveys per la stima del valore monetario del danno ambientale: il metodo della valutazione contingente*

15/2006 – Carlo Boselli – *Le piccole imprese leggere esportatrici e non esportatrici: differenze di struttura e di comportamento*

16/2006 – Carlo De Greogorio – *Il nuovo impianto della rilevazione centralizzata del prezzo dei medicinali con obbligo di prescrizione*

1/2007 – Paolo Roberti, Maria Grazia Calza, Filippo Oropallo e Stefania Rossetti – *Knowledge Databases to Support Policy Impact Analysis: the EuroKy-PIA Project*

2/2007 – Ciro Baldi, Diego Bellisai, Stefania Fivizzani, e Marina Sorrentino – *Production of job vacancy statistics: coverage*

3/2007 – Carlo Lucarelli e Giampiero Ricci – *Working times and working schedules: the framework emerging from the new Italian lfs in a gender perspective*

4/2007 – Monica Scannapieco, Diego Zardetto e Giulio Barcaroli – *La Calibrazione dei Dati con R: una Sperimentazione sull'Indagine Forze di Lavoro ed un Confronto con GENESEES/SAS*

5/2007 – Giulio Barcaroli e Tiziana Pellicciotti – *Strumenti per la documentazione e diffusione dei microdati d'indagine: il Microdata Management Toolkit*

6/2007 – AA.VV. – *Seminario sulla qualità: l'esperienza dei referenti del sistema informativo SIDI - 1ª giornata*

7/2007 – Raffaella Cianchetta, Carlo De Gregorio, Giovanni Seri e Giulio Barcaroli – *Rilevazione sulle Pubblicazioni Scientifiche Istat*

8/2007 – Emilia Arcaleni, e Barbara Baldazzi – *Vivere non insieme: approcci conoscitivi al Living Apart Together*

9/2007 – Corrado Peperoni e Francesca Tuzi – *Trattamenti monetari non pensionistici metodologia sperimentale per la stima degli assegni al nucleo familiare*

10/2007 – AA.VV – *Seminario sulla qualità: l'esperienza dei referenti del sistema informativo SIDI - 2ª giornata*

11/2007 – Leonello Tronti – *Il prototipo (numero 0) dell'Annuario di statistiche del Mercato del Lavoro (AML)*

12/2007 – Daniele Frongia, Raffaello Martinelli, Fernanda Panizon, Bruno Querini e Andrea Stanco – *Il nuovo Sistema informatico Altri Servizi. Progetto di reingegnerizzazione dei processi produttivi delle indagini trimestrali di fatturato degli altri servizi*

1/2008 – Carlo De Gregorio, Stefania Fatello, Rosanna Lo Conte, Stefano Mosca, Francesca Rossetti – *Sampling design and treatment of products in Istat centralised CPI surveys*

2/2008 – Mario Albisinni, Elisa Marzilli e Federica Pintaldi – *Test cognitivo e utilizzo del questionario tradotto: sperimentazioni dell'indagine sulle forze di lavoro*

3/2008 – Franco Mostacci – *Gli aggiustamenti di qualità negli indici dei prezzi al consumo in Italia: metodi, casi di studio e indicatori impliciti*

4/2008 – Carlo Vaccari e Daniele Frongia – *Introduzione al Web 2.0 per la Statistica*

5/2008 – Antonio Cortese – *La conta degli stranieri: una bella sfida per il censimento demografico del 2011*

6/2008 – Carlo De Gregorio, Carmina Munzi e Paola Zavagnini – *Problemi di stima, effetti stagionali e politiche di prezzo in alcuni servizi di alloggio complementari: alcune evidenze dalle rilevazioni centralizzate dei prezzi al consumo*

7/2008 – AA.VV. – *Seminario: metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche strutturali*

8/2008 – Monica Montella – *La nuova matrice dei margini di trasporto*

9/2008 – Antonia Boggia, Marco Fortini, Matteo Mazziotta, Alessandro Pallara, Antonio Pavone, Federico Polidoro, Rosabel Ricci, Anna Maria Sgamba e Angela Seeber – *L'indagine conoscitiva della rete di rilevazione dei prezzi al consumo*

10/2008 – Marco Ballin e Giulio Barcaroli – *Optimal stratification of sampling frames in a multivariate and multidomain sample design*

11/2008 – Grazia Di Bella e Stefania Macchia – *Experimenting Data Capturing Techniques for Water Statistics*

(*) ultimi cinque anni

12/2008 – Piero Demetrio Falorsi e Paolo Righi – *A Balanced Sampling Approach for Multi-way Stratification Designs for Small Area Estimation*

13/2008 – AA.VV. – *Seminario: Strategie e metodi per il controllo e la correzione dei dati nelle indagini sulle imprese: alcune esperienze nel settore delle statistiche congiunturali*

14/2008 – Francesco Chini, Marco Fortini, Tiziana Tuoto, Sara Farchi, Paolo Giorgi Rossi, Raffaella Amato e Piero Borgia – *Probabilistic Record Linkage for the Integrated Surveillance of Road Traffic Injuries when Personal Identifiers are Lacking*

15/2008 – Sonia Vittozzi – *L'attività editoriale e le sue regole: una ricognizione e qualche proposta per l'Istat editore*

16/2008 – Giulio Barcaroli, Stefania Bergamasco, Michelle Jouvenal, Guido Pieraccini e Leonardo Tininini – *Generalised software for statistical cooperation*

1/2009 – Gianpiero Bianchi, Antonia Manzari, Alessandra Reale e Stefano Salvi – *Valutazione dell'idoneità del software DIESIS all'individuazione dei valori errati in variabili quantitative*

2/2009 – Silvia Pacini – *Indicatori territoriali su retribuzioni e costo del lavoro una sperimentazione basata sui dati Inps*

3/2009 – Mauro Tibaldi – *L'occupazione femminile nella Pubblica amministrazione: un'analisi dei dati della Ragioneria Generale dello Stato*

4/2009 – Veronica Rondinelli – *La calibrazione dei pesi campionari delle aziende Rica nell'indagine sui risultati Economici delle Aziende Agricole*

5/2009 – Domenico Tebala – *Distribuzione territoriale del rischio di usura in Calabria: una cluster analysis comunale*

6/2009 – Carolina Corea, Incoronata Donnarumma e Antonio Frenda – *La stima dello stock di beni durevoli delle famiglie: un primo contributo sperimentale*

7/2009 – Massimo Costanzo, Rosalba Filippello e Marco Marini – *La contabilità nazionale verso l'ATECO 2007: alcune considerazioni sull'uso di matrici di conversione nel periodo di transizione*

8/2009 – Anna Ciammola, Francesca Ceccato, Maria Carla Congia, Silvia Pacini, Fabio Massimo Rapiti e Donatella Tuzi – *The Italian Labour Cost* Index (LCI): sources and methods

1/2010 – Antonio Cortese, Gerardo Gallo e Evelina Paluzzi – *Il censimento della popolazione straniera: opinioni a confronto sul principale aspetto definitorio*

2/2010 – Ciro Baldi e Marina Sorrentino – *I posti vacanti in Italia e in Europa. Le nuove statistiche trimestrali armonizzate: prime analisi delle serie storiche*

3/2010 – Fabio Bacchini, Anna Ciammola, Roberto Iannaccone e Marco Marini – *Combining forecasts for a flash estimate of Euro area GDP*

4/2010 – Alessandra Burgio, Alessandra Battisti, Alessandro Solipaca, Simona Colosimo, Lorella Sicuro, Gianfranco Damiani, Giordana Baldassarre Giulia Milan, Tiziana Tamburrano, R. Crialesi e Walter Ricciardi – *La relazione tra offerta di servizi di Long Term Care ed i bisogni assistenziali dell'anziano*

5/2010 – Flavio Foschi e Brunero Liseo – *Artificial Continuous Data for SDC*

6/2010 – Flavio Foschi e Luisa Franconi – *Tutela statistica della riservatezza: una proposta metodologica per la valutazione del rischio orientata all'indagine sulle forze di lavoro*