*A cura di*: Luisa Franconi e Giovanni Seri

*Capitoli redatti da*: Luisa Franconi (paragrafi 1.1, 1.2, 1.4.3, 5.4, 9.1)
Giovanni Seri (paragrafi 1.3, 1.4 eccetto il sotto paragrafo 1.4.3, 1.5, capitolo 5 eccetto i paragrafi 5.4 e 5.5, capitolo7, paragrafo 8.4, capitolo 9 eccetto il paragrafo 9.1)
Giovanni Maria Merola (capitoli 2, 3 e 4)
Alessandra Capobianchi (paragrafo 5.5, capitolo 6)
Silvia Polettini (capitolo 8 eccetto il paragrafo 8.4)

*Hanno collaborato*: Daniela De Luca
Maurizio Lucarelli

Per chiarimenti sul contenuto della pubblicazione rivolgersi a:
Istat, Direzione Centrale per le Metodologie Statistiche
e le Tecnologie Informatiche
Via C. Balbo, 16
00184 Roma
Tel 06 4673 2306

# Indice

**Presentazione** 9

**PARTE PRIMA La tutela della riservatezza nel rilascio dell'informazione statistica**

**Capitolo 1. La tutela della riservatezza: concetti di base** 13

1.1. Introduzione 13

1.2. Contenuto del volume 13

1.3. La tutela della riservatezza dal punto di vista legislativo 14
1.3.1. Riferimenti normativi 15
1.3.2. Definizioni 16

1.4. La tutela della riservatezza dal punto di vista statistico 19
1.4.1. La violazione della riservatezza nel rilascio di tabelle 23
1.4.2. La violazione della riservatezza nel rilascio di dati elementari 24
1.4.3. Interrogazioni on line di banche dati su internet 26

1.5. Tipologie di rilascio e tipologie di utenti 27

**PARTE SECONDA La tutela statistica della riservatezza per dati aggregati**

**Capitolo 2. Rischio di violazione per dati aggregati** 33

2.1. Introduzione 33

2.2. Tabelle di frequenza 34

2.3. Tabelle di intensità 35

2.4. Misure di rischio di violazione per tabelle di intensità 36

2.5. Esempi di misure di rischio lineari subadditive 37

2.6. Tabelle campionarie 42

2.7. Tabelle personalizzate 43

2.8. Tabelle collegate 43

2.9. Tabelle di intensità con valori negativi 46

**Capitolo 3. La protezione statistica di tabelle** 47

3.1. Tecniche di protezione non perturbative: soppressione 48
3.1.1. Soppressione secondaria marginale 48
3.1.2. Soppressione secondaria 49
3.1.3. Soppressione parziale 53
3.1.4. Aspetti computazionali della soppressione 54
3.1.4.1. Soppressione secondaria ottimale 55
3.1.4.2. Soppressione parziale 57

3.2. Tecniche di protezione non perturbative: riarrangiamento delle categorie delle variabili classificatrici 58
3.3. Tecniche di protezione perturbative: arrotondamento 59
3.3.1. Arrotondamento deterministico 60
3.3.2. Arrotondamento stocastico non controllato 61
3.3.3. Arrotondamento stocastico controllato 62
**Capitolo 4. Tutela statistica della riservatezza per dati rilasciati da siti Web** 63
4.1. Introduzione 63
4.2. Approcci alla protezione dei Siti Web per la Diffusione di Dati (Swdd) 65
4.2.1. Limitazioni di ingresso e di uscita 65
4.2.2. Metodi statistici 66
4.3. Protezione dei Swdd che rilasciano tabelle 66
4.3.1. Metodi perturbativi: riarrangiamento delle modalità 67
4.3.2. Metodi perturbativi casuali 68
4.3.3. Metodi soppressivi 69
4.3.3.1. Tabelle collegate gerarchiche 70
4.3.3.2. Metodi per il calcolo degli intervalli di esistenza delle celle di una tabella dato un insieme di sue marginali 72
4.4. Sistemi automatici per la protezione soppressiva di Swdd 77
4.4.1. Protezione soppressiva dei Swdd statici: selezione di un insieme di tabelle di frequenza rilasciabili 78
4.4.2. Protezione soppressiva dei Swdd dinamici: valutazione *al volo* del rischio di rilasciare tabelle aggiuntive 79
4.5. Laboratori virtuali 80
4.5.1. Laboratori virtuali server di programmi degli utenti 81
4.6. Strategie per l'applicazione della tutela statistica della riservatezza ai Swdd 82
**PARTE TERZA La tutela statistica della riservatezza per dati individuali**
**Capitolo 5. Il rilascio dei dati individuali** 89
5.1. Introduzione 89
5.2. I dati individuali 89
5.3. Violazione della riservatezza nel caso del rilascio di dati individuali 92
5.4. Tipi di rilascio per utente 96
5.5. Fattori che influenzano il rischio di re-identificazione 98
**Capitolo 6. Rischio di violazione di dati individuali in ambito sociale** 101
6.1. Introduzione 101
6.2. Funzioni di Rischio globale 101
6.2.1. Stima del numero dei casi unici 102

6.2.2. Modelli di rischio 104
6.3. Rischio Individuale 105
6.3.1. Rischio Individuale probabilistico 106
6.3.2. Metodologia Cbs 112
6.3.2.1. Stimatori delle frequenze nella popolazione nella metodologia Cbs 114
6.3.2.2. Stimatore indiretto 114
**Capitolo 7. Rischio di violazione di dati elementari di impresa** 117
7.1. Introduzione 117
7.2. Fattori di rischio nei dati di impresa 117
**Capitolo 8. Le tecniche di protezione per i dati individuali** 121
8.1. Metodi di protezione: una visione d'insieme. 121
8.2. Modelli di protezione 122
8.2.1. Metodi non parametrici per la tutela della riservatezza 123
8.2.2. Metodi semi parametrici per la tutela della riservatezza 124
8.2.3. Metodi parametrici per la tutela della riservatezza 125
8.3. Una rassegna dei metodi di protezione per microdati 126
8.3.1. Ricodifica globale, *topcoding* e arrotondamento 126
8.3.2. Soppressione locale 127
8.3.3. *Data Swapping* 127
8.3.4. Aggiunta di disturbo 128
8.3.5. Imputazione multipla 132
8.3.6. Proposta di Fienberg, Makov e Steele 133
8.3.7. Post-Randomizzazione (Pram) 134
8.3.8. Metodi di imputazione da modelli di tipo regressivo 135
8.3.9. Metodo di Dandekar basato sul campionamento da ipercubo latino (*latin hypercube sampling*) 137
8.3.10. Modello mistura di Grim, Boček e Pudil 138
8.3.11. Mascheramento Matriciale 139
8.4. La microaggregazione 140
**PARTE QUARTA Alcune esperienze in Istat**
**Capitolo 9. Comunicazione di dati a soggetti non Sistan** 149
9.1. Il rilascio dei *file standard* in Istat 149
9.2. La microaggregazione dei dati microaggregati del sistema dei conti economici delle imprese italiane. Anni 1995 e 1996 151
9.3. Il Laboratorio Adele per l'analisi dei dati elementari 155
**APPENDICE A.1 Decreto legislativo 6 settembre 1989, n. 322** 159

**APPENDICE A.2 Codice di deontologia e di buona condotta per i trattamenti di dati personali a scopi statistici e di ricerca scientifica effettuati nell'ambito del sistema statistico nazionale** 169

**Riferimenti bibliografici** 177

# Presentazione

Il problema della tutela della riservatezza nel rilascio di informazione statistica da parte dell'Istituto nazionale di statistica (Istat) e del Sistema statistico nazionale (Sistan) è un problema complesso che interseca materie anche molto diverse tra loro e che richiede di conseguenza l'interazione fra competenze specifiche dei vari settori coinvolti.

Innanzitutto, vi sono problemi giuridici legati alle norme di garanzia dei diritti dei rispondenti alle indagini e di assolvimento dei doveri dell'Istituto nazionale di statistica. Il quadro di riferimento normativo deve successivamente essere interpretato da parte dello statistico cui spetterà il compito di definire modelli e di stimare quantità che possano misurare il grado di protezione dei dati o di rischio di violazione della riservatezza.

Vicini ai problemi statistici vi sono quelli computazionali di creazione di algoritmi efficienti che portino al raggiungimento di soluzioni ottimali o migliori. Infine, con l'avvento di nuove forme di diffusione principalmente legate alle possibilità di comunicazione telematica, sono sorti nuovi problemi di tipo informatico conseguenti alla creazione di reti protette ove permettere l'accesso controllato ad utenti esterni al Sistan o, ancora, alla gestione di database di grandi dimensioni con relativi sistemi di interrogazione *sicura*.

La recente evoluzione della legislazione sul trattamento dei dati personali (dalla Legge 675/1996 al D.lgs 30 giugno 2003, n. 196, Codice in materia di protezione dei dati personali[1] che ha assorbito e risistemato tutte le previgenti norme) ha creato una legittima maggiore attenzione da parte dei cittadini alla tutela del proprio diritto individuale alla privacy. Sotto questa spinta anche la statistica ufficiale ha dovuto adeguare le proprie procedure, definizioni e criteri soprattutto sotto l'aspetto qualitativo piuttosto che quantitativo. Infatti, il diritto alla tutela della riservatezza dei cittadini e delle imprese chiamate a collaborare per il raggiungimento dei fini della statistica ufficiale era già riconosciuto e garantito come "segreto statistico" nel decreto costitutivo del Sistan (D.lgs. 322/1989).

Per questo motivo il nuovo quadro di riferimento normativo ha fornito uno stimolo per rinnovare l'interesse da parte degli addetti ai lavori che si sono trovati a compiere uno sforzo notevole per adeguare e adeguarsi alle nuove procedure, ma anche a trarre nuove indicazioni per migliorare e aumentare il livello di informazione statistica rilasciata. In particolare, con la scrittura del *Codice di deontologia e di buona condotta per i trattamenti di dati personali a scopi statistici e di ricerca scientifica effettuati nell'ambito del sistema statistico nazionale* (in breve Codice deontologico) si è definito un quadro di riferimento normativo specifico che aiuterà i produttori di statistica ufficiale a interpretare le nuove esigenze e i nuovi strumenti a disposizione particolarmente nella fase di rilascio dell'informazione prodotta. Anche dal punto di vista della metodologia statistica la rinnovata attenzione nei confronti della materia ha prodotto importanti sviluppi convogliando l'interesse di numerosi studiosi. Negli ultimi

[1] Disponibile sul sito http://www.garanteprivacy.it

anni, infatti, è cresciuto notevolmente il numero di contributi metodologici nell'ambito della tutela della riservatezza che hanno portato alla definizione di nuovi e più accurati modelli per l'individuazione e il trattamento del rischio di violazione in dati predisposti per il rilascio.

Il presente volume, oltre a presentare una descrizione delle metodologie più attuali per la tutela della riservatezza dei rispondenti alle indagini nella fase di rilascio dell'informazione statistica, intende fornire agli enti e uffici del Sistan gli strumenti necessari per una introduzione al quadro normativo sull'argomento e indicazioni con esempi pratici di tecniche e procedure attualmente utilizzate in Istat. Il quadro metodologico che viene descritto comprende anche le recenti innovazioni di processo, come le nuove tecniche di produzione dei *file standard*, e di prodotto, come i file di dati microaggregati e il Laboratorio per l'Analisi di dati elementari (Adele) che gli utenti esterni al Sistan possono utilizzare per motivi di ricerca.

In questa attività ci si è avvalsi della cooperazione di molte strutture dell'Istat tra cui il Servizio Condizioni economiche delle famiglie (Cef) che ha fornito i dati dell'indagine sui Consumi delle famiglie, il Servizio Statistiche strutturali sulle imprese dell'industria e dei servizi (Ssi) che ha fornito i dati delle indagini sul Sistema dei conti di impresa (Sci) e della *Community Innovation Survey* nonché l'Ufficio di gabinetto e in particolare la dott.ssa Maria Rosaria Simeone per la continua opera di supporto nell'interpretazione degli aspetti legislativi. I ringraziamenti miei e degli autori vanno, inoltre, al dott. Giulio Barcaroli e alla dott.ssa Vittoria Buratta per la loro opera di revisione del testo nei suoi aspetti tecnico scientifici e al dott. Fabio Albo e alla dott.ssa Giovanna Bellitti per la revisione del testo negli aspetti giuridici e normativi.

Questo lavoro testimonia lo sforzo che l'Istat ha prodotto e continuerà a produrre affinché nel rispetto della normativa sulla tutela della riservatezza degli individui sia migliorata la fruibilità e la qualità dell'informazione statistica da parte del mondo della ricerca scientifica. Per tale motivo l'Istituto sostiene lo studio di nuove metodologie e strumenti anche nell'ottica più ampia della Comunità europea. L'Istat, infatti, è stato promotore e partecipa ad un progetto europeo per l'implementazione delle nuove metodologie per la tutela della riservatezza in un software. Obiettivo principale di questo progetto denominato Casc (*Computational Aspects of Statistical Confidentiality*)[2] è lo sviluppo del software Argus per la produzione di tabelle e di file di dati elementari "protetti" da possibili violazioni del segreto statistico. Tale software rappresenterà uno strumento indispensabile per l'applicazione delle metodologie standard di cui si è dotato l'Istat per la tutela della riservatezza nel rilascio di informazione statistica.

Francesco Zannella
Dipartimento per la produzione statistica e il coordinamento tecnico-scientifico

[2] Per maggiori informazioni si rimanda al sito http://neon.vb.cbs.nl/casc/default.htm

# PARTE PRIMA

## La tutela della riservatezza nel rilascio dell'informazione statistica

# Capitolo 1. La tutela della riservatezza: concetti di base[(*)]

## 1.1 Introduzione

La funzione primaria di un sistema statistico pubblico è quella di produrre statistica ufficiale per il proprio Paese. Infatti, il Decreto legislativo 6 settembre 1989, n.322, costitutivo del Sistema statistico nazionale (Sistan), cita: "L'informazione statistica ufficiale è fornita al Paese e agli organismi internazionali attraverso il Sistema statistico nazionale" (art.1 comma 2) e ancora "I dati elaborati nell'ambito delle rilevazioni statistiche comprese nel programma statistico nazionale sono patrimonio della collettività e vengono distribuiti per fini di studio e di ricerca a coloro che li richiedono secondo la disciplina del presente decreto, fermi restando i divieti di cui all'art.9" riguardanti il segreto statistico (art.10 comma 1).

La fase del processo produttivo di un'indagine statistica in cui si concretizza questa funzione è detta "diffusione". La diffusione è, difatti, la fase nella quale vengono poste in essere le modalità con cui l'obiettivo conoscitivo di un'indagine viene reso ai fruitori di informazione statistica ed è in questa fase che la tutela della riservatezza dei rispondenti alle indagini si presenta nei suoi aspetti tecnici e metodologici. Infatti, il compito di diffondere informazione statistica entra spesso in contrasto con il tradizionale obbligo morale, ma anche legale, degli uffici statistici di mantenere la privacy dei singoli rispondenti a garanzia di una più attiva cooperazione dei partecipanti alle indagini.

In questo primo capitolo vengono trattati i riferimenti normativi nel Paragrafo 1.3 mentre, nei paragrafi successivi, viene fornita una visione d'insieme sintetica del problema della tutela della riservatezza nel rilascio di informazione statistica. In particolare, nel Paragrafo 1.4 vengono esaminati i concetti fondamentali e le varie fasi del processo di tutela della riservatezza da un punto di vista statistico, dalla definizione del problema alla protezione del dato. Nel Paragrafo 1.5 si classificano le procedure statistiche e tecnico-legali adottate dall'Istat per tutelare la riservatezza dell'unità statistica a seconda delle differenti tipologie di dato rilasciato e dei possibili utenti di informazione statistica. Il prossimo Paragrafo 1.2, infine, contiene una visione d'insieme del volume con suggerimenti su eventuali percorsi di lettura differenti a seconda delle esigenze del lettore.

## 1.2 Contenuto del volume

Con il presente lavoro si intende fornire principalmente un quadro di riferimento sia normativo che metodologico ad uso di chi si occupa di diffusione, o più in generale di rilascio di informazione statistica, nell'ambito della statistica ufficiale, ma anche un manuale di indirizzo metodologico per chi intendesse approfondire gli aspetti teorico-statistici della materia.

---

(*) I paragrafi 1.1 e 1.2 sono stati redatti da Luisa Franconi. I paragrafi 1.3 e 1.5 sono stati redatti da Giovanni Seri. Il paragrafo 1.4 è stato redatto da Giovanni Seri eccetto il sotto-paragrafo 1.4.3 redatto da Luisa Franconi

L'intento è quello di assolvere a un duplice scopo in funzione degli utenti a cui è principalmente destinato. Il primo obiettivo che ci si è posti, quindi, è quello di fornire un punto di riferimento tecnico per gli operatori della diffusione, cioè, per chi, all'interno del Sistema statistico nazionale è coinvolto nella diffusione di informazione statistica ufficiale e, pertanto, può incontrare almeno potenzialmente problemi legati alla tutela della riservatezza dei dati. Il secondo scopo è di tipo più prettamente metodologico e, da questo punto di vista, il manuale è stato pensato per quegli studiosi che intendono occuparsi degli aspetti statistico metodologici della materia.

La struttura del manuale tiene in considerazione questa duplice funzione cercando di suggerire diversi percorsi di lettura a seconda delle esigenze del lettore.

La Prima Parte del manuale può essere considerata un'introduzione al problema della tutela della riservatezza dal punto di vista statistico inquadrata nello schema tecnico legislativo di riferimento per l'Istat e il Sistema statistico nazionale italiano. Tutti gli aspetti della tutela della riservatezza statistica vengono toccati, ma senza approfondire particolarmente gli aspetti metodologici, occupandosi di quello che prevalentemente viene messo in pratica nell'esperienza italiana.

La Parte Seconda e Terza si occupano rispettivamente dei dati aggregati (tabelle) e dei dati elementari curando in maniera più approfondita gli aspetti metodologici. Viene presentata, in questa parte del manuale, un'ampia panoramica dei metodi e delle tecniche presenti in letteratura cercando di tenere in debito conto gli avanzamenti della ricerca fatti negli ultimi anni che hanno modificato alcune metodologie e alcuni approcci al problema della tutela della riservatezza nella diffusione di dati nelle sue diverse forme.

L'ultima parte del manuale è dedicata, invece, alla descrizione di alcune esperienze di rilascio di informazione statistica da parte dell'Istat e, in particolare, al Laboratorio per l'analisi dei dati elementari (Adele), dove il controllo della tutela della riservatezza è legato prevalentemente a soluzioni di tipo tecnico/amministrativo oltre che statistico. Soluzioni di questo tipo sono state considerate con crescente interesse da parte degli Istituti nazionali di statistica in proporzione alla crescente richiesta di informazioni sempre più dettagliate proveniente dal mondo della ricerca scientifica. D'altro canto, soluzioni di tipo amministrativo sono le uniche adottabili nei casi in cui il rispetto del vincolo del segreto statistico non consentirebbe di rilasciare l'informazione richiesta nonostante l'applicazione dei metodi statistici di tutela della riservatezza.

## 1.3 La tutela della riservatezza dal punto di vista legislativo

Negli ultimi anni una crescente presa di coscienza da parte dei cittadini del diritto individuale alla tutela della riservatezza e la necessità di assolvere ad improcrastinabili impegni assunti in ambito comunitario, hanno spinto il legislatore italiano a delineare il quadro di principi e norme che regolano il trattamento di dati personali. L'oggetto specifico sottoposto a tutela è il diritto della persona (fisica o giuridica) a che i dati che la riguardano siano trattati in modo legittimo in un contesto globale che riconosce il ruolo dell'informazione in una società moderna come fattore di sviluppo della persona e della collettività. Più specificatamente, il diritto alla tutela della riservatezza non è

inteso come esclusività della conoscenza di ciò che attiene alla sfera privata ma è contemperato con il diritto di informare e essere informati secondo un principio di trasparenza, con la necessità di dotare le amministrazioni pubbliche delle informazioni necessarie a migliorare il governo degli interessi loro affidati e, di particolare interesse per quanto attiene a questo manuale, con la libertà di ricerca scientifica. Lo sforzo legislativo è stato, pertanto, indirizzato alla definizione di regole che consentano di individuare e rispettare il confine tra necessità e abuso nel trattamento di dati personali. Quantomeno, questo è il compito che viene assegnato allo statistico e al ricercatore in ogni fase della loro attività

Va osservato che, dal punto di vista della ricerca statistica ufficiale, il quadro normativo che si è delineato non ha alterato significativamente le possibilità di produrre informazione statistica. Il Sistan, infatti, ha sempre agito sin dalla sua costituzione in un quadro istituzionale di autonomia e rispetto delle regole largamente compatibile con le nuove norme soprattutto per quanto riguarda la diffusione dell'informazione statistica. Inoltre, occorre dire anche, che lo stesso Istituto nazionale di statistica (Istat) ha contribuito direttamente alla definizione del quadro normativo partecipando alla stesura del *Codice di deontologia e di buona condotta per i trattamenti di dati personali a scopi statistici e di ricerca scientifica effettuati nell'ambito del sistema statistico nazionale* richiesto dalla legge per il trattamento dei dati personali a fini statistici. Per commenti e interpretazioni in tale ambito si rimanda a Zucchetti *et al*. (2004).

### 1.3.1 Riferimenti normativi

Attualmente, la complessiva disciplina giuridica in tema di privacy è contenuta nel Decreto legislativo 30 giugno 2003, n. 196 (Codice in materia di protezione dei dati personali) che ha assorbito e risistemato tutte la normativa adottata al riguardo in Italia sin dalla Legge 31 dicembre 1996 n.675.

Con la Legge n.675/1996 ("tutela delle persone e di altri soggetti rispetto al trattamento dei dati personali") è stata originariamente regolata la materia della tutela della privacy con la finalità di garantire "che il trattamento dei dati personali si svolga nel rispetto dei diritti, delle libertà fondamentali, nonché della dignità delle persone fisiche, con particolare riferimento alla riservatezza e all'identità personale; e di garantire altresì i diritti delle persone giuridiche e di ogni altro ente o associazione" (art.1 comma 1). La stessa legge definisce la figura del Garante per la protezione dei dati personali (art.30). Contemporaneamente, il legislatore ha dato delega al Governo, con la Legge n.676/1996, per l'emanazione di disposizioni integrative in materia, in particolare per specificare le modalità di trattamento dei dati personali utilizzati a fini storici, di ricerca scientifica e di statistica, tenendo conto dei principi contenuti nella Raccomandazione n. R10 del 1983 (adottata e sostituita con la R18 del 1997 dal Consiglio d'Europa).

Fra i decreti emessi in attuazione della legge delega citiamo, come rilevanti per il presente contesto: il D.lgs 11 maggio 1999, n.135 e il D.lgs 30 luglio 1999, n.281. Il primo dispone sul trattamento di dati sensibili da parte di soggetti pubblici ed individua come "di rilevante interesse pubblico i trattamenti svolti dai soggetti pubblici che fanno parte del Sistema statistico nazionale …" (art.22). Il secondo dispone in materia di

trattamento dei dati personali per finalità storiche, statistiche e di ricerca scientifica e recepisce le raccomandazioni del Consiglio d'Europa n.R10 del 1983 e n.R18 del 1997. In particolare, questo decreto, demanda a specifici codici deontologici (da emanarsi a cura del Garante) la disciplina delle modalità di trattamento per scopi statistici fissandone principi e contenuti, introducendo per l'Italia il principio di "ragionevolezza" nella definizione dei mezzi che possono essere utilizzati per identificare un interessato.

Le norme sul Sistema statistico nazionale (Sistan) e sull'Istituto nazionale di statistica (Istat) sono contenute nel D.lgs 6 settembre 1989, n.322. In particolare, gli art.9 e 10 dispongono rispettivamente in materia di segreto statistico e di accesso ai dati statistici. Il Decreto legislativo n.281/99 ha modificato ed integrato il D.lgs n.322/1989 – in particolare, l'art. 9 riformulando proprio la nozione di segreto statistico in relazione al criterio della "identificabilità".

Tra i codici di deontologia di cui sopra, il *Codice di deontologia e di buona condotta per i trattamenti di dati personali a scopi statistici e di ricerca scientifica effettuati nell'ambito del sistema statistico nazionale* (da qui in avanti, per brevità, verrà indicato semplicemente come Codice deontologico), alla cui stesura ha contribuito direttamente l'Istat, è stato pubblicato – sotto forma di Provvedimento del Garante per la protezione dei dati personali - sulla Gazzetta Ufficiale del 1 ottobre 2002. Il Codice deontologico tratta anche i criteri e le modalità per l'interscambio dei dati individuali nell'ambito del Sistema statistico nazionale.

Di notevole rilevanza a livello europeo, è il Regolamento (Ce) 322/1997 del Consiglio, relativo alle statistiche comunitarie, attraverso il quale è stata definita la nozione di segreto statistico e sono state dettate le regole generali per la sua tutela nel processo di produzione delle statistiche comunitarie (articoli 13-19). In attuazione di tale regolamento è stato adottato il recente Regolamento (Ce) 831/2002 del Consiglio concernente l'accesso per finalità scientifiche ai dati riservati trasmessi all'Istituto centrale di statistica europeo (Eurostat) dagli Istituti nazionali di statistica. Più specificatamente, in quest'ultimo regolamento vengono delineate le norme per l'adozione da parte di Eurostat di due strumenti statistico-amministrativi che consentano l'utilizzo di dati statistici da parte della comunità scientifica. Questi due strumenti consistono nel rilascio di file di dati elementari opportunamente trattati (resi anonimi) e nella costituzione di un *Safe Data Center* presso Eurostat. L'Italia, come altri Paesi della Comunità, ha già adottato a livello nazionale questo tipo di pratiche che verranno trattate in maniera approfondita in questo manuale (si veda il Paragrafo 1.5.1).

Un ultimo riferimento normativo, è il Regolamento (Euratom/Cee) n.1588/1990 del Consiglio dell'11 giugno 1990 relativo alla trasmissione all'Istituto statistico delle Comunità europee di dati statistici protetti dal segreto.

### 1.3.2 Definizioni

Ai fini del presente manuale adottiamo le definizioni riportate di seguito.

- *Trattamento*: qualunque operazione o complesso di operazioni, effettuati anche senza l'ausilio di strumenti elettronici, concernenti la raccolta, la registrazione, l'organizzazione, la conservazione, la consultazione, l'elaborazione, la

modificazione, la selezione, l'estrazione, il raffronto, l'utilizzo, l'interconnessione, il blocco, la comunicazione, la diffusione, la cancellazione e la distruzione di dati anche se non registrati in una banca dati (definizione contenuta nel Decreto legislativo n.196/2003 );

- *Dato personale*: qualunque informazione relativa a persona fisica, persona giuridica, ente od associazione, identificati o identificabili, anche indirettamente, mediante riferimento a qualsiasi altra informazione, ivi compreso un numero di identificazione personale (definizione contenuta nel Decreto legislativo n.196/2003);
- *Utilizzazione a fini statistici*: utilizzazione per la sola elaborazione di tavole statistiche o di analisi statistico economiche; i dati non possono essere utilizzati a fini amministrativi, giudiziari, fiscali o di controllo nei confronti delle unità rilevate (definizione contenuta nel Regolamento n.1588/90 - Euratom/Cee);
- *Trattamento per scopi statistici*: qualsiasi trattamento di dati personali necessario per analisi statistiche o per la produzione di risultati statistici. In particolare, nel codice deontologico, quindi con riferimento al Sistan, è definito come: "qualsiasi trattamento effettuato per finalità di indagine statistica o di produzione, conservazione e diffusione di risultati statistici in attuazione del programma statistico nazionale o per effettuare informazione statistica in conformità agli ambiti istituzionali dei soggetti …" che fanno parte o partecipano al Sistema statistico nazionale;
- *Titolare*: la persona fisica, la persona giuridica, la pubblica amministrazione e qualsiasi altro ente, associazione od organismo cui competono, anche unitamente ad altro titolare, le decisioni in ordine alle finalità, alle modalità del trattamento di dati personali, e agli strumenti utilizzati, ivi compreso il profilo della sicurezza (definizione contenuta nel Decreto legislativo n.196/2003);
- *Responsabile*: la persona fisica, la persona giuridica, la pubblica amministrazione e qualsiasi altro ente, associazione od organismo preposti dal titolare al trattamento di dati personali (definizione contenuta nel Decreto legislativo n.196/2003);
- *Incaricati*: le persone fisiche autorizzate a compiere operazioni di trattamento dal titolare o dal responsabile (definizione contenuta nel Decreto legislativo n.196/2003);
- *Interessato*: la persona fisica, la persona giuridica, l'ente o l'associazione cui si riferiscono i dati personali (definizione contenuta nel Decreto legislativo n.196/2003;
- *Comunicazione*: il dare conoscenza dei dati personali a uno o più soggetti determinati diversi dall'interessato, dal rappresentante del titolare nel territorio dello Stato, dal responsabile e dagli incaricati in qualunque forma, anche mediante la loro messa a disposizione o consultazione (definizione contenuta nel Decreto legislativo n.196/2003);
- *Diffusione*: il dare conoscenza dei dati personali a soggetti indeterminati, in

qualunque forma, anche mediante la loro messa a disposizione o consultazione (definizione contenuta nel Decreto legislativo n.196/2003);

- *Dato anonimo*: il dato che in origine, o a seguito di trattamento, non può essere associato ad un interessato identificato o identificabile (definizione contenuta nel Decreto legislativo n.196/2003);
- *Unità statistica*: entità alla quale sono riferiti o riferibili i dati trattati (definizione contenuta nel Codice deontologico); unità elementare alla quale si riferisce l'informazione statistica (definizione contenuta nel Regolamento n.1588/1990 - Euratom/Cee);
- *Identificazione diretta*: identificazione di una unità statistica in base al suo nome o al suo indirizzo ovvero ad un numero di identificazione ufficialmente attribuito e reso pubblico (definizione contenuta nel Regolamento n.1588/1990 - Euratom, Cee);
- *Dati identificativi*: i dati personali che permettono l'identificazione diretta dell'interessato (definizione contenuta nel Decreto legislativo n.196/2003)
- *Dato identificativo indiretto*: insieme di modalità di caratteri associate o associabili ad una unità statistica che ne consente l'identificazione con l'impiego di tempi e risorse ragionevoli;
- *Identificabilità di un interessato*: un interessato si ritiene identificabile quando, con un elevato grado di probabilità, è possibile stabilire, con un impiego di mezzi ragionevoli, una relazione biunivoca tra la combinazione delle modalità delle variabili relative a una unità statistica e i suoi dati identificativi;
- *Rischio di identificazione*: probabilità di identificare l'interessato. L'interessato si ritiene non identificabile se il rischio di identificazione, tenendo conto dei dati comunicati o diffusi, è tale da far ritenere sproporzionati i mezzi eventualmente necessari per procedere all'identificazione, rispetto alla lesione o al pericolo di lesione dei diritti degli interessati che può derivarne, avuto altresì riguardo al vantaggio che se ne può trarre (definizione contenuta nel Codice deontologico);
- *Mezzi* (ragionevolmente utilizzabili per identificare un interessato): afferiscono, in particolare, alle seguenti categorie:
  - risorse economiche;
  - risorse di tempo;
  - archivi nominativi o altre fonti di informazione contenenti dati identificativi congiuntamente ad un sottoinsieme delle variabili oggetto di comunicazione o diffusione;
  - archivi anche non nominativi, che forniscano ulteriori informazioni oltre a quelle oggetto di comunicazione o diffusione;
  - risorse hardware e software per effettuare le elaborazioni necessarie per collegare informazioni non nominative ad un soggetto identificato, tenendo anche conto delle effettive possibilità di pervenire in modo illecito alla sua identificazione in rapporto ai sistemi di sicurezza ed ai software di controllo adottati;

- conoscenza delle procedure di estrazione campionaria, imputazione, correzione e protezione statistica adottate per la produzione dei dati.
  (definizione contenuta nel Codice deontologico)

- *Variabile pubblica*: il carattere o la combinazione di caratteri, di tipo qualitativo o quantitativo, oggetto di una rilevazione statistica che faccia riferimento ad informazioni presenti in pubblici registri, elenchi, atti, documenti o fonti conoscibili da chiunque (definizione contenuta nel Codice deontologico);
- *Istituto o ente di ricerca*: organismo pubblico o privato per il quale la finalità di ricerca risulti dagli scopi societari o istituzionali e la cui attività scientifica sia documentabile.

## 1.4 La tutela della riservatezza dal punto di vista statistico

La traduzione dei concetti enunciati nella legge in regole operative dal punto di vista statistico avviene all'interno di un quadro di riferimento metodologico caratterizzato essenzialmente da una precisa definizione di cosa si intenda per violazione della riservatezza, dalle specificazioni di come questa possa verificarsi, dallo sviluppo di metodi che possano quantificare la probabilità di violazione e la messa a punto di tecniche di protezione dei dati. E' evidente che queste problematiche sono di natura prettamente statistica.

Nell'ottica sopra esposta, il processo di tutela della riservatezza del dato da rilasciare comporta comunque sempre due fasi ben distinte: nella prima, che potremmo identificare come la *fase definitoria* del processo (Franconi, 1999), occorre identificare, in base alla definizione di violazione della riservatezza adottata, le unità statistiche soggette a rischio di identificazione fornendo una misura del rischio stesso, spesso attraverso modelli probabilistici; nella seconda fase, che chiameremo la *fase operativa*, si adottano i provvedimenti necessari per la protezione del dato. Le tecniche di protezione generalmente comportano una riduzione del contenuto informativo dei dati rilasciati (perdita di informazione) e alle volte richiedono che la valutazione del rischio di violazione sia effettuata anche *a posteriori* per verifica, all'interno di un processo ciclico delle due fasi.

Affinché si possa verificare una violazione della riservatezza occorre che esista un "intruso" ovvero un utente cui siano stati rilasciati dati statistici (sotto qualsiasi forma: tabelle, file, eccetera) e che abbia intenzione di ricavare da questi informazioni riservate. Inoltre, l'intruso deve avere le capacità e delle informazioni ulteriori per poter mettere in atto i suoi propositi. Più operativamente, dato per scontato che il rilascio di informazione statistica nella fase di diffusione in nessun caso riguarda dati identificativi diretti, si assume che un eventuale intruso abbia a disposizione un archivio nominativo, contenente cioè gli estremi identificativi di individui o imprese (nome, cognome, codice fiscale eccetera) e informazioni tali da consentirgli di attribuire uno o più dati statistici rilasciati ai nominativi in suo possesso. Si noti che in questa definizione non ha rilevanza né la dimensione (è sufficiente che venga re-identificata una unità statistica per avere una violazione) né la natura (cartacea, informatizzata, informazioni di

pubblico dominio, eccetera) dell'archivio a disposizione dell'intruso.

In letteratura esistono differenti definizioni di violazione della riservatezza, si veda Duncan e Lambert (1989) per una trattazione completa; quella cui faremo riferimento in questo manuale è quella più comunemente utilizzata dagli Istituti nazionali di statistica, basata sul concetto di "identificabilità di un interessato", che implica un'identificazione di un'unità statistica. Si verifica una identificazione quando, con un certo grado di sicurezza, si riesce a stabilire una relazione biunivoca tra la combinazione delle modalità dei dati identificativi indiretti di un'*unità* presente nel file di dati rilasciati e l'unità dell' archivio esterno in possesso dell'intruso. La violazione della riservatezza contraddistingue la possibilità per l'intruso di acquisire informazioni riservate relative ad un'unità statistica a partire dai dati pubblicati, siano questi collezioni campionarie di dati elementari o dati aggregati.

Vediamo più in dettaglio come questo può accadere. Innanzitutto è utile ribadire che la statistica ha come obiettivo naturale quello di sintetizzare l'informazione raccolta. Soprattutto nel caso della statistica ufficiale, che tratta grandi quantità di dati, la singola informazione e, a maggior ragione, la singola unità statistica non hanno significato statistico autonomo, ma contribuiscono in piccola parte alla produzione di indici o aggregati sintetici. Esistono delle eccezioni in particolare fra le imprese. Si pensi ad esempio alla Fiat nel campo della produzione di automobili in Italia, o alla Benetton nel settore tessile del Nord-est italiano. Il contributo di imprese di tali dimensioni è spesso predominante nella costruzione di aggregati specifici. Vale, comunque, il principio per cui in fase di rilascio di informazione statistica l'identificazione delle singole unità non è utile per gli utenti e pertanto vengono omesse tutte quelle informazioni che identificano direttamente le stesse (nome, ragione sociale, codice fiscale, eccetera). Un intruso che voglia re-identificare una unità statistica deve, quindi, basarsi esclusivamente sulle informazioni rilasciate o più precisamente su quelle che vengono definite identificativi indiretti. Il meccanismo con cui una re-identificazione può avvenire può essere immediato o affidato a più o meno complessi algoritmi di abbinamento di informazioni (*record linkage*, *statistical matching*, eccetera). Per chiarire facciamo alcuni semplici esempi. Supponiamo che vengano rilevate presso gli ospedali le "cause di ricovero" per "comune di residenza", "età" e "titolo di studio" del ricoverato e ne risulti che nel comune di Montemignaio risieda un solo laureato in statistica trentacinquenne ricoverato per un'infezione alle vie respiratorie (i dati riportati sono puro frutto di fantasia). Supponiamo inoltre che per la specificità della laurea e la dimensione molto piccola in termini di numero di abitanti del comune di Montemignaio esista alla data del ricovero un solo abitante laureato in statistica e trentacinquenne. In tal caso è lecito immaginare che la persona in questione sia facilmente riconoscibile da parte di molti suoi compaesani, conoscenti e probabilmente da chiunque sia interessato a riconoscerlo con un modesto impegno di risorse. Trattandosi di un esempio, non sono qui rilevanti i motivi per cui qualcuno vorrebbe tentare un'identificazione; né il tipo di diffusione che viene data dell'informazione (sui mass media, su internet, in una pubblicazione, per comunicazione diretta, eccetera), di cui è giusto normalmente tenere conto. Pertanto, il rilascio di questa combinazione di dati consentirebbe a un eventuale intruso di conoscere informazioni sullo stato di salute di questa persona senza che la stessa sia informata o consenziente. In questo caso la violazione della riservatezza è

evidente e particolarmente grave perché riguarda informazioni delicate come quelle sullo stato di salute degli individui. Nell'esempio il comune di residenza, il titolo di studio e l'età hanno svolto la funzione di identificativi indiretti in quanto informazioni facilmente in possesso di un eventuale intruso associate o associabili al nome di una persona (identificativo diretto dell'interessato). La causa del ricovero invece è l'informazione acquisita indebitamente ma che in generale non è nota al generico intruso (rappresenta il contenuto della violazione). La causa del ricovero non è di per sé un identificativo indiretto in quanto, non essendo nota, non consente il riconoscimento dell'interessato. Questo tipo di informazioni (nell'esempio fatto "la causa di ricovero") vengono identificate come "sensibili", "confidenziali" o "riservate".

Occorre specificare che il termine "sensibile" è utilizzato per identificare certe categorie di dati personali particolarmente delicate (art.4 comma 1 lettera d del Codice unico), tra le quali le condizioni di salute di una persona, per cui sono previste cautele particolari (anche per le fasi di raccolta e conservazione delle informazioni) mentre nell'ambito della riservatezza statistica tutte le informazioni raccolte sugli interessati vengono considerate sensibili, cioè riservate e da proteggere contro eventuali tentativi di violazione salvo che non siano di dominio pubblico. Nel nostro ambito, quindi, non esistono variabili più o meno sensibili, lo sono tutte allo stesso modo. Nel Codice deontologico l'introduzione del "principio di ragionevolezza" nel valutare i mezzi che un intruso è disposto a utilizzare per tentare una violazione consente almeno in parte di mediare il livello di protezione richiesta al tipo di informazione rilasciata contemperandolo alla "lesione o il pericolo di lesione ai diritti degli interessati".

Dal punto di vista metodologico le variabili presenti in un file o in una tabella da rilasciare vengono classificate in due gruppi:

- variabili identificative (*identifying variables*): sono quelle che contengono informazioni utili per un tentativo di identificazione. Vengono indicate anche come "variabili chiave" (*key variables*), o "variabili note" proprio per mettere in evidenza il fatto che possono essere informazioni conosciute, cioè in possesso dell'intruso, attraverso le quali è possibile riconoscere un'unità statistica;
- variabili riservate (*confidential data*) per indicare che non essendo note all'intruso, questi non può usarle per tentare una re-identificazione. Vengono indicate anche con i termini "confidenziali" o "informazioni sensibili".

Alle volte è possibile individuare un sottoinsieme delle variabili identificative come variabili "pubbliche" ossia relative ad informazioni di pubblico dominio, quindi non soggette a tutela. Per la loro caratteristica di essere pubbliche, quindi note o disponibili per chiunque, sono naturalmente delle variabili identificative. Il termine "pubblico" è stato spesso associato alla caratteristica di essere noto e questo ha portato in alcuni casi a usare il termine "variabile pubblica" come sinonimo di "variabile identificativa", ma i due insiemi non necessariamente coincidono ed è bene distinguerli.

Continuando ad analizzare l'esempio precedente, proviamo ad immaginare che la stessa informazione venga rilasciata in maniera meno dettagliata. Supponiamo che da una tabella pubblicata si evinca che vi sia una sola persona ricoverata per un'infezione alle vie respiratorie residente nella provincia di Arezzo, laureata e di età compresa fra i 30 e i 40 anni. In base alle informazioni precedenti sappiamo che è la stessa persona

perché vi è un solo ricoverato per quella causa ma un eventuale intruso stavolta avrebbe a disposizione come identificativi indiretti i seguenti dati: laureato, residente nella provincia di Arezzo, di età compresa fra i 30 e i 40 anni. E' ovvio che il numero di persone che rispondono a questa descrizione si contano in migliaia ed è, quindi, "ragionevolmente" impossibile pensare che un individuo possa essere riconosciuto per questo. Cioè, le informazioni rilasciate non consentono all'intruso di associare il nome di una persona agli identificativi indiretti se non con una probabilità molto bassa di effettuare la scelta giusta. In questo caso il rilascio dei dati viene considerato "sicuro". Si potrebbe obiettare che alcune persone vicine al ricoverato in questione (amici, parenti, eccetera), potrebbero avere conoscenza diretta della causa del ricovero e ancor più facilmente delle altre caratteristiche (residenza, età e titolo di studio) e in base a ciò riconoscere dai dati rilasciati l'interessato. Tuttavia, una tale identificazione non viene considerata propriamente una violazione in quanto il riconoscimento verrebbe effettuato proprio in virtù dell'informazione che doveva essere riservata (la causa del ricovero). In pratica gli intrusi in questione non acquisirebbero dai dati rilasciati alcuna informazione ulteriore a quelle già in loro possesso e, pertanto, si tratterebbe di una violazione senza contenuto.

Esempi di possibili violazioni della riservatezza sono ancora più evidenti nel caso delle imprese proprio a causa delle caratteristiche peculiari che presentano alcune di esse. In particolare sono le grandi imprese ad essere facilmente riconoscibili a causa delle loro dimensioni. Per fare un esempio più specifico, anche in questo caso del tutto inventato, supponiamo che venga rilasciata una tabella relativa alle indagini sul commercio estero dalla quale si evince che le esportazioni di una particolare fibra tessile in un paese del sud-est asiatico ammontano a una certa cifra. Supponiamo, anche, che le aziende che producono quella particolare fibra siano solo due in Italia, in concorrenza tra loro. Supponiamo, infine, che solo una sia azienda esportatrice e abbia mantenuta riservata questa informazione per non svelare le proprie strategie di espansione in paesi che rappresentano un potenziale mercato per quello specifico prodotto. Dalla tabella rilasciata l'azienda che non ha esportato (o ha esportato in altri paesi) riconosce o potrebbe riconoscere la sua concorrente in quanto sono solo in due a produrre quella specifica fibra. In tal modo potrebbe venire a conoscenza del fatto che la sua concorrente effettua delle esportazioni in un paese che rappresenta un nuovo mercato per quello specifico prodotto. In questo caso il rilascio di una tale informazione potrebbe comportare non solo una violazione della riservatezza ma anche una violazione delle norme sulla concorrenza in quanto verrebbe messa un'azienda in situazione di privilegio (informativo) rispetto all'altra. Anche in questo caso possiamo distinguere gli identificativi indiretti dall'informazione sensibile. I primi sono rappresentati semplicemente dal "tipo di prodotto esportato" che da solo consente il riconoscimento di un'azienda da parte della sua concorrente.[3] Mentre i dati riservati consistono nel fatto che il prodotto viene esportato, per quale ammontare e verso quale destinazione. Da notare che nell'esempio l'unità statistica di rilevazione cui si riferiscono le informazioni sono i prodotti esportati riportati nelle bollette doganali, mentre l'interessato, cioè il soggetto da tutelare, è l'azienda esportatrice. Quindi, non

[3] Implicitamente è presente il dettaglio territoriale che, se non indicato, è il più ampio possibile, ad esempio il livello nazionale o quello di riferimento dell'indagine.

necessariamente il soggetto "interessato" cui si riferiscono o possono riferirsi i dati rilasciati coincide con l'unità di rilevazione o è direttamente coinvolto nella rilevazione.

### 1.4.1 La violazione della riservatezza nel rilascio di tabelle

La tipologia di diffusione classica delle statistiche ufficiali è sempre stata la forma aggregata sotto forma ad esempio di indici o tabelle. Ci riferiremo esclusivamente al caso delle tabelle sia perché normalmente le altre forme di diffusione di dati aggregati sono sufficientemente sintetici da non comportare problemi di riservatezza (si pensi ad esempio alla stima degli indici dei prezzi) sia perché i concetti che vengono espressi per le tabelle possono essere facilmente estesi alle altre forme di diffusione.

Il concetto di violazione della riservatezza del rispondente, comunque, non dipende dal tipo di dati rilasciati, aggregati o dati elementari. Pertanto, coerentemente con quanto scritto sopra, anche per i dati tabellari si verificherà una violazione quando si riescano a trarre informazioni riservate aventi carattere individuale a partire dalla tabella pubblicata.

Una tabella costituita esclusivamente di variabili pubbliche non ha bisogno di essere tutelata in quanto, per definizione, non contiene informazioni riservate. Ad esempio, le liste elettorali sono considerati elenchi pubblici e contengono informazioni in merito al comune di nascita e di residenza, al sesso e alla data di nascita di tutti gli abitanti in età di voto. Pertanto una tabella che riporti le frequenze della popolazione per sesso o per classi di età a livello comunale non ha bisogno di essere protetta in quanto le informazioni che contiene sono di pubblico dominio anche se riferite agli individui. Inoltre, la facilità con cui è possibile reperire tali informazioni su un individuo (ad esempio interrogando l'anagrafe comunale) rende decisamente sproporzionati i mezzi che un eventuale intruso dovrebbe impiegare per riconoscere un individuo da una tabella statistica del tipo descritto sopra.

Specularmente, nella maggioranza dei casi, non necessitano di protezione le tabelle relative alle sole variabili riservate, se analizzate singolarmente, perché tramite tali variabili per definizione non è possibile identificare alcuna unità. Sono, come detto, variabili riservate che non contengono identificativi indiretti e quindi non associabili agli identificativi diretti di chicchessia.

La maggioranza delle tabelle di interesse presenta, tuttavia, entrambe le tipologie di variabili congiuntamente e tale incrocio permette, in alcuni casi, di poter giungere ad identificare unità e, di conseguenza, violare la riservatezza (come negli esempi precedenti il contenuto della violazione è dato dal valore delle variabili riservate che potrebbe essere indebitamente conosciuto da un intruso). Per quanto concerne la *fase definitoria* del problema si noti che l'identificazione può avvenire se la cella contiene una sola unità — come nel primo esempio — o due unità — in tal caso una riconosce l'altra come nel secondo esempio — e, per tale motivo, una regola accettata da molti istituti nazionali di statistica impone una frequenza maggiore o uguale a tre per le celle di tabelle pubblicate.

Questa è la cosiddetta **regola della soglia** per cui si considerano dati aggregati le combinazioni di modalità alle quali è associata una frequenza non inferiore a una soglia

prestabilita, ovvero un'intensità data dalla sintesi dei valori assunti da un numero di unità statistiche pari alla suddetta soglia. Il valore minimo attribuibile alla soglia è pari a tre.

Le celle che non rispondono al criterio della soglia o dalle quali si possono trarre riferimenti individuali secondo altri criteri, vengono definite "sensibili" o "a rischio" (*unsafe*).

Ferma restando l'integrità del dato pubblicato — ovvero non facendo ricorso a metodi perturbativi che alterano il dato elementare *a priori* cioè prima della pubblicazione della tabella, o *a posteriori* cioè che modificano la tabella contenente delle celle sensibili, (Cox, 1987) — il problema della protezione di una tabella, nella *fase operativa*, si riduce ad oscurare le celle sensibili e, al tempo stesso, verificare che il valore relativo a tali celle non possa essere ricavato in altro modo. Per evitare ciò si ricorre se necessario a delle ulteriori soppressioni (soppressioni secondarie).

I metodi di perturbazione, invece, consistono nel rilasciare un'informazione diversa da quella vera in modo da limitare le possibilità di identificazione delle unità statistiche e per ridurre il contenuto informativo di un'eventuale violazione. Per le tabelle i metodi più spesso proposti consistono nell'arrotondare secondo qualche criterio i valori nelle celle, (vedi Paragrafo 3.3).

Tutto ciò è valido quando si analizzano e si proteggono tabelle prese singolarmente; tuttavia, la crescente potenza di calcolo e facilità di uso dei computer negli ultimi anni ha reso evidente la limitatezza dell'approccio al problema della tutela della riservatezza dei dati aggregati se ci si limita alla protezione delle *singole* tabelle pubblicate.

Infatti, poiché le tabelle provengono da un medesimo file di dati elementari queste possono essere collegate tra loro e da tale incrocio spesso è possibile ottenere informazioni riservate che non erano desumibili da ciascuna tabella presa singolarmente. Questo è noto come "problema delle *linked tables*". Ovvero, la tutela della riservatezza delle singole tabelle non implica necessariamente la tutela della riservatezza di un insieme di tabelle tratte da uno stesso file.

Un ultimo aspetto da tenere in considerazione riguarda la natura, campionaria o meno, della rilevazione statistica della quale si vogliono diffondere i risultati. Se la rilevazione è campionaria i dati riportati nelle tabelle sono generalmente frutto di un riproporzionamento dei valori rilevati in base ai coefficienti di riporto all'universo. L'individuazione delle celle sensibili in base alla regola della soglia deve allora tenere conto del fatto che i numeri pubblicati sono stime statistiche e potrebbero non corrispondere con le unità rilevate.

Il problema della tutela della riservatezza delle tabelle viene affrontato in maniera approfondita nel Parte Seconda del manuale.

### 1.4.2 La violazione della riservatezza nel rilascio di dati elementari

I dati elementari possono essere definiti come il prodotto finale di una rilevazione statistica dopo le fasi di progettazione, raccolta, controllo e correzione. Dai dati elementari viene prodotto quanto previsto nel piano di pubblicazione di una rilevazione

che, come abbiamo visto, consiste nel rilascio di informazioni sintetiche (dati aggregati) ma che può prevedere anche il rilascio di un file di dati elementari. Ovviamente non tutte le variabili rilevate possono essere rilasciate ma, in generale, il file così prodotto costituisce il massimo contenuto informativo prodotto da una rilevazione statistica cui possa accedere un utente. I dati elementari sono un archivio di record ciascuno relativo a una unità statistica. Ogni record contiene tutte le informazioni (variabili) rilevate sulla relativa unità. Tali variabili, al pari delle tabelle, possono essere classificate tra le variabili chiave in quanto identificativi indiretti, oppure come variabili riservate. Naturalmente gli identificativi diretti (nome, cognome, indirizzo) che ordinariamente non sono ammessi nemmeno alla fase di registrazione dei dati sono eliminati dall'archivio dei dati rilasciati.

L'Istat, in accordo con il D.lgs 322/1989, può distribuire, "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 o giuridiche". Quello che la legge indica come *collegamento* consiste in quello che finora abbiamo indicato con *identificazione* di un'unità statistica.

Rispetto al caso di rilascio di tabelle non cambia il concetto di violazione della riservatezza ma cambiano sostanzialmente sia l'insieme delle variabili chiave che, in generale, nei dati elementari saranno più numerose (in alcuni casi questo comporta che siano meno dettagliate), sia il contenuto di un'eventuale violazione in quanto le variabili riservate nei dati elementari sono presenti tutte insieme.

Per contro, il rilascio di microdati riguarda esclusivamente le collezioni campionarie e l'accesso ai file è molto più controllato (per soli motivi di ricerca e dietro la sottoscrizione di un modulo/contratto).

Tuttavia, non v'è dubbio che il rilascio di dati elementari è questione più delicata rispetto alla diffusione di tabelle, non fosse altro che l'eliminazione di riferimenti individuali da un insieme di record individuali appare sin dalla descrizione un problema difficile se non contraddittorio. Per questo la fase definitoria prevede dei modelli di misurazione del rischio di identificazione specifici rispetto alle tabelle e spesso basati su modelli probabilistici. Mentre per la fase operativa possiamo ricondurre i metodi di protezione a tre categorie:

1. ricodifica di variabili (*global recoding*): consiste nel ridurre il dettaglio di rilascio di alcune variabili (ad esempio l'età in classi quinquennali anziché annuali);
2. soppressione di informazioni (*local suppression*): per eliminare alcune caratteristiche che rendono alcuni record più facilmente identificabili;
3. perturbazione dei dati pubblicati: con metodi diversi ma con le stesse finalità viste per le tabelle, ossia rendere difficile la re-identificazione e alterarne i contenuti per dissuadere eventuali tentativi.

Il rilascio di collezioni campionarie di dati elementari, comunque, è prerogativa dell'Istat e deve essere autorizzata dal suo Presidente, quindi rimandiamo ogni altro approfondimento al Capitolo 3, dedicato a questo argomento.

### 1.4.3 Interrogazioni on line di banche dati su internet

Se i problemi menzionati finora presentano un grado elevato di difficoltà, la vera sfida del futuro è rappresentata dalla creazione di una strategia di rilascio estremamente innovativa che permetta una più veloce e flessibile diffusione dei dati secondo le reali necessità degli utenti esterni. Tale strategia di diffusione comporta la predisposizione di una banca dati che possa essere interrogata on line così che gli utenti possano ottenere in tempo reale i dati aggregati di interesse. Ciò si è già verificato nel caso del Censimento intermedio dell'industria e dei servizi,[4] indagine *short-form*, perché i dati raccolti erano soggetti a uno specifico regolamento che ne ha consentito il rilascio con un tale strumento senza la necessità di adottare misure di protezione di tipo statistico (L'impianto normativo, metodologico e organizzativo – Censimento intermedio dell'industria e dei servizi 31 dicembre 1996, Istat 1998). La medesima strategia di diffusione nel caso in cui siano presenti sia dati pubblici che dati personali comporta la soluzione di problemi statistici e computazionali estremamente complessi. Una soluzione ovvia è quella di proteggere la banca dati *a priori* ovvero prima dell'interrogazione da parte degli utenti ma questo comporta in molti casi una eccessiva riduzione del contenuto informativo della base di dati rispetto a quello che potrebbe essere disponibile. Un'altra soluzione è rappresentata dalla perturbazione dei dati originali ovvero l'introduzione di errori sia addizionando alle variabili quantitative variabili aleatorie con media nulla e varianza dipendente dalla varianza della variabile originale (Duncan e Mukherjee, 2000), sia ricorrendo a tecniche di *data swapping* (Dalenius e Reiss, 1982) che consistono nello scambiare unità all'interno di prefissate classificazioni territoriali. Come ovvio tale soluzione comporta l'adozione della perturbazione della banca dati originaria che costituisce la base di *tutti* i possibili *output* dell'indagine. Questa soluzione protegge la riservatezza del dato in quanto l'utente esterno anche nel caso in cui riesca ad identificare un'unità non è in grado di stabilire se questa rappresenti effettivamente un individuo o un'impresa nella popolazione o piuttosto non sia frutto della perturbazione apportata ai dati. Gli svantaggi di tale approccio sono evidenti; la qualità dei dati è necessariamente compromessa dal livello di perturbazione apportato ai dati elementari e studi hanno dimostrato che tali metodi hanno un effetto distorcente sia sulle distribuzioni che sulle statistiche di base (Little, 1993). Una strategia di protezione simile a questa ora descritta prevede la perturbazione non della banca dati iniziale quanto delle risposte alle interrogazioni degli utenti (Keller-McNulty e Unger, 1998); simili svantaggi possono essere evidenziati in questo caso. Se non si procede alla perturbazione dei dati originali o delle risposte alle interrogazioni è necessario limitare i quesiti posti alla banca dati perché è noto in letteratura che se le interrogazioni alla banca dati non sono vincolate e le risposte alle interrogazioni sono "esatte" (cioè, i dati non sono preventivamente modificati) allora è sempre possibile trarre riferimenti individuali dalle tabelle richieste (Domingo-Ferrer, 1998).

In letteratura sono stati proposti numerosi metodi di tutela della riservatezza basati sulla restrizione dei quesiti (per un quadro generale si veda Adam e Wortmann, 1989), tuttavia il problema risulta estremamente complesso.

[4] http://cens.istat.it/

In Istat una delle prime esperienze di soluzione del problema della tutela della riservatezza nelle interrogazioni a banche dati su internet è fornita da Coeweb,[5] la banca dati per la diffusione on line delle statistiche del commercio con l'estero. L'idea di base è stata la predisposizione di un accurato piano di diffusione che ha permesso di evidenziare gli interessi di utenti nazionali e internazionali e, al tempo stesso, di chiarire in funzione degli incroci delle modalità di variabili interessate il tipo di protezione necessaria. Limitando gli incroci e il dettaglio delle singole variabili e definendo dei criteri più stringenti di protezione è stato possibile tutelare la riservatezza dei rispondenti (Fazio e Menghinello, 2001 e Cicalini, 2001). In un certo senso la base di dati di Coeweb è protetta *a priori* dal punto di vista della tutela della riservatezza, poiché sono permessi solo alcuni tipi di incroci; di conseguenza le interrogazioni effettuate sulla base di dati sono anch'esse protette.

## 1.5 Tipologie di rilascio e tipologie di utenti

Le metodologie, le tecniche e le procedure per la tutela della riservatezza nella fase di rilascio di informazione statistica differiscono sensibilmente a seconda della tipologia del dato — dato elementare o dato aggregato — delle unità statistiche coinvolte — persone fisiche o imprese — e della tipologia di utente. Per affrontare in modo organico le problematiche connesse alla tutela della riservatezza è sempre consigliabile procedere alla definizione di un piano di pubblicazione che disciplini le possibilità di compiere incroci e di raggiungere specifici livelli di dettaglio (geografico, territoriale, merceologico, eccetera). Il piano di pubblicazione può consentire la riduzione, entro soglie ragionevoli, del numero di tabelle che necessitano di protezione mantenendo comunque un contenuto informativo rilevante per gli utenti. Naturalmente, il piano di pubblicazione deve essere coerente ed integrare tutte le forme di rilascio di informazione statistica previste, soprattutto, in ragione del crescente utilizzo di sistemi interrogabili dall'esterno tramite internet. L'efficacia delle regole statistiche per la tutela della riservatezza può essere vanificata se la loro applicazione è indipendente dal complesso delle tipologie di rilascio ed è effettuata in modo non coerente. Per chiarire con un esempio, si pensi al rilascio di due distinte tabelle protette indipendentemente l'una dall'altra e che tale protezione comporti la soppressione di alcune celle (si veda in proposito il Paragrafo 3.1). Se l'applicazione delle regole di protezione non è coerente fra le due tabelle può verificarsi l'eventualità che un valore soppresso in una tabella sia invece pubblicato nell'altra. Analogamente si deve tener conto di tutte le forme di rilascio previste per ciascuna indagine.

### Comunicazione di dati individuali a soggetti non facenti parte del Sistan

Nel caso in cui gli utenti siano soggetti non facenti parte del Sistema statistico nazionale e l'oggetto del rilascio siano dati individuali di cui è titolare un soggetto Sistan, le ipotesi percorribili sono le seguenti:

1. comunicazione di collezioni campionarie di dati individuali resi anonimi (art.10, comma 2 D.lgs. n.322/89 e art.7, comma 1 Codice deontologico)

[5] http://www.coeweb.istat.it

2. comunicazione di dati personali privi di dati identificativi a ricercatori di università o istituti ed enti di ricerca nell'ambito di specifici laboratori (art.7, comma 2 Codice deontologico)
3. comunicazione di dati personali privi di dati identificativi a ricercatori di università istituzioni ed organismi con finalità di ricerca nell'ambito di progetti congiunti e previa sottoscrizione di appositi protocolli di ricerca (art.7, comma 3 Codice deontologico).

Nella comunicazione di collezioni campionarie di dati individuali, che genericamente vengono chiamati *file standard*, il rischio di identificazione deve essere contenuto entro un limite accettabile. Tale limite e la metodologia per la stima del rischio di identificazione sono individuati dall'Istat, che definisce anche le modalità di rilascio dei dati dandone comunicazione alla Commissione per la garanzia dell'informazione statistica (art.4, comma 3 Codice deontologico). Lo stesso vale per i cosiddetti *file ad hoc*, che hanno le stesse caratteristiche dei *file standard* ma derivano dalla richiesta specifica di un utente.

Per la comunicazione di dati personali nell'ambito di specifici laboratori costituiti dai soggetti del Sistema statistico nazionale, il Codice deontologico stabilisce delle condizioni (si veda l'art.7, comma 2 del Codice deontologico) volte in particolare ad assicurare il controllo dell'ambiente di lavoro in cui vengono prodotti i risultati e la successiva verifica degli stessi preventiva al rilascio. Il Paragrafo 9.3 descrive il Laboratorio per l'analisi dei dati elementari (Adele) costituito dall'Istat.

La terza fattispecie di comunicazione di dati personali a soggetti non Sistan prevede la definizione di protocolli di ricerca (a volte detti convenzioni o *fellowship*) nell'ambito di progetti che devono avere la caratteristica peculiare di essere finalizzati anche al perseguimento di compiti istituzionali del soggetto titolare del trattamento che ha originato tali dati. Le altre condizioni essenziali sono fissate nell'art.7, comma 3 del Codice deontologico.

## Comunicazione dei dati tra i soggetti del Sistan

Nell'ambito del Sistan la comunicazione dei dati è regolata da norme specifiche che, per certi aspetti, definiscono un circuito "privilegiato" di circolazione delle informazioni. La materia della comunicazione dei dati tra i soggetti del Sistan è trattata nell'art. 8 del Codice deontologico. E' in fase di definizione una direttiva "attuativa" delle norme contenute nel Codice deontologico che sarà il punto di riferimento procedurale per l'interscambio dei dati all'interno del Sistan.

La comunicazione di dati personali è prerogativa del titolare anche in ambito Sistan. Pertanto l'ufficio di statistica partecipante al Sistan che riceve comunicazione di dati personali da altro ufficio partecipante al Sistan e titolare del trattamento dei dati comunicati, non può ulteriormente comunicarli a terzi compresi altri uffici della stessa amministrazione di appartenenza.

## Diffusione di informazione statistica

I dati raccolti nell'ambito di rilevazioni statistiche comprese nel Programma statistico nazionale da parte degli uffici di statistica non possono essere esternati se non

in forma aggregata, in modo che non se ne possa trarre alcun riferimento relativamente a persone identificabili e possono essere utilizzati solo per scopi statistici (D.lgs. 322/1989, art.9 comma 1).

Nel Codice deontologico (art.4) sono specificati alcuni criteri per la valutazione del rischio di identificazione nella diffusione di risultati statistici qui di seguito riportati:

a) si considerano dati aggregati le combinazioni di modalità alle quali è associata una frequenza non inferiore a una soglia prestabilita, ovvero un'intensità data dalla sintesi dei valori assunti da un numero di unità statistiche pari alla suddetta soglia. Il valore minimo attribuibile alla soglia è pari a tre;

b) nel valutare il valore della soglia si deve tenere conto del livello di riservatezza delle informazioni;

c) i risultati statistici relativi a sole variabili pubbliche non sono soggetti alla regola della soglia;

d) la regola della soglia può non essere osservata qualora il risultato statistico non consenta ragionevolmente l'identificazione di unità statistiche, avuto riguardo al tipo di rilevazione e alla natura delle variabili associate;

e) i risultati statistici relativi a una stessa popolazione possono essere diffusi in modo che non siano possibili collegamenti tra loro o con altre fonti note di informazione, che rendano possibili eventuali identificazioni;

f) si presume che sia adeguatamente tutelata la riservatezza nel caso in cui tutte le unità statistiche di una popolazione presentino la medesima modalità di una variabile.

Nel programma statistico nazionale sono individuate le variabili che possono essere diffuse in forma disaggregata, ove ciò risulti necessario per soddisfare particolari esigenze conoscitive anche di carattere internazionale o comunitario.

**Prospetto 1.1 Schema riepilogativo delle regole, norme o procedure per tipologia di rilascio e tipologia di utenti**

| TIPOLOGIA DI DATI | | TIPOLOGIA DI UTENTI | | | |
|---|---|---|---|---|---|
| | | SISTAN | SOGGETTI NON SISTAN | EUROSTAT E ORGANISMI INTERNAZIONALI | PUBBLICO |
| MICRODATI | ECONOMICO | Codice deontologico | Protocolli ricerca | Regolamenti | NO |
| | SOCIALE | Codice deontologico | Protocolli ricerca<br>File standard (ad hoc) | Regolamenti | NO |
| TABELLE | FREQUENZA | Soglia (a) | | | |
| | INTENSITA' | Soglia (a)<br>Dominanza | | | |
| BASI DI DATI SU WEB | PROTETTE | (b) | | | |
| | NON PROTETTE | Accesso controllato tramite password<br>Interrogazioni vincolate<br>Protezione sui risultati dell'interrogazione | | | NO |
| ON SITE | LABORATORIO ADELE | NO | (c) | | NO |

[(a)] In alcuni casi leggi e regolamenti prevedono deroghe alla regola della soglia per alcune informazioni; in altri casi la soglia può essere fissata a un valore superiore a 3; in alcuni casi si adotta anche la regola della dominanza (si veda in proposito il Paragrafo 2.5).
[(b)] tipo di rilascio non soggetto a vincoli in quanto la base di dati è protetta per costruzione.
[(c)] L'accesso al Laboratorio è consentito ai soli ricercatori appartenenti ad organismi o enti aventi finalità di ricerca (al Laboratorio Adele è dedicato il Paragrafo 9.3; informazioni sono disponibili sul sito http://www.istat.it seguendo il percorso PRODOTTI E SERVIZI/LABORATORIO ANALISI DATI ELEMENTARI, oppure scrivendo a adele@istat.it).

# PARTE SECONDA

## La tutela statistica della riservatezza per dati aggregati

# Capitolo 2. Rischio di violazione per dati aggregati(*)

## 2.1 Introduzione

In questo capitolo si affronta il problema del tutela statistica della riservatezza per dati aggregati in tabelle. Una tabella consiste nell'aggregazione di dati elementari all'interno di celle. Ogni cella è definita da una combinazione di categorie di una o più variabili categoriche, dette classificatrici.

Le tabelle si distinguono in tabelle di frequenza e tabelle di intensità. Le celle delle tabelle di frequenza contengono la numerosità delle unità che vi appartengono, anche dette rispondenti, mentre le celle delle tabelle di intensità contengono le somme dei valori, detti contributi, che una variabile numerica assume per ogni rispondente. Quindi, ad ogni tabella di intensità ne corrisponde una di frequenze, che può essere nota o meno, mentre ad ogni tabella di frequenze possono corrispondere una o più tabelle di intensità.

Le tabelle maggiormente trattate nell'ambito della tutela della riservatezza sono quelle di intensità, in quanto contengono maggiore informazione rispetto a quelle di frequenza e gli scenari di violazione sono molteplici. In particolare, le tabelle di dati di impresa spesso richiedono una tutela maggiore di quelle di dati sociali. Questo per diversi motivi, tra cui ricordiamo: perché spesso la minore numerosità della popolazione porta anche all'identificazione dell'impresa e perché è plausibile ipotizzare che ci siano tentativi sistematici di violazione da parte di imprese concorrenti.

La tutela statistica della riservatezza per le tabelle consta di due fasi: 1) la valutazione del rischio di violazione di ogni cella; 2) la protezione dell'intera tabella. Il rischio di violazione per una cella di una tabella di intensità è definito in termini dell'accuratezza con cui un intruso può determinare i valori di un contributo o della somma di più di uno di essi. La soglia minima di accuratezza ammessa determina il grado di protezione richiesto. Nei casi in cui la conoscenza dei singoli contributi non è sufficiente all'identificazione degli individui, una tabella di intensità è protetta se lo è la tabella di frequenze corrispondente. Il rischio per le celle di una tabella di frequenze, è legato alla possibilità di identificare uno o più individui appartenenti a modalità sensibili. In questo capitolo tratteremo prima la valutazione del rischio e poi le tecniche di protezione. Discuteremo separatamente, per quanto possibile, le tabelle di frequenze e quelle di intensità, cominciando dalle prime.

Una cella di una tabella con *p* dimensioni è identificata dall'insieme degli indici delle categorie delle variabili classificatrici a cui corrisponde, così la cella corrispondente alle categorie $(X_1=x_{i1},X_2=x_{j2},\ldots,X_p=x_{rp})$ è identificata dal simbolo $C_{ij\ldots r}$. Comunque, per semplicità, ci riferiremo a tabelle con una o due dimensioni. Quando possibile, una generica cella sarà indicata semplicemente come $C_i$ o anche solo come C. Le variabili classificatrici si definiscono note quando sono variabili pubbliche o comunque che si suppongono conosciute dall'intruso. Con questa definizione, quindi, si vuole estendere il concetto di variabile pubblica (cioè conoscibile da chiunque) a quello di variabile conosciuta o conoscibile anche solo da un ipotetico intruso nello scenario di violazione adottato.

(*) Capitolo redatto da Giovanni M. Merola

## 2.2 Tabelle di frequenza

I valori delle celle di una tabella di frequenza sono le numerosità dei rispondenti che vi appartengono e non sono sensibili in quanto tali. Ovviamente, se tutte le variabili classificatrici sono note o se la tabella è stata ottenuta aggregando dati già noti, la tabella non va protetta in quanto non aggiunge informazioni ma semplicemente sintetizza informazioni disponibili. Le tabelle in cui tutte le modalità delle variabili classificatrici sono sensibili richiedono minor protezione in quanto non è possibile identificare i rispondenti, però occorre proteggere le celle con zero rispondenti. Quindi, le tabelle di frequenza che richiedono protezione sono quelle in cui sono presenti variabili classificatrici sensibili o che hanno almeno una modalità sensibile.

Il controllo statistico della riservatezza per tabelle di frequenza non ha regole precise, ma si basa su regole generali che vanno applicate discrezionalmente per ogni tabella (Willenborg e de Waal, 2001). Esemplificheremo alcune situazioni con la Tabella 2.1, che è un esempio fittizio di tabella ad alto rischio di violazione. La tabella è sensibile in quanto riporta un'informazione riservata su ognuna delle imprese del ramo di attività X rilevate, come l'essere stati denunciati (in attesa di sentenza) o condannati per violazioni alle leggi per la tutela dell'ambiente a fronte della classificazione per area geografica di attività, che è pubblica.

**Tabella 2.1 Precedenti per reati ambientali delle imprese nel ramo X per area**

| Imprese ramo attività X | | Area geografica di attività | | | | Totale |
|---|---|---|---|---|---|---|
| | | Nord est | Nord ovest | Centro | Sud | |
| Precedenti per reati Ambientali | Nessuno | 0 | 1 | 2 | 6 | 9 |
| | Denuncia | 0 | 0 | 7 | 1 | 8 |
| | Condanna | 9 | 5 | 3 | 4 | 21 |
| | Totale | 9 | 6 | 12 | 11 | 48 |

Si noti che le categorie "Denuncia" e "Condanna" della variabile classificatrice "Precedenti per reati ambientali" si considerano sensibili in quanto forniscono informazioni riservate, anche se, in principio, sono disponibili. Infatti, anche il facilitare informazioni non facilmente reperibili su di un rispondente costituisce violazione. La categoria "Nessuno" si può ritenere sensibile in quanto è un'informazione riservata, e perché alcune imprese potrebbero essere svantaggiate dall'identificazione di un'altra impresa in questa categoria.

Nella Tabella 2.1 i dati relativi alla regione "Nord-est" non sono rilasciabili in quanto rendono noto che tutte le imprese di quell'area sono state condannate. Questo è un esempio di violazione di gruppo (*group disclosure*). I dati del "Nord-ovest" costituiscono un'altra violazione di gruppo, infatti l'unica impresa che non ha precedenti è posta a conoscenza che tutte le altre hanno subito condanne. Le imprese del "Centro" non sono a rischio di identificazione diretta però le due imprese senza precedenti sanno che quasi tutte le altre hanno precedenti (cioè sono nella categoria aggregata "Denunciati o condannati per violazioni alle leggi per la tutela dell'ambiente"). Inoltre, coalizzandosi queste due imprese potrebbero violare la colonna. In questo caso le unità sono troppo concentrate in categorie sensibili;

l'esistenza di casi come questo rende necessario adattare i parametri delle regole di tutela ad ogni tabella secondo la natura dei dati. I dati del "Sud" forniscono un esempio di come non sempre le tabelle che presentano frequenze pari ad uno siano violabili. Infatti, è vero che si rende noto che solo un'impresa del Sud è stata denunciata, però è impossibile identificare quale essa sia, così come essa non può identificare le imprese nelle altre celle.

Alcuni criteri a cui ci si deve attenere per la protezione delle tabelle di frequenza sono elencati di seguito.

1. Evidentemente tutte le categorie definite in corrispondenza di una modalità di una variabile nota che ha frequenza marginale minore di tre sono violabili. Quando le categorie sensibili sono due anche la frequenza marginale pari a tre comporta rischio, come mostrato nella Tabella 2.2. In generale, quando la frequenza marginale nota è bassa relativamente al numero delle categorie della variabile sensibile, le unità saranno sparse ed il rischio di violazione (specialmente di gruppo) può essere elevato.
2. In alcuni casi il valore 0 in una cella è informativo in quanto identifica tutte le unità come non appartenenti ad una categoria della variabile sensibile. Nell'esempio della Tabella 2.1, si sa con certezza che nessuna impresa del "Nord" ha denunce in corso. La frequenza 0 può comportare rischio se cade in una categoria sensibile o nell'unica categoria non sensibile.

**Tabella 2.2 Possibili distribuzioni di tre rispondenti in due categorie sensibili**

| | | Nota | | | | |
|---|---|---|---|---|---|---|
| | | A | B | C | D | Totale |
| Sensibile | Sì | 3 | 2 | 1 | 0 | 6 |
| | No | 0 | 1 | 2 | 3 | 6 |
| | Totale | 3 | 3 | 3 | 3 | 12 |

3. Una regola generale per la tutela delle tabelle di frequenze è che la tabella è a rischio di violazione se le unità in una categoria sensibile sono più del $p$% del totale marginale noto. Il valore di $p$ deve essere scelto in base al numero e alla sensibilità delle categorie.

## 2.3 Tabelle di intensità

Ogni cella C di una tabella di intensità contiene il valore $z$ che è la somma dei contributi, cioè dei valori di una variabile assunti dagli $n$ rispondenti di quella cella. I contributi si assumono non negativi e saranno indicati con $z_j$ e, senza perdere in generalità, saranno indicizzati in ordine non crescente, $z_1 \geq z_2 \geq \cdots \geq z_n \geq 0$. Cosicché il valore della cella è $z = \sum_{j=1}^{n} z_j$. Le somme parziali dei contributi di una cella saranno denotati nel seguente modo: $t_m = \sum_{j=1}^{m} z_j$ e $r_m = \sum_{j=m+1}^{n} z_j$, cosicché risulta $z = t_m + r_m$.

Il problema della tutela della riservatezza per questo tipo di tabelle è diverso da quello per le tabelle di frequenze. Nelle tabelle di intensità le variabili classificatrici possono essere tutte note e si devono proteggere i singoli contributi delle celle. Le

tabelle che richiedono maggior tutela sono quelle in cui un intruso conosce la tabella di frequenze (specialmente quando tutte le variabili classificatrici sono note) e può identificare gli individui appartenenti alle celle in base al valore del contributo. Anche nei casi in cui si può ragionevolmente supporre che le informazioni a disposizione di intrusi siano più limitate, è comunque necessario adottare delle precauzioni che tutelino la riservatezza. Magari rilassando i criteri di protezione.

Nella prossima sezione si considerano tabelle di intensità per l'intera popolazione con contributi non negativi e possibilità di identificare i rispondenti. Più avanti saranno trattate le tabelle campionarie e quelle in cui i contributi possono essere negativi.

## 2.4 Misure di rischio di violazione per tabelle di intensità

Le misure del rischio di violazione per le celle si basano sulla accuratezza della miglior stima di uno o più dei contributi da parte di un intruso, si parla infatti di violazione predittiva. Il rischio viene misurato in termini del minimo errore relativo di stima (ER) per una data funzione dei contributi, per cui, una cella si considera a rischio se il minimo ER è minore di una certa soglia, che indicheremo con *p*. Nelle misure di rischio più stringenti si presuppone che un intruso abbia delle informazioni ausiliarie (informazioni *a priori*) e sia interessato a conoscere uno o più contributi di una cella. Il rischio di violazione per una cella si misura rispetto al caso più favorevole all'intruso (caso peggiore per la protezione). In pratica, i casi peggiori sono quelli in cui l'intruso è uno dei rispondenti e, quindi, possiede almeno il proprio contributo come informazione ausiliaria.

Si definisce misura di rischio di violazione per la cella C una funzione $S(C)=f(z_1, z_2, \ldots, z_n)$ per cui una cella non è sicura se $S_d(C) \geq d$, dove *d* è una costante. Di solito ci si riferisce alla misura $S(C)= S_d - d \geq 0$ per cui una cella è a rischio se $S(C) \geq 0$.

Le misure di rischio maggiormente utilizzate sono quelle lineari (Cox , 1981) definite da:

$$S(C) = \sum_{j=1}^{n} z_j \lambda_j ,$$

dove i $\lambda_j$ sono pesi associati ai singoli contributi che caratterizzano le diverse misure. Si noti che le misure di rischio sono definite solo rispetto al loro segno e, quindi, misure proporzionali sono equivalenti. Una proprietà auspicabile di una misura di rischio è la sub-additività, cioè che non cresca se una cella è aggregata ad un'altra, ovvero che: $S(C_i \cup C_k) \leq S(C_i) + S(C_k)$. Per le misure lineari la sub-additività è verificata se e solo se $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n_i}$ (si veda Cox, 1981). Questa proprietà garantisce che l'aggregazione di celle della tabella non aumenti il rischio di violazione ed è di fondamentale importanza per la protezione delle tabelle per mezzo della soppressione. Si tenga presente, però, che le misure di rischio sub-additive non coprono tutti i possibili scenari di violazione predittiva, come si vedrà più avanti.

## 2.5 Esempi di misure di rischio lineari subadditive

### 1) Regola della soglia

Questa misura è analoga a quella di cui si è parlato alla fine della sezione sulle tabelle di intensità e richiede che il numero dei rispondenti di una cella, $n$, sia maggiore di una certa soglia. Una cella non è mai sicura se $n < 3$. Infatti, quando $n$=1 la pubblicazione di $z$ equivale a dichiarare il valore da proteggere e quando $n$=2 entrambi i rispondenti sono in grado di conoscere esattamente il valore dell'altro. La soglia di tre garantisce contro la predizione esatta (errore di stima uguale a zero) ed è presente negli standard dell'Istat. Comunque, la regola della soglia può essere estesa ad un numero minimo di contributi maggiore di tre.
Denotando con $m$ il valore della soglia, si pone $\lambda_j = -1/z_j$ così che la misura di rischio è:

$$S = \sum_{j=1}^{n} -\frac{z_j}{z_j} + m = m - n.$$

Quando $n < m$, $S > 0$ e quindi la cella sarà dichiarata a rischio. La sub-additività di questa misura è implicata dai pesi non crescenti ed è verificata immediatamente osservando che:

$$\lambda_1 = -\frac{1}{z_1} \geq \lambda_2 = -\frac{1}{z_2} \geq \cdots \geq \lambda_n = -\frac{1}{z_n}.$$

La regola della soglia è basata solo sul numero dei contributi di una cella e quindi, per una maggiore protezione, dovrebbe essere usata in combinazione con un'altra misura basata sul valore dei contributi.
Si consideri la Tabella 2.3 come esempio fittizio di tabella di intensità. I dati si riferiscono al fatturato di imprese che producono strumenti musicali; le variabili classificatrici sono la regione di attività e la tipologia di strumento prodotto. Le variabili classificatrici si suppongono note ed alcune delle frequenze associate alla tabella, che sono quindi note, sono state incluse ed indicate in parentesi. Le celle relative ai produttori di pianoforti nelle aree B e C non sono pubblicabili in quanto hanno frequenza minore di tre.

### 2) Regola della dominanza

Questa misura, quasi sempre accettata, è una misura di concentrazione dei contributi basata sul principio che una cella non è sicura se il rapporto della somma di $m$ contributi sul totale è maggiore di $k$, con $0 \leq k \leq 1$. Poiché i contributi sono ordinati, sarà sufficiente verificare che i primi $m$ soddisfano questa condizione, e cioè che $t_m / z \leq k$. Quindi la misura di rischio è:

$$S = \sum_{j=1}^{m} z_j - kz = (1-k)\sum_{j=1}^{m} z_j - k \sum_{j=m+1}^{n} z_j = (1-k)t_m - kr_m$$

per cui i pesi per questa misura sono:

$$\begin{cases} \lambda_j = (1-k) & j \le m \\ \lambda_j = -k & j > m \end{cases}$$

Poiché questi pesi sono non crescenti anche questa misura è sub-additiva.

**Tabella 2.3 Fatturato di imprese produttrici di strumenti musicali per regione e tipo di strumento prodotto**

| Strumenti | Regione | | | | | | Totale | |
|---|---|---|---|---|---|---|---|---|
| | A | | B | | C | | | |
| | Fatturato | Freq | Fatturato | Freq | Fatturato | Freq | Fatturato | Freq |
| Arpe | 58 | ($f_{11}$) | 47 | ($f_{21}$) | 36 | ($f_{31}$) | 141 | ($f_{.1}$) |
| Organi | 71 | | 124 | | 24 | | 219 | ($f_{.2}$) |
| Pianoforti | 92 | (5) | 157 | (2) | 59 | (1) | 308 | (8) |
| Altro | 800 | | 934 | | 651 | | 2385 | ($f_{.4}$) |
| Totale | 1021 | ($f_{1.}$) | 1262 | ($f_{2.}$) | 770 | ($f_{3.}$) | 3053 | ($f_{..}$) |

La scelta dei parametri *m* e *k* può essere fatta in base ad una misura di concentrazione massima richiesta. Occorre però che la regola della dominanza sia soddisfatta almeno nel caso di minima concentrazione (cioè *n* osservazioni uguali), perciò i parametri *m* e *k* devono essere tali che $n > \lfloor m/k \rfloor$, in quanto è sempre verificato che $\frac{t_m}{z} \ge \frac{m}{n}$, con uguaglianza nel caso di equidistribuzione. Quindi, per esempio, la dominanza con parametri *m*=3 e *k*=0,6 può essere applicata solo nelle celle in cui $n \ge 5$; infatti, in una cella con tre osservazioni uguali risulta $t_3 / z = z / z = 1$, così come in una cella con quattro osservazioni uguali a $\bar{z}$ si ha $t_3 / z = 3\bar{z} / 4\bar{z} = 0.75$ e in una cella con cinque osservazioni uguali $t_3 / z = 0.6$.

La scelta di *m* e *k* può anche essere fatta in funzione del grado di protezione contro la violazione predittiva voluto, però l'interpretazione di questa regola in termini predittivi non è univoca. Ad esempio si può pensare di scegliere *m* e *k* per proteggere la violazione di un contributo contro una coalizione di (*m*-1) rispondenti (Cox, 1981). A nostro avviso, però, esistono altre forme di protezione più adeguate contro questo tipo di violazione, come si vedrà più avanti (Merola, 2003a, e 2003b). Si consideri invece il caso in cui un intruso estraneo alla cella voglia stimare la somma dei primi *m* contributi, $t_m$. In questo caso, la condizione che l'errore relativo di stima sia maggiore di *p* è $\frac{z - t_m}{t_m} > p$ ovvero $t_m - \frac{1}{1+p} z < 0$.

Confrontando quest'ultima espressione con quella definitoria della misura si ha $k = \frac{1}{1+p}$ (Merola, 2003a e 2003b). Nella Figura 2.1 è mostrato l'andamento di *p* rispetto a *k* e nella Tabella 2.4 sono mostrati i valori di *p* corrispondenti a diverse

scelte di *k*. Quindi, per esempio, per avere una protezione dell'errore relativo tra 0.25 e 1 occorrerà scegliere *k* tra 0.8 e 0.5.
Si consideri, come esempio, una cella con contributi (23, 19, 13, 10, 8, 5, 2) e quindi con totale di *z*=80. Se applicassimo la regola della dominanza con *m*=3 e *k*=0,8, la cella risulterebbe sicura, infatti $t_3 / z = 0.688$ ; l'errore relativo della stima $\hat{t}_3 = z$ è pari a 0.453, superiore al valore di *p*=0.25 implicato dalla scelta di *k*.

**Figura 2.1 Relazione tra rapporto di concentrazione *k* e protezione *p***

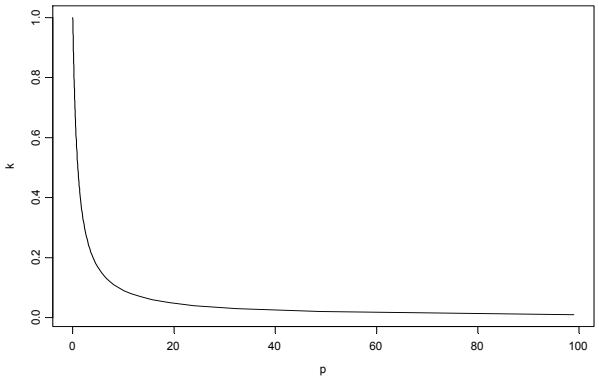



**Tabella 2.4 Valori di *p* corrispondenti a diverse scelte di *k***

| *k* | *p* |
|---|---|
| 0.1 | 9.00 |
| 0.2 | 4.00 |
| 0.3 | 2.33 |
| 0.4 | 1.50 |
| 0.5 | 1.00 |
| 0.6 | 0.67 |
| 0.7 | 0.43 |
| 0.8 | 0.25 |
| 0.9 | 0.11 |

## 3) Regola del rapporto

Questa misura protegge contro il rischio che un intruso possa stimare un contributo con errore relativo inferiore a *p*, con $0 \le p \le 1$. La misura viene costruita ipotizzando che uno dei rispondenti, diciamo il *k*-esimo, voglia stimare uno degli altri contributi con $z$-$z_k$. Per questo scenario il caso peggiore (minimo ER) è quello in cui il secondo rispondente vuole stimare il contributo massimo, $z_1$. In questo caso la stima sarà $\hat{z}_1 = z - z_2$ e l'errore relativo sarà maggiore di *p* se $\frac{\hat{z}_1 - z_1}{z_1} > p$.
Quindi la misura di rischio rispetto a questo criterio è:

$$S = pz_1 - \sum_{l=3}^{n} z_l = pz_1 - r_2 ,$$

per cui i pesi sono

$$\begin{cases} \lambda_1 = p \\ \lambda_2 = 0 \\ \lambda_j = -1 \quad j > 2. \end{cases}$$

Poiché i pesi sono non crescenti questa misura è sub-additiva. Considerando l'esempio precedente dei 6 contributi (23, 19, 13, 10, 8, 5, 2), la cella risulterebbe

sicura per valori di $p$ < 1.65, in quanto risulta $\hat{z}_1 = 61$ con $\mathrm{ER} = (\hat{z}_1 - z_1)/z_1 = 1.65$. La regola del rapporto può essere estesa a coalizioni di ($m$-1) rispondenti prendendo per pesi

$$\begin{cases} \lambda_1 = p \\ \lambda_j = 0 \quad 2 \le j \le m \\ \lambda_j = -1 \quad j > m \end{cases}$$

In questo caso la scelta di $m$ deve corrispondere ad una precisa esigenza, in quanto l'ER decresce con $m$ fino al limite 0 per $m$=$n$.

**4) Regola priori-posteriori**

Questa regola, anche conosciuta come regola ($p$,$q$) (($p$,$q$)-rule), è analoga alla precedente ma presuppone che uno qualsiasi dei rispondenti della cella, poniamo il $k$-esimo, prima della pubblicazione di $z$ possa stimare gli altri con una precisione $q>0$ (cioè che conosca $(1 \pm q)z_j$, per ogni $j \neq k$). Anche in questo caso ci si vuole proteggere dal rischio che l'intruso possa usare $z$ per stimare un altro contributo con ER inferiore a $p$, dove, naturalmente, $p < q$.
Per questo scenario il caso peggiore è sempre quello in cui il secondo rispondente vuole stimare $z_1$. Quindi, prendendo $\hat{z}_1 = z - z_2 - (1 \pm q)\sum_{j=3}^{n} z_j = z_1 \pm q\, r_2$ come miglior stima di $z_1$, la cella non è sensibile se $\frac{|\hat{z}_1 - z_1|}{z_1} \ge p$. Quindi, si ottiene la misura di rischio:

$$\mathrm{S} = pz_1 - q\sum_{j=3}^{n} z_j = pz_1 - qr_2,\ q > p,$$

perciò, i pesi per questa misura sono

$$\begin{cases} \lambda_1 = p \\ \lambda_2 = 0 \\ \lambda_j = -q \quad j > 2 \end{cases}$$

e quindi la misura è sub-additiva ed estendibile a coalizioni di ($m$-1) rispondenti. La misura del rapporto è un caso particolare della regola ($p$,$q$) con $q$=1.
Come esempio si consideri una cella con contributi (20, 17, 7, 5, 2) e $q$=0.5. In questo caso risulta $\hat{z}_1 = 59 - 17 - 0.5 \times 14 = 27$ (13 se sottostima) e, quindi, la cella risulterebbe sicura per ogni $p$<0.35 in quanto $\mathrm{ER} = \frac{|\hat{z}_1 - z_1|}{z_1} = 0.35$.
In Tabella 2.5 sono riassunte le varie misure di rischio descritte sopra.

**Tabella 2.5 Dati riassuntivi di alcune misure di rischio lineari subadditive**

| Nome | Misura | Protezione | Parametri | Vincoli |
|---|---|---|---|---|
| Soglia | $m$-$n$ | | $k$ intero | $k \geq 3$ |
| Dominanza | $(1-k)t_m - kr_m$ | $\mathrm{ER}(t_m) > (1-k)/k$ | $0< k < 1,\ m < n$ | $n > \lfloor m/k \rfloor$ |
| Del rapporto | $pz_1 - r_2$ | $\mathrm{ER}(z_1) > p$ | $p > 0$ | |
| Priori-posteriori | $pz_1 - qr_2$ | $\mathrm{ER}(z_1) > p$ | $p,q > 0$ | $q > p$ |

## Commenti e critiche

Le misure di rischio lineari costituiscono uno strumento di facile implementazione ed interpretazione. Inoltre la subadditività garantisce che l'aggregazione di celle non aumenti il rischio di violazione. Spesso, però, la soggettività della scelta dei parametri rende difficile stabilire una linea di demarcazione netta tra celle a rischio e non. Si può dimostrare che se una cella è sicura rispetto alla dominanza con parametri $k$ e $m$>1, essa è sicura rispetto alla dominanza con lo stesso parametro $k$ per tutti gli $m'$<$m$ e per la regola del rapporto con $p \geq (1-k)/k$. Quindi la regola della dominanza può essere applicata per garantirsi rispetto alle altre scegliendo opportunamente i valori dei parametri $m$ e $k$. Si consideri, comunque, che il parametro $k$ dovrebbe essere scelto anche in funzione del numero di contributi considerati (così come si fa negli indici di concentrazione più comunemente utilizzati, come quello del Gini e dello Spearman).

Le misure di rischio discusse non coprono tutti i possibili rischi di violazione; anzi queste misure sono ottenute considerando tentativi di intrusione piuttosto semplici. Inoltre, anche quando un criterio di rischio è accettato, la conoscenza del numero di rispondenti della cella e l'esistenza di altre variabili note, che diano un'indicazione sul rango dei rispondenti o del valore della variabile da proteggere, dovrebbero essere tenute in debito conto per valutare i rischi di violazione di una cella.

Di seguito elenchiamo alcuni problemi relativi all'uso delle misure di rischio lineari.

1. La regola del rapporto non garantisce che un rispondente non possa comunque stimare un contributo con ER<p. Per esempio, si consideri una cella con contributi (15, 12, 11, 2, 1). Questa cella è sicura rispetto alla misura 100p% per tutti i valori di p < 0.933 (quindi molto sicura). Supponiamo ora che il terzo rispondente conosca il proprio rango e che voglia stimare z1 e z2 senza conoscerne la distribuzione. Indicando con ${}^k\hat{z}_j$ la stima del j-esimo contributo da parte del k-esimo rispondente, le stime che minimizzano l'errore quadratico medio (Merola, 2003a, e 2003b) sono:

$$ {}^3\hat{z}_1 = {}^3\hat{z}_2 = \frac{(z - z_3)}{2} = \frac{1}{2}(z_1 + z_2 + r_3) = 15 $$

Quindi il primo contributo è perfettamente violabile mentre il secondo può essere stimato con ER=0.25. Questo esempio, volutamente estremo, mostra come non sempre le misure discusse sopra forniscano un'adeguata protezione. Risulta però che la misura di rischio per questa violazione non è subadditiva e perciò di non facile trattazione.

2. La misura priori-posteriori è molto sensibile ai valori di $p$ e $q$. E' possibile che una cella risulti sicura per dei valori di questi parametri e non sicura per valori molto simili. In realtà questa misura è equivalente alla misura del rapporto con soglia minima di errore relativo $p' = p/|q| < 1$. Quest'ultima misura sembra offrire maggiore interpretabilità ai parametri.
3. Nella misura priori-posteriori la conoscenza *a priori* di $(1 \pm q)z_1$ non viene utilizzata, inoltre risulta che l'ER è maggiore di $p$ ogni qual volta $z_1 < r_2$ a prescindere dai valori di $p$ e $q$.
4. In alcuni casi è più probabile che un intruso abbia una idea (stima *a priori*) del rapporto di grandezza tra contributi piuttosto che del loro valore.

L'esempio al punto 1. è un caso particolare di violazione da parte di un rispondente che conosce il proprio rango. In alcuni casi è verosimile ipotizzare che un tale intruso possa anche avere un'idea sulla distribuzione di alcuni degli altri contributi e che quindi possa ottenere stime puntuali abbastanza precise di essi. Il calcolo del rischio di violazione per scenari di predizione complessi è ancora oggetto di studio.

## 2.6 Tabelle campionarie

In alcuni casi, le tabelle pubblicate contengono dati stimati da un campione, per esempio stratificato. Per le celle contenenti dati ottenuti campionando uno strato al 100 per cento si devono applicare le regole riportate sopra. Invece, il rischio di violazione per le celle che riportano dati stimati sarà nullo, o comunque basso, sempre che non si rilasci troppa informazione sul piano campionario e sulle unità campionate. Per esempio, supponiamo che il valore di una cella sia stato ottenuto da un campione di due unità da uno strato con dieci unità. Il totale $z$ sarà stato ottenuto come $5z_s$, dove $z_s$ è il totale campionario. Entrambe le unità campionate possono ottenere il valore dell'altra unità che però, se non si è rivelato quale essa sia, non può essere identificata tra le altre nove della popolazione. Oltretutto, poiché il vero totale della popolazione non è noto un intruso non potrà stabilire un limite massimo per gli altri contributi. Quindi il rischio per questa cella si può considerare nullo. Comunque, se nel campione sono presenti valori estremi della popolazione anche il valore campionario può comportare rischio. Quindi, spesso, si sente l'esigenza di una regola automatica che classifichi una cella come "a rischio" o meno. Per le tabelle di intensità è possibile adattare la regola della dominanza a celle ottenute da valori campionari: occorre rimpiazzare il valore totale della cella e la somma dei primi $m$ contributi con le loro stime. Denotando con $w_j$ i pesi campionari dei contributi campionati, $z_j$, nella cella C, la stima del totale della cella (valore rilasciato) è $\hat{z} = \sum_{j \in C} w_j z_j$. Come stima della somma dei primi $m$ contributi si può prendere il valore:[6]

$$\hat{t}_m = \sum_{j=1}^{J-1} w_j z_j + \beta w_J z_J$$

[6] Metodo attribuito ad A. Hundepool in Willenborg e de Waal, 2001

dove $J$ e $\beta$ sono tali che $0 \le \beta < 1$ e $\sum_{j=1}^{J-1} w_j + \beta w_J = m$. Per esempio si consideri di voler applicare la misura della dominanza con $m$=3 e $k$=0.7 ad una cella con contributi come in Tabella 2.6.

**Tabella 2.6 Esempio di tabella con dati campionari**

| Unità | 1 | 2 | 3 | 4 | 5 | Totale |
|---|---|---|---|---|---|---|
| Contributi | 40 | 16 | 6 | 4 | 4 | |
| Pesi | 2.5 | 1.5 | 1 | 2 | 1 | 8 |
| Contributi ponderati | 100 | 24 | 6 | 8 | 4 | 150 |

In questo caso $\hat{z} = 150$ (con $\hat{n} = 8$). Per la stima di $t_3$, $\beta$ si ottiene come soluzione di $2.5 + 1.5\beta = 3$ e quindi risulta $\beta = 1/3$. Perciò, si ha $\hat{t}_3 = 2.5 * 40 + 0.5 * 16 = 108$. Poiché 108 è il 72 per cento di 150, la cella sarà dichiarata a rischio secondo questa regola. Analoghi metodi empirici possono essere usati per l'applicazione di altre misure di rischio.

Naturalmente, nel caso di dati campionari si può seguire un approccio probabilistico per la valutazione del rischio di identificazione.

## 2.7 Tabelle personalizzate

A volte capita che degli utenti richiedano delle tabelle personalizzate, cioè ricavate da una tabella solo per unità di una sottopopolazione specificata. In questo si può ipotizzare che un richiedente malintenzionato possieda già informazioni su alcuni contributi della sottopopolazione richiesta e sia interessato a stimare solo gli altri contributi. Quando questo è possibile, si può rilasciare una tabella per una sottopopolazione più ampia di quella richiesta, in cui le unità siano state aggiunte in modo che ogni singolo contributo possa essere stimato con un errore relativo maggiore di una certa soglia, ipotizzando che il richiedente conosca tutti i valori richiesti meno uno. Ad esempio, se il massimo dei contributi della sottopopolazione richiesta è $z_{\max}$ allora, per ottenere un livello di sicurezza $p$, il totale delle unità aggiunte, diciamo $z_a$, dovrà essere tale che $z_a / z_{\max} > p$.

## 2.8 Tabelle collegate

Il problema delle tabelle collegate sorge quando, a partire dal medesimo file di dati elementari, sono rilasciate più tabelle che hanno variabili classificatrici in comune per la stessa variabile sensibile. Il caso più comune è quello in cui vengano rilasciate più tabelle bidimensionali, ognuna con una variabile classificatrice in comune con un'altra. Altri casi sono quelli di tabelle in cui le variabili classificatrici in comune sono categorizzate in maniera differente. In quest'ultima casistica rientrano le tabelle gerarchiche, cioè tabelle annidate in cui le classi di una sono sottoclassi dell'altra (è il caso di tabelle per classificazione Ateco a diverse cifre).

La misura del rischio di violazione delle celle di tabelle collegate non è stata

compiutamente risolta, anche per la complessità e diversità dei casi possibili. Di seguito si riporteranno alcuni esempi esplicativi dell'approccio al problema.

Il caso delle tabelle collegate può essere inquadrato come un caso in cui vengano rilasciati solo i totali marginali di una tabella multidimensionale (supertabella). Per valutare il rischio di violazione è necessario determinare se e quali celle della supertabella possono essere ricostruite o, comunque, definite in un intervallo e se sulla base di questi risultati c'è rischio di violazione o no. Si consideri il caso di tre variabili classificatrici ($x_1$,$x_2$,$x_3$) e si supponga che siano rilasciate due tabelle, $T_1$ e $T_2$, la prima definita dalle variabili classificatrici ($x_1$,$x_2$) e la seconda da ($x_1$,$x_3$), quindi con la variabile classificatrice $x_1$ in comune. La tabella $T_3$, definita da ($x_2$,$x_3$), invece, non viene rilasciata. La conoscenza di $T_1$ e $T_2$ comporta la conoscenza delle marginali di $T_3$. Un intruso che voglia ricostruire $T_3$ deve, perciò, ricostruire una tabella dalle marginali, cosa che, come è noto, non è in genere possibile fare esattamente; l'intruso potrà però stabilire dei margini inferiori e superiori per il valore delle celle.

Si considerino per esempio le Tabelle 2.7 e 2.8 di seguito. Queste sono due tabelle fittizie collegate che riportano il fatturato di imprese secondo: $x_1$=Settore (I=1, II=2), $x_2$=Dimensione(Piccola/Media=1, Grande=2) e $x_3$=Area geografica (Nord=1, Centro-Sud=2).

**Tabella 2.7 Prima tabella rilasciata, Fatturato per Settore e Dimensione**

| $T_1$ | Piccola/Media | Grande | Totale |
|---|---|---|---|
| Settore I | 190 | 60 | 250 |
| Settore II | 30 | 60 | 90 |
| Totale | 220 | 120 | 340 |

**Tabella 2.8 Seconda tabella rilasciata, Fatturato per Settore e Area geografica**

| $T_2$ | Nord | Centro-Sud | Totale |
|---|---|---|---|
| Settore I | 130 | 120 | 250 |
| Settore II | 30 | 60 | 90 |
| Totale | 160 | 180 | 340 |

La Tabella 2.9 riporta i dati della tabella del Fatturato per Dimensione ed Area geografica ricavabili dalle altre due, dove i valori incogniti sono indicati come $z_{.jk}$, usando il punto per denotare la sommatoria rispetto all'indice sostituito.

**Tabella 2.9 Dati deducibili direttamente dalle tabelle 2.7 e 2.8**

| $T_3$ | Nord | Centro-Sud | Totale |
|---|---|---|---|
| Piccola/Media | $z_{.11}$ | $z_{.12}$ | 220 |
| Grande | $z_{.21}$ | $z_{.22}$ | 120 |
| Totale | 160 | 180 | 340 |

L'intruso, sapendo che il fatturato è non negativo, può evincere degli intervalli per i valori incogniti in Tabella 2.9, osservando che questi devono soddisfare i seguenti vincoli:

$$z_{\bullet 1\bullet} = 220 = z_{\bullet 11} + z_{\bullet 12}$$
$$z_{\bullet 2\bullet} = 120 = z_{\bullet 21} + z_{\bullet 22}$$
$$z_{\bullet\bullet 1} = 160 = z_{\bullet 11} + z_{\bullet 21}$$
$$z_{\bullet\bullet 2} = 180 = z_{\bullet 12} + z_{\bullet 22}$$

Da questi vincoli è facile ricavare i seguenti intervalli di esistenza per i valori cercati:

$$40 \le z_{\bullet 11} \le 160$$
$$0 \le z_{\bullet 21} \le 120$$
$$60 \le z_{\bullet 12} \le 180$$
$$0 \le z_{\bullet 22} \le 120$$

La tabella originale con i dati cercati è riportata in Tabella 2.10 e, come si vede, i valori rientrano negli intervalli calcolati.

**Tabella 2.10 Tabella completa del Fatturato per Settore e Dimensione, non rilasciata**

| $T_4$ | Nord | Centro-Sud | Totale |
|---|---|---|---|
| Piccola/Media | 70 | 150 | 220 |
| Grande | 90 | 30 | 120 |
| Totale | 160 | 180 | 340 |

Si supponga ora che tutte e tre le tabelle siano state rilasciate perché non considerate a rischio. In questo caso l'intruso cercherà di ricostruire le singole celle $z_{ijk}$ o, almeno di individuarne un intervallo di esistenza, sfruttando i totali marginali e la non negatività dei valori. Si supponga, ad esempio, che l'intruso sia interessato al fatturato delle imprese dell'area I, di grande dimensione situate al centro-sud, cioè a $z_{122}$. Dunque porrà:

$$z_{12\bullet} = 60 = z_{121} + z_{122}\,.$$

Il primo termine dopo il segno di uguale può essere scritto come:

$$z_{121} = z_{\bullet 21} - z_{221} = 90 - z_{221}\,;$$

alla stessa maniera risolverà il secondo membro dell'equazione sopra come:

$$z_{221} = z_{2\bullet 1} - z_{211} = 30 - z_{221}\,.$$

Unendo le tre uguaglianze ottiene:

$$60 = 90 - 30 + z_{211} + z_{122}\,.$$

Per cui l'intruso può determinare con certezza che $z_{122} = z_{211} = 0$.

Con questi esempi si vuole mettere in evidenza che il rilascio di tabelle collegate, individualmente non a rischio, possa portare alla violazione di informazione per mezzo della loro unione. Le misure di rischio per le tabelle collegate si basano sull'ampiezza dell'intervallo di esistenza (*feasibility interval*) che deve rispettare alcune regole di

protezione. Questo argomento sarà affrontato di seguito, nel Capitolo 3. relativo ai metodi di protezione.

## 2.9 Tabelle di intensità con valori negativi

A volte le variabili riportate nelle tabelle di intensità possono assumere valori negativi, come per esempio il profitto delle imprese. In questo caso non è possibile applicare le misure di protezione esposte sopra. In pratica non esiste una regola generale per tutti i casi. Si consideri, comunque, che se è vero che la protezione di queste tabelle è più difficile che per le tabelle con contributi non negativi, anche la violazione lo è. Possibili procedure per celle con contributi negativi sono:

1. considerare i valori assoluti dei contributi ed applicare una delle misure di rischio viste;
2. dividere i contributi negativi da quelli positivi ed applicare una delle misure di rischio per ciascun gruppo (prendendo il valore assoluto di quelli negativi). Se uno dei due gruppi risulta a rischio, allora tutta la cella lo sarà;
3. sottrarre il valore del contributo minore (cioè aggiungere il valore assoluto) da tutti i contributi.

L'applicazione di queste regole, però, stravolge l'interpretazione delle misure di rischio applicate. Un altro metodo usato per queste tabelle è quello di applicare le misure di rischio ad una variabile a valori non negativi che sia correlata a quella da proteggere (una *proxy*). Tale variabile viene chiamata variabile ombra (*shadow*) e può anche essere usata per la successiva protezione. La procedura al punto 3. è un esempio di variabile ombra artificiale.

# Capitolo 3. La protezione statistica di tabelle[(*)]

Una volta individuate le celle a rischio occorre proteggere la tabella affinché sia possibile pubblicarla senza correre rischi di violazione. Un approccio possibile è quello di proteggere i microdati prima della creazione della tabella, ma non entreremo in questo merito (per una trattazione di tale argomento si veda Keller-McNulty e Unger, 1998): ci occuperemo solo di tecniche atte a proteggere le celle di tabelle derivate da microdati non modificabili.

Le tabelle sono tradizionalmente il "prodotto finale" rilasciato da un'indagine statistica e, pertanto, dovrebbero contenere il massimo di informazione possibile. Tutti i diversi metodi per la protezione delle tabelle esistenti comportano la perdita di una parte dell'informazione. L'informazione persa varia a seconda del livello di rischio accettato, stabilito dalla scelta della regola di sicurezza e dei suoi parametri, e del metodo di protezione scelto. Intuitivamente si può dire che minore è il rischio accettato e maggiore sarà la perdita di informazione. Quest'ultima, però, non è, e forse non può esserlo, univocamente definita. Infatti, oltre a diverse misure quantitative possibili, si possono concepire anche misure dell'informazione che si basino su diverse esigenze analitiche; si pensi, per esempio, a dover stabilire se la soppressione dei valori di poche celle di una tabella comporti maggiore perdita di informazione di una perturbazione di tutte le celle, che però garantisca la significatività dei parametri di un modello log-lineare. In questa sede ci limiteremo a considerare la misure di informazione più comunemente utilizzate nell'ambito della protezione delle tabelle, che consistono nell'assegnare dei pesi di contenuto informativo ad ogni cella che possono essere scelti secondo diversi criteri. Esiste anche un criterio basato sull'entropia (Willenborg e de Waal, 2001, pag. 167) ma, poiché risulta complesso ed in genere non porta a soluzioni praticabili, non verrà trattato. Si tenga presente che la scelta della misura dell'informazione determina anche il risultato della tecnica di protezione adottata. Infatti, il risultato della protezione, cioè la tabella da rilasciare, viene determinato scegliendo tra tutte le soluzioni possibili quella che minimizza l'informazione persa.

Nel prosieguo illustreremo prima vari metodi di protezione delle tabelle e poi le misure di perdita di informazione più utilizzate.

Le tecniche di protezione delle tabelle sono di due tipi: perturbative e non perturbative. Le protezioni perturbative prevedono la modificazione dei valori rilasciati mentre le tecniche non perturbative consistono nel riarrangiare le categorie delle tabelle oppure nel sopprimere il valore di alcune celle. Le tecniche di protezione di cui ci occuperemo sono:

- non perturbative: soppressione e riarrangiamento;
- perturbative: arrotondamento.

[(*)] Capitolo redatto da Giovanni M. Merola

## 3.1 Tecniche di protezione non perturbative: soppressione

La soppressione dei valori sensibili è la tecnica di protezione più usata. In principio questa consente di eliminare i valori sensibili rilasciando la maggior parte dei valori non a rischio inalterati. Un valore soppresso non è ricavabile direttamente se nella riga e nella colonna corrispondenti c'è almeno un altro valore soppresso. Quando la soppressione delle celle sensibili di una tabella non soddisfa questa condizione o se, come vedremo, la somma dei valori soppressi non è sufficientemente alta, occorre sopprimere altre celle non sensibili in modo che i valori sensibili non siano più ricavabili. La soppressione dei valori sensibili viene chiamata *primaria* e quella successiva *secondaria*. In genere esistono diversi schemi di soppressione secondaria possibili.

### 3.1.1 Soppressione secondaria marginale

Un tipo di soppressione secondaria molto efficace è quello in cui si sopprimono i totali marginali delle colonne e delle righe che contengono valori soppressi primariamente ed il totale generale. Questa procedura è correntemente utilizzata per il rilascio di tabelle dall'Istat. Questo tipo di soppressione elimina tutti i termini che contengono i valori sensibili (come addendi) ed è definita *soppressione marginale*. In Tabella 3.1 è mostrato un esempio di soppressione secondaria *marginale* per il caso in cui siano soppresse primariamente le celle delle importazioni dei fagioli per i paesi A e C.

**Tabella 3.1 Soppressione secondaria marginale. I valori soppressi primariamente sono stati sostituiti con il simbolo ♥ e quelli soppressi secondariamente con il simbolo ♠**

| Importazioni | Paesi | | | |
|---|---|---|---|---|
| | A | B | C | Totale |
| Ceci | 20 | 50 | 10 | 80 |
| Fagioli | ♥ | 19 | ♥ | ♠ |
| Fave | 17 | 32 | 12 | 61 |
| Totale | ♠ | 101 | ♠ | ♠ |

La soppressione secondaria marginale elimina la dipendenza dei valori soppressi dai valori rilasciati (cioè non è possibile calcolare nessun tipo di limite superiore per i valori soppressi), però è molto dispendiosa in termini di perdita di informazione. In molti casi si può raggiungere una protezione sufficiente sopprimendo secondariamente solo alcune celle interne. In questo caso però la soppressione secondaria deve essere scelta massimizzando la protezione e minimizzando una "perdita di informazione". Ci riferiremo a questo tipo di soppressione secondaria semplicemente come *soppressione secondaria.*

### 3.1.2 Soppressione secondaria

Si consideri come esempio la Tabella 3.2 in cui i valori delle importazioni di fagioli per i paesi A e C, indicati con $z_1$ e $z_2$ rispettivamente, sono considerati a rischio. Se fossero soppressi solo questi valori, si riotterrebbero immediatamente come differenza dei totali marginali dai valori non soppressi. Nell'esempio si avrebbe: $z_1 = 45 - 17 - 20 = 8$ e $z_2 = 44 - 10 - 12 = 22$.

**Tabella 3.2 Esempio di tabella da proteggere. I valori $z_1$ e $z_2$ sono considerati a rischio**

| Importazioni | Paesi | | | Totale |
|---|---|---|---|---|
| | A | B | C | |
| Ceci | 20 | 50 | 10 | 80 |
| Fagioli | $z_1$=8 | 19 | $z_2$=22 | 49 |
| Fave | $z_3$=17 | 32 | $z_4$=12 | 61 |
| Totale | 45 | 101 | 44 | 190 |

Naturalmente è possibile scegliere tra più soppressioni secondarie, tenendo conto che neanche queste devono essere ricavabili. Da una tabella con valori soppressi, in ogni riga e in ogni colonna è possibile ricavare il valore della somma dei valori soppressi. Confrontando questi valori è possibile ottenere degli intervalli, detti *intervalli di esistenza* (*feasibility intervals*) del tipo $[z_i - L_i, z_i + U_i]$, con $L_i, U_i \geq 0$, per ogni valore soppresso $z_i$. Uno schema di soppressione secondaria è ritenuto sufficientemente protettivo per una cella a rischio solo se l'ampiezza dell'intervallo di esistenza ricavabile è maggiore di una certa soglia o, altrimenti, se l'intervallo di esistenza contiene un *intervallo di protezione*. Le ampiezze minime e gli intervalli di protezione possono essere determinati secondo vari criteri, che si discuteranno più avanti.

Si supponga che nella Tabella 3.2 non si effettuino soppressioni secondarie di riga ma solo di colonna; una scelta possibile è quella di sopprimere i valori indicati con $z_3$ e $z_4$, come mostrato in Tabella 3.3.

**Tabella 3.3 Esempio di soppressione secondaria. I valori soppressi primariamente sono stati sostituiti con il simbolo ♥ e quelli soppressi secondariamente con il simbolo ♠**

| Importazioni | Paesi | | | Totale |
|---|---|---|---|---|
| | A | B | C | |
| Ceci | 20 | 50 | 10 | 80 |
| Fagioli | ♥ | 19 | ♥ | 49 |
| Fave | ♠ | 32 | ♠ | 61 |
| Totale | 45 | 101 | 44 | 190 |

In questo caso un intruso può calcolare che: (i) $z_1 + z_2 = 30$, (ii) $z_1 + z_3 = 25$, (iii) $z_2 + z_4 = 34$ e (iv) $z_3 + z_4 = 29$. Queste informazioni, insieme al vincolo di non negatività dei contributi, gli permettono poi di ricavarsi gli intervalli di esistenza. Per

esempio, da (ii) si ricava che $z_1 \leq 25$ e, quindi, da (i) segue che $z_2 \geq 5$. Date le soppressioni, gli intervalli di esistenza si ricavano come soluzioni di problemi di programmazione lineare.

Nell'esempio si possono ricavare i seguenti intervalli di esistenza:

$$\begin{aligned} 0 &\leq z_1 \leq 25 \\ 5 &\leq z_2 \leq 30 \\ 0 &\leq z_3 \leq 25 \\ 4 &\leq z_4 \leq 29 \end{aligned}$$

Si noti che, in questo esempio, l'ampiezza degli intervalli di esistenza è costante ed uguale al minimo dei vincoli ricavabili, ciò non è pero necessariamente vero in tutti i casi. Una volta determinati gli intervalli di esistenza ricavabili da uno schema di soppressione, occorre determinare se questi rispettano un criterio di sicurezza. Tra i criteri di sicurezza più comunemente utilizzati ricordiamo:

1. L'ampiezza dell'intervallo di esistenza dovrebbe essere almeno il 100p% del valore della cella. Questo metodo è molto semplice da applicare ma ha lo svantaggio di non considerare il livello di rischio della cella. Nell'esempio di Tabella 3.3, nel caso di $z_1$ risulta $25/8 = 3.125$, per $z_2$ risulta $25/22 = 1.14$ e quindi la soppressione secondaria in Tabella 3.3 sarebbe sufficiente per tutti i valori di $p \leq 1.14$.
2. L'ampiezza dell'intervallo di esistenza dovrebbe essere almeno il 100p% del contributo maggiore della cella. In questo caso si intendono proteggere i singoli contributi più che il valore della cella. L'ampiezza minima così determinata sarà minore di quella al punto 1. Supponiamo che $z_1 = 6+1+1$ e $z_2 = 12+6+2+1+1$, allora la proporzione dell'ampiezza degli intervalli di esistenza rispetto ai contributi maggiori sono $25/6 = 4,2$ per $z_1$ e 25/12=2.1 per $z_2$ e quindi soddisfano questo criterio per valori $p \leq 2.1$
3. L'ampiezza dell'intervallo deve essere tale che almeno il $d$% dei valori in esso contenuti siano sicuri rispetto alla misura di rischio adottata. Supponiamo che si voglia determinare l'intervallo di sicurezza per $z_2$ rispetto alla regola del rapporto con $p$=0.5. Indicando la generica stima di $z_2$ con $\hat{z}_2^U = 22+U$, l'errore relativo di stima di $z_2$ sarà minore di $p$ quando $U > 0.5 z_{2,1} - r_{2,2} = 6\text{-}4 = 2$. Se $d$=50 allora l'intervallo dovrà contenere il 50 per cento di punti che hanno distanza da $z_2$=22 superiore a 2. La proporzione dei punti dell'intervallo di esistenza determinato sopra che hanno distanza superiore a due dal valore $z_2$=22 è $(20-5)+(30-24)/25 = 0.84$, che essendo maggiore di 0.50 soddisfa la condizione.
4. Uno degli estremi dell'intervallo di esistenza deve soddisfare una regola di sicurezza. Nel caso della regola della dominanza dovrà essere $t_{m,i}/(z_i + U_i) \leq k$ e, quindi $t_{m,i}/k - z_i \leq U_i$. Supponiamo che si scelga $m$=2 e $k$=0.7, valori per cui $z_2 = 12+6+2+1+1$ in Tabella 3.3 non soddisfa la regola della dominanza. Allora, affinché la soppressione secondaria sia accettabile, l'estremo superiore deve risultare $U_2 \geq 18/0.7 - 22 = 3.71$, come è verificato per l'esempio. Nel caso si adottasse la regola priori-posteriori, deve essere verificato che

$pz_{1,i} - (z_i + U_i - z_{1,i} - z_{2,i} - (1+q)r_{2,i}) \leq 0$. Quindi l'estremo superiore deve soddisfare $U_i \geq pz_{1,i} + qr_{2,i}$. Nel caso della regola priori-posteriori con $p$=0.25 e $q$=0.6 (cosicché risulta $p/q \simeq (1-0.7)/0.7 \simeq 0.43$), l'estremo superiore ricavabile è $U_2 \geq 0.25*12 + 0.6*4 = 5.4$, che, anche in questo caso, è soddisfatta.

Ovviamente, gli stessi esempi avrebbero potuto essere fatti rispetto al limite di esistenza inferiore, *L*. I criteri ai punti 3. e 4. sono basati su regole di sicurezza per le celle e, pertanto, sembrano preferibili agli altri. Gli intervalli di sicurezza per un valore $z_i$ possono essere rappresentati con i valori $LPL_i$, $UPL_i$ e $SPL_i$, rispettivamente livello di protezione inferiore, livello di protezione superiore e livello di protezione scorrevole (*sliding protection level*), che è l'ampiezza dell'intervallo. Uno schema di soppressione è ammissibile se gli intervalli di esistenza, $[z_i - L_i, z_i + U_i]$ soddisfano: $L_i \leq LPL_i$, $U_i \geq UPL_i$ *e* $|U_i - L_i| \geq SPL_i$. Per le regole in cui è sufficiente che $SPL_i > K > 0$, $LPL_i$ e $UPL_i$ sono fissati a $\pm\infty$.

Nella maggior parte dei casi esisteranno più schemi di soppressione secondaria che soddisfano il criterio di protezione scelto. Allora la soppressione secondaria viene scelta in modo di minimizzare la perdita di informazione, specificata da pesi assegnati alle celle.

La scelta dei pesi per l'informazione per le celle non sensibili può essere fatta in base a diversi criteri. Si consideri l'esempio della Tabella 3.2: se i valori delle importazioni dei ceci fossero ritenuti meno informativi di quelli delle importazioni di fave, si potrebbero assegnare pesi maggiori alle celle corrispondenti a quest'ultimo legume. Qualora la soppressione dei valori delle importazioni dei ceci per A e C portassero a degli intervalli di esistenza sicuri, questa soppressione verrebbe preferita a quella in Tabella 3.3. I pesi però possono essere adottati rispetto a criteri oggettivi, i più comuni dei quali sono elencati di seguito:

- *pesi uguali per ogni cella*
  Con questa scelta si minimizza il numero delle celle soppresse. Spesso ammette più soluzioni e quindi una può essere individuata rapidamente anche con metodi euristici. I totali marginali devono avere pesi maggiori o essere esclusi dalle possibili soppressioni secondarie.
- *pesi uguali ai valori delle celle*
  Questo sistema di pesi dovrebbe essere adottato solo se le celle assumono valori positivi e se le celle con bassi valori non hanno particolare rilievo, in quanto saranno queste quelle preferite per la soppressione secondaria.
- *pesi uguali al numero dei contributi*
  Questo sistema penalizza le celle con bassa frequenza minimizzando il numero totale dei contributi soppressi. Raramente porta alla soppressione di totali marginali e può essere applicato alle tabelle di frequenza.
- *principio della minima sicurezza*[7]
  Avendo adottato una regola di sicurezza per la misura del rischio, le celle ritenute

[7] Msp: Minimum Safety Principle

sicure possono essere ponderate con pesi che siano funzione del proprio rischio. Per esempio, nel caso della dominanza, ogni cella non sensibile avrà peso $w = 1 - t_m / z$, cosicché le celle con concentrazione maggiore avranno peso minore e saranno le prime candidate ad essere soppresse. Se si vuole considerare anche il valore delle celle si potrà porre $w = z - t_m = r_m$, cosicché le celle con minore somma degli ultimi *m* contributi avranno peso minore. In modo analogo si possono assegnare pesi alle celle rispetto ad altre regole di sicurezza, come quella del rapporto o quella priori-posteriori. In questo caso i pesi saranno tanto minori quanto minore è il minimo errore relativo.

La scelta di pesi uguali per tutte le celle sembra essere quella più comune e, per così dire, quella più "democratica". Questi pesi comportano la minimizzazione del numero delle soppressioni secondarie. L'attribuzione dei pesi secondo il principio della minima sicurezza è quella metodologicamente più coerente; si tenga presente, però, che con tale scelta è difficile predire il risultato della soppressione secondaria.

Ad ogni modo, la scelta dei pesi fatta in base ad uno di questi principi può sempre essere aggiornata modificando i pesi di celle che non si vuole che siano soppresse o che, viceversa, si preferisce che vengano soppresse.

Una volta assegnati i pesi alle celle, la scelta ottimale della soppressione secondaria è un problema di programmazione lineare che, specie per tabelle di grandi dimensioni, può essere molto dispendiosa in termini computazionali. I metodi per la determinazione della soppressione secondaria ottimale saranno trattati più avanti.

La protezione delle tabelle con la soppressione non è esente da inconvenienti. Ad esempio la soppressione di due celle nella stessa riga potrebbe non essere una protezione adeguata per i singoli contributi, anche se l'intervallo di esistenza per la cella sensibile è sufficientemente ampio. Infatti, la cella unione delle due celle soppresse, di cui si conosce il totale, non è necessariamente sicura. Si consideri il caso in cui siano state soppresse le celle $C_1$: $z_1$=32+4+4+2+1=43 e $C_4$: $z_4$=5+1+1+1+1=9 perché la cella $C_1$ è sensibile rispetto al rischio priori-posteriori con $p$=0.25 e $q$=0.5. Il totale della cella $C=(C_1 \cup C_4)$ è $z$=32+5+4+4+2+1+1+1+1+1=52 per cui la misura di sensibilità per la regola detta è $S = 0,25 * 32 - 0,5(4 + 4 + 2 + 1 + 1 + 1 + 1 + 1) = 0,5 > 0$ e quindi C è a rischio.

In alcuni casi anche più di due soppressioni per riga o colonna potrebbero non essere sufficienti. Si consideri la tabella del fatturato delle imprese produttrici di strumenti musicali riportata in Tabella 3.4 in cui i due valori sensibili sono in neretto.

**Tabella 3.4 Fatturato imprese produttrici di strumenti musicali. I valori a rischio sono indicati in neretto**

| Strumenti | Regione | | | | Totale |
|---|---|---|---|---|---|
| | A | B | C | D | |
| Arpe | 58 | **47** | 36 | 89 | 230 |
| Organi | 71 | 124 | 24 | 31 | 250 |
| Pianoforti | 92 | 157 | 59 | **28** | 454 |
| Altro | 800 | 934 | 651 | 742 | 3127 |
| Totale | 1021 | 1262 | 770 | 890 | 4061 |

Si supponga che la soppressione secondaria scelta abbia portato al rilascio della Tabella 3.5.

**Tabella 3.5 Fatturato imprese produttrici di strumenti musicali**

| Strumenti | Regione | | | | |
|---|---|---|---|---|---|
| | A | B | C | D | Totale |
| Arpe | $x_{11}$ | $\mathbf{x_{12}}$ | $x_{13}$ | 89 | 230 |
| Organi | $x_{21}$ | 124 | $x_{23}$ | 31 | 250 |
| Pianoforti | 92 | $x_{32}$ | 59 | $\mathbf{x_{34}}$ | 454 |
| Altro | 800 | $x_{42}$ | 651 | $x_{44}$ | 3127 |
| Totale | 1021 | 1262 | 770 | 890 | 4061 |

Questa soppressione secondaria non è efficace in quanto il valore sensibile dei produttori di arpe nella regione B è ricavabile. Infatti, si ha: $x_{11} + x_{21} + x_{13} + x_{23} = 129 + 60 = 189$, $x_{11} + \boldsymbol{x}_{12} + x_{13} + x_{21} + x_{23} = 141 + 95 = 236$ da cui si ricava esattamente $\boldsymbol{x}_{12} = 236 - 189 = 47$.

Anche la conoscenza della regola di sicurezza adottata può ridurre l'efficacia della protezione. Si supponga che la cella relativa al fatturato di alcune imprese sia stata soppressa perché non soddisfa la regola della dominanza con $m$=5 e $k$=0.8. Se l'impresa A sa di essere il rispondente di questa cella con il maggiore contributo, pari a 8, e se conosce la regola di sicurezza utilizzata, allora sa che $(8 + z_2 + z_3 + z_4 + z_5)/z \geq 0.8$. Inoltre, poiché tutti i contributi devono essere non maggiori di 8, l'impresa A può ricavare che $z \leq 50$. Questo esempio, che è solo uno dei tanti possibili, dovrebbe mettere in guardia contro la diffusione dei parametri adottati per la protezione delle tabelle. La regola generale è che si dovrebbe sempre dichiarare se una tabella è stata protetta o meno ma non si dovrebbero dichiarare i parametri delle regole di sicurezza adottati.

### 3.1.3 Soppressione parziale

Per quanto visto sopra, la pubblicazione di una tabella protetta con soppressione secondaria equivale alla pubblicazione degli intervalli di esistenza per i valori soppressi. Quindi Tabella 3.3 è equivalente alla Tabella 3.6.

Partendo da questo punto di vista, Fischetti e Salazar (2002) propongono di proteggere le tabelle con la "soppressione parziale", anziché con la soppressione secondaria, cioè rilasciando degli intervalli di esistenza per i valori altrimenti soppressi con la soppressione secondaria. Quindi, nella soppressione parziale le celle da "sopprimere parzialmente" e gli intervalli da sostituire ai veri valori vengono determinati in modo ottimale. Tabella 3.7 è un esempio di soppressione parziale, comparandola con Tabella 3.6 si nota come le celle modificate siano diverse e l'ampiezza degli intervalli rilasciati sia notevolmente inferiore.

Per l'applicazione della soppressione parziale si assume che per ogni valore di cella $z_i$ siano stati fissati gli intervalli di sicurezza, rappresentati con i valori $LPL_i$, $UPL_i$ e $SPL_i$. Uno schema di soppressione parziale, quindi, consiste dall'insieme degli indici, $SUP$, delle celle da modificare e dagli estremi degli intervalli di esistenza, $L_i$ e $U_i$. La

perdita di informazione viene definita come $\sum_{i \in SUP} w_i \left| U_i + L_i \right|$, dove $w_i$ è la perdita unitaria per ogni cella, e lo schema di soppressione parziale ottimale è quello che, tra gli schemi ammissibili, che minimizza la perdita di informazione.

**Tabella 3.6 Esempio di soppressione secondaria. I valori soppressi sono stati sostituiti con i corrispondenti intervalli di esistenza**

| Importazioni | Paesi | | | Totale |
|---|---|---|---|---|
| | A | B | C | |
| Ceci | 20 | 50 | 10 | 80 |
| Fagioli | [0,25] | 19 | [5,30] | 49 |
| Fave | [0,25] | 32 | [4,29] | 61 |
| Totale | 45 | 101 | 44 | 190 |

**Tabella 3.7 Esempio di soppressione parziale. I valori sensibili ed altri sono stati sostituiti con intervalli di esistenza**

| Importazioni | Paesi | | | Totale |
|---|---|---|---|---|
| | A | B | C | |
| Ceci | [18,24] | 50 | [6,12] | 80 |
| Fagioli | [4,10] | 19 | [20,26] | 49 |
| Fave | 17 | 32 | 12 | 61 |
| Totale | 45 | 101 | 44 | 190 |

La soppressione parziale costituisce una generazione continua della soppressione secondaria, rispetto alla quale presenta diversi vantaggi; in quanto:

- la perdita di informazione è ottimizzata in modo continuo e quindi le tabelle rilasciate saranno più informative di quelle soppresse secondariamente;
- la soluzione ottimale può essere calcolata molto più efficientemente;
- costituisce un metodo molto flessibile con cui l'effettiva protezione dei valori sensibili può essere controllata.

Maggiori dettagli su questo metodo saranno dati nella sezione sui modelli di soppressione.

### 3.1.4 Aspetti computazionali della soppressione

La determinazione dello schema di soppressione ottimo è un problema di programmazione lineare abbastanza complesso. Una volta individuate le celle a rischio, occorre determinare la soppressione complementare "migliore" tra tutte quelle ammissibili, cioè che garantiscono l'inviolabilità delle celle. Per fare ciò sono stati sviluppati degli approcci che ora analizzeremo brevemente.

### 3.1.4.1 Soppressione secondaria ottimale

L'individuazione dello schema di soppressione secondaria ottimale, cioè quello che, tra tutte le soppressioni secondarie che danno intervalli di esistenza sufficientemente ampi (rispetto ad uno dei criteri visti sopra), minimizza l'informazione persa, è un problema di programmazione lineare di tipo fortemente *NP*-completo. In realtà, gli algoritmi che forniscono soluzioni ottimali funzionano solo per tabelle con meno di tre dimensioni, sono lenti e di difficile implementazione (si veda, ad esempio, de Wolf, 2001 e Willenborg e de Waal, 2001). Sono stati perciò sviluppati diversi algoritmi euristici computazionalmente più efficienti, che però risultano subottimali. Tra questi ricordiamo quelli proposti da Cox (1995a), Sande (1984), Kelly *et al.* (1992) e Carvalho *et al.* (1994).

Un altro approccio al problema della scelta della soppressione secondaria ottimale è quello del Mip (*mixed integer programming*), proposto per primo da Kelly (1990). In questo approccio la tabella data è definita da un vettore $\boldsymbol{z}$, i cui elementi sono i valori delle celle e i totali marginali, e dai vincoli di additività. Questi vincoli sono espressi attraverso la matrice $\boldsymbol{M}$ ed il vettore dei totali marginali, $\boldsymbol{b}$, come $\boldsymbol{Mz} = \boldsymbol{b}$. La matrice $\boldsymbol{M}$ in genere ha valori in $\{-1, 0, 1\}$ e $\boldsymbol{b}$ è composto di soli zero. Per la Tabella 3.2, il vettore $\boldsymbol{z}$ può essere ottenuto dalla sequenza delle righe, per cui la relazione $\boldsymbol{Mz} = \boldsymbol{b}$ è definita da:

$$\boldsymbol{z} = (20,\ 50,\ 10,\ 8,\ 19,\ 22,\ 17,\ 32,\ 12,\ 80,\ 49,\ 61,\ 45,\ 101,\ 44,\ 190)'$$

$$\boldsymbol{M} = \begin{pmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & -1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1
\end{pmatrix}$$

$$\boldsymbol{b} = (0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0)'$$

Le prime tre righe della matrice $\boldsymbol{M}$ esprimono i vincoli per gli elementi delle righe, le seconde tre i vincoli per gli elementi delle colonne, le due successive i vincoli per i totali marginali di riga e di colonna e l'ultima il vincolo per il totale generale. La matrice $\boldsymbol{M}$ può essere espressa anche in altre forme, si noti, inoltre, come alcuni vincoli siano ridondanti. Questa rappresentazione può essere generalizzata a tabelle a più di due dimensioni e collegate.

Nel Mip l'idea della soppressione secondaria viene esplicitata considerando che i valori soppressi possono assumere un valore qualsiasi nell'intervallo di esistenza. L'insieme delle tabelle così ottenute è perciò l'insieme dei vettori $\boldsymbol{y}$ che soddisfano i vincoli:

$$\begin{cases} \boldsymbol{My} = \boldsymbol{b} \\ l_i \le y_i \le u_i \quad i = 1,...,n, \end{cases}$$

dove $l_i$ e $u_i$ sono i limiti *pubblici*, per esempio, $l_i$=0 e $u_i = \infty$. Uno schema di soppressione è così generalizzato a tutti i possibili valori di cella e marginali che sono consistenti con quella data; la tabella originale $\boldsymbol{y} = \boldsymbol{z}$ è una delle possibili soluzioni. Dato l'insieme degli indici delle tabelle sensibili, $S = \{i_1,...,i_p\}$, uno schema di soppressione completo, $SUP \supseteq S$, è l'insieme degli indici dei valori che saranno soppressi. Le tabelle consistenti con questo *SUP* sono definite da:

$$\begin{cases} \boldsymbol{My} = \boldsymbol{b} \\ l_i \le y_i \le u_i \quad i \in SUP \\ y_i = a_i \quad i \notin SUP. \end{cases}$$

Quindi, per i valori non soppressi $l_i = u_i = Z_i$. Per ogni *SUP* gli intervalli di esistenza per i valori sensibili, $[y_{i_k}^{-}, y_{i_k}^{+}]$, $i_k \in S$, vengono determinati da

$$y_{i_k}^{-} = \min y_{i_k}$$
$$\begin{cases} \boldsymbol{My} = \boldsymbol{b} \\ l_i \le y_i \le u_i \quad i \in SUP \\ y_i = a_i \quad i \notin SUP \end{cases}$$

e

$$y_{i_k}^{+} = \max y_{i_k}$$
$$\begin{cases} \boldsymbol{My} = \boldsymbol{b} \\ l_i \le y_i \le u_i \quad i \in SUP \\ y_i = a_i \quad i \notin SUP. \end{cases}$$

Un *SUP* è ammissibile se per ogni $i_k \in S$ gli intervalli di esistenza soddisfano le condizioni richieste per la protezione. Cioè se, dati i valori $LPL_i$, $UPL_i$ e $SPL_i$ definiti nella sezione della soppressione, sono verificate le condizioni: $y_{i_k}^{-} \le a_{i_i} - LPL_{i_k}$, $y_{i_k}^{+} \ge a_{i_i} + UPL_{i_k}$ e $y_{i_k}^{+} - y_{i_k}^{-} \ge SPL_{i_k}$ per ogni $i_k \in S$.

Il *SUP* ottimo viene determinato tra quelli ammissibili minimizzando la perdita $\sum_{i \in SUP} w_i$ Ciò viene fatto associando ad ogni cella una variabile binaria che indica lo stato della cella (1 se soppressa e 0 altrimenti), per cui la perdita diventa $\sum_{\forall i} w_i x_i$. Le variabili $x_i$ vengono poi inglobate anche nel sistema di vincoli sugli intervalli di esistenza. La soluzione del problema della soppressione secondaria richiede, pertanto, la soluzione di un numero molto grande di equazioni. Fischetti e Salazar (1999) hanno proposto un algoritmo efficiente per tabelle bidimensionali. Basato sulla soluzione del problema duale, e attraverso l'uso di tecniche euristiche di *branch-and-cut*, l'algoritmo porta a soluzioni di "provata ottimalità" in tempo polinomiale. Tale approccio è poi stato generalizzato (Fischetti e Salazar, 2000) a tabelle di dimensioni realistiche anche a

quattro dimensioni e a tabelle collegate e gerarchiche. Questo algoritmo è stato implementato anche nel software τ-Argus.

Il metodo dell'ipercubo è l'approccio euristico alla soppressione secondaria maggiormente utilizzato ed è implementato anche in τ-Argus. In questo approccio, vengono considerati solo gli schemi di soppressione che formano un ipercubo con la cella sensibile ad un vertice e con tutti gli altri vertici maggiori di zero. In due dimensioni un ipercubo è un rettangolo e, quindi, per ogni soppressione primaria si considerano tutti gli schemi che sopprimono una cella nella stessa riga, una nella stessa colonna e l'altra all'intersezione di queste due; in più dimensioni, i vertici dell'ipercubo contigui alla cella sensibile devono essere nella stessa direzione assiale. Tra tutti gli ipercubi di soppressione che danno intervalli di esistenza accettabili viene poi selezionato quello che minimizza la funzione di perdita specificata. L'efficienza del metodo è dovuta al numero ridotto di soluzioni possibili ed alla semplificazione del calcolo degli intervalli di esistenza ottenuti, dovuta allo schema ipercubico. Il processo viene poi ripetuto per tutte le celle sensibili. Lo svantaggio di questo metodo rispetto agli algoritmi Mip è che le soppressioni secondarie sono calcolate una alla volta e, quindi, le soluzioni finali possono contenere un numero di soppressioni secondarie maggiore del necessario. I vantaggi sono la maggior efficienza computazionale, l'interpretabilità dell'algoritmo e l'applicabilità a tabelle di qualsiasi dimensione.

### 3.1.4.2 Soppressione parziale

La soppressione parziale consiste nel pubblicare intervalli del tipo $[z_i - L_i, z_i + U_i]$ per alcuni valori, anziché sopprimerli come nella soppressione secondaria ottimale. Per cui, le variabili binarie $x_i$, definite nella sezione precedente, vengono eliminate e sono sostituite con le variabili continue $z_i^-$ e $z_i^+$; le tabelle consistenti con quella data vengono definite dal vettore $\boldsymbol{y} = \left\{ y_i \in [z_i - z_i^-, z_i + z_i^+] \right\}$. Dato uno schema di soppressione parziale, $S \subseteq \{1,2,\ldots,n\}$, la perdita dovuta alla soppressione parziale è definita come $\sum_{i \in S} w_i \left| z_i^- + z_i^+ \right|$ ed è minimizzata sotto i vincoli:

$$\begin{cases} \boldsymbol{My} = \boldsymbol{b} \\ UPL_i \le z_i^+ \le ub_i \quad i = 1,\ldots,n \\ lb_i \le z_i^- \le LPL_i \quad i = 1,\ldots,n \\ z_i^- + z_i^+ \ge SPL_i \quad i = 1,\ldots,n. \end{cases}$$

Questo problema di programmazione lineare risulta notevolmente più semplice di quello della soppressione secondaria ottimale (per l'assenza di variabili binarie) e può essere risolto efficientemente nello spazio duale (Per una descrizione più completa del modello di programmazione si rimanda a Fischetti e Salazar, 2002). Inoltre, la soppressione parziale può essere applicata a tabelle collegate e di grande dimensione.

## 3.2 Tecniche di protezione non perturbative: riarrangiamento delle categorie delle variabili classificatrici

Una cella che risulta sensibile rispetto ad una misura di rischio subadditiva può essere ridistribuita all'interno della tabella cambiando le categorie di una o più modalità delle variabili classificatrici. Se le modalità della variabile classificatrice sono sconnesse allora si deve accorpare la modalità contenente la cella sensibile ad un'altra, se le modalità sono trasferibili si possono cambiare i limiti di alcune classi per "mischiare" i rispondenti. Si consideri l'esempio del fatturato dei produttori di strumenti musicali in Tabella 3.8 in cui una variabile classificatrice è il numero di addetti suddivisa nelle classi $[(<20),(20 \mapsto 50),\geq 50]$. I dati dei produttori di pianoforti nelle classi di addetti [25-50) e $[\geq 50]$ non sono rilasciabili in quanto sono in numero inferiore a 3.

**Tabella 3.8 Produttori di strumenti musicali per addetti, in neretto sono evidenziate le frequenze della riga con celle a rischio.**

| Strumenti | Numero addetti | | | |
|---|---|---|---|---|
| | <20 | 20-50 | ≥ 50 | Totale |
| Arpe | 58 | 47 | 36 | 141 |
| Organi | 71 | 124 | 24 | 219 |
| Pianoforti | 92 **(5)** | 157 **(2)** | 59 **(1)** | 308 **(8)** |
| Altro | 800 | 934 | 651 | 2385 |
| Totale | 1021 | 1262 | 770 | 3053 |

Se il numero di addetti viene ripartito diversamente, per esempio in due classi $[(<18),(\geq 18)]$ come mostrato in Tabella 3.9, le celle a rischio sono eliminate e la tabella risulta sicura. Naturalmente questo riarrangiamento degli estremi degli intervalli è solo uno dei tanti possibili.

**Tabella 3.9 Riarrangiamento delle classi del numero di addetti. Tabella ottenuta riarrangiando le classi della variabile "NUMERO ADDETTI" della Tabella 3.8**

| Strumenti | Numero addetti | | |
|---|---|---|---|
| | <18 | ≥ 18 | Totale |
| Arpe | 58 | 83 | 141 |
| Organi | 71 | 148 | 219 |
| Pianoforti | 92 **(4)** | 216 **(4)** | 308 **(8)** |
| Altro | 800 | 1585 | 2385 |
| Totale | 1021 | 2032 | 3053 |

Anziché riarrangiare la variabile "numero di addetti", si sarebbe potuto riarrangiare la variabile "strumenti prodotti", per esempio raggruppando i produttori di organi con quelli di pianoforti, ottenendo la Tabella 3.10, che risulta sicura. La scelta del riarrangiamento può essere fatta secondo vari criteri: nel caso in cui esistano delle

classificazioni gerarchicamente superiori prestabilite, sembra opportuno attenersi a queste, sia per favorire il confronto con altri dati che per evitare che il confronto con altre tabelle collegate crei possibilità di violazione; altrimenti si possono seguire criteri di omogeneità (come nell'esempio in Tabella 3.10 dove si è posto: [produttori di organi e di pianoforti] = produttori di strumenti a tastiera); oppure si può scegliere un riarrangiamento che minimizzi una perdita di informazione.

**Tabella 3.10 Riarrangiamento delle classi degli strumenti prodotti. Tabella ottenuta aggregando le classi "Organi" e "Pianoforti" della variabile "STRUMENTI" della Tabella 3.8**

| Strumenti | Numero addetti | | | |
|---|---|---|---|---|
| | <20 | 20-50 | ≥50 | Totale |
| Arpe | 58 | 47 | 36 | 141 |
| Organi e Pianoforti | 163 **(10)** | 281**(5)** | 83 **(4)** | 527 **(19)** |
| Altro | 800 | 934 | 651 | 2385 |
| Totale | 1021 | 1262 | 770 | 3053 |

Quest'ultimo criterio può attuarsi considerando che l'accorpamento di due categorie è equivalente alla soppressione di tutte le celle delle due righe (o colonne). In questi termini è possibile modificare le procedure di calcolo della soppressione secondaria ottimale. Si consideri per esempio di assegnare pesi pari al valore delle celle alla Tabella 3.8; la riga con cui l'accorpamento minimizza la perdita di informazione è quello corrispondente ai produttori di arpe, in quanto è quella che ha totale marginale minore. Per le variabili continue non sembra essere stata sviluppata una metodologia standard. Una possibilità è quella di definire un numero di classificazioni possibili e paragonare poi le varie tabelle rispetto ad una misura di informazione.

## 3.3 Tecniche di protezione perturbative: arrotondamento

L'arrotondamento consiste nel rilasciare tabelle in cui i valori veri sono stati cambiati con multipli di una certa base. Questo metodo di protezione è esteticamente migliore della soppressione ma di dubbia efficacia. Infatti, la necessità di rilasciare tabelle che siano additive, cioè in cui i totali marginali siano la somma dei valori interni, richiede tecniche di calcolo alquanto sofisticate che rendono il controllo della perturbazione difficile. Inoltre le tecniche ottimali non sono applicabili a tabelle con tre o più dimensioni, le quali richiedono metodi euristici che possono avere alta complessità computazionale e che non assicurano la convergenza. Per questi motivi è difficile mantenere il controllo del livello di protezione ottenuto e dell'informazione persa.

L'arrotondamento fornisce una protezione tanto più alta quanto più grande è la base, in quanto l'intervallo di esistenza dei valori delle celle è funzione crescente di questa. Quindi, il valore della base di arrotondamento dovrebbe essere scelto rispetto all'intervallo di esistenza ricavabile, che è costante per ogni cella ma differisce a seconda del tipo di arrotondamento scelto.

Data una base di arrotondamento intera, *b*, il valore di una cella, z, si può scrivere

$z = kb + r$, dove k è un intero e $0 \le r \le (b-1)$ è il resto in base *b*. Il valore arrotondato in base *b* di *z* è

$$[z] = kb + \Phi(r),$$

dove $\Phi(r) = \{0, b\}$.

Un arrotondamento è ristretto a zero quando i valori multipli della base scelta vengono lasciati invariati, cioè se Φ(0)=0. Un arrotondamento è stocastico se Φ(*r*) assume il valore b con probabilità *p* ed il valore 0 con probabilità (1-*p*) altrimenti è deterministico. Quando il valore atteso dell'arrotondamento è il valore vero, cioè se si verifica E([*z*])=*z*, l'arrotondamento è detto non distorto. Un arrotondamento ristretto a zero che produce tabelle additive è detto controllato.

Esistono diversi metodi di arrotondamento. Presenteremo ora brevemente quelli più comunemente utilizzati; per maggiori dettagli sui metodi di arrotondamento qui presentati si veda Willenborg e de Waal (2001).

### 3.3.1 Arrotondamento deterministico

L'arrotondamento deterministico convenzionale è quello più semplice ma anche quello più grezzo, si attua sostituendo i valori delle celle con i loro multipli di una base *b* più prossimi. Quindi il valore arrotondato $[z]$ si ottiene ponendo

$$\Phi(r) = \begin{cases} 0 & r/b < 1/2 \\ b & r/b \ge 1/2. \end{cases}$$

E' facile vedere che l'ampiezza dell'intervallo di esistenza per ogni cella è uguale a *b*.

Come tutti i metodi di arrotondamento non controllati, questa tecnica non assicura l'additività della tabella protetta. Come esempio si consideri la Tabella 3.12 ottenuta arrotondando in base cinque la Tabella 3.11. Dai totali marginali evidenziati in neretto si nota come la tabella protetta non sia additiva.

**Tabella 3.11 Esempio di tabella prima dell'arrotondamento**

| | X | Y | W | Totale |
|---|---|---|---|---|
| A | 12 | 9 | 16 | 37 |
| B | 6 | 8 | 12 | 26 |
| C | 2 | 20 | 13 | 35 |
| Totale | 20 | 37 | 41 | 98 |

**Tabella 3.12 Tabella ottenuta arrotondando in base cinque la Tabella 3.11. I valori marginali che non rispettano l'additività sono evidenziati in neretto**

| | X | Y | W | Totale |
|---|---|---|---|---|
| A | 10 | 10 | 15 | 35 |
| B | 5 | 10 | 10 | 25 |
| C | 0 | 20 | 15 | 35 |
| Totale | **20** | **35** | 40 | **100** |

La perdita di informazione dovuta all'arrotondamento convenzionale è bassa, se paragonata a quella degli altri metodi di arrotondamento, ma anche la protezione offerta è inferiore; per alcune indicazioni sugli effetti della scelta della base nell'arrotondamento deterministico convenzionale si rimanda a Shackis (1993). La Tabella 3.13 mostra un arrotondamento convenzionale in base cinque di una tabella d'intensità con i totali marginali non arrotondati. Poiché i valori originali devono essere compresi tra zero e due, è facile desumere dai totali marginali che devono essere tutti uguali a due. Quindi, in quest'esempio la protezione non è adeguata.

**Tabella 3.13 Esempio di arrotondamento convenzionale in base 5**

| | $C_1$ | $C_2$ | Totale |
|---|---|---|---|
| $R_1$ | 0 | 0 | 4 |
| $R_2$ | 0 | 0 | 4 |
| Totale | 4 | 4 | 8 |

### 3.3.2 Arrotondamento stocastico non controllato

L'arrotondamento stocastico è spesso preferito a quello deterministico in quanto, per l'aleatorietà della regola, l'intervallo di esistenza dei valori delle celle sarà uguale a $2b$, quindi doppio a quello ottenibile con i metodi deterministici. Si sottolinea, però, che, per lo stesso motivo, la perdita di informazione non è prevedibile.

Il metodo stocastico più semplice è quello in cui si sceglie arbitrariamente una probabilità $p$ uguale per tutte le celle, e si pone

$$\Phi(r) = \begin{cases} 0 & \text{con probabilità } 1-p \\ b & \text{con probabilità } p. \end{cases}$$

L'arrotondamento stocastico non distorto e controllato a zero differisce da quello sopra per il fatto che si richiede che $\Phi(0) = 0$ e che il valore atteso dell'arrotondamento sia uguale al valore del resto, cioè che $\Phi(r) = r$. Perciò le probabilità di arrotondamento superiore o inferiore dipendono dal valore del resto. Dato che $E[\Phi(r)] = pb$, la non distorsione si ottiene ponendo

$$\Phi(r) = \begin{cases} 0 & r = 0 \\ \left.\begin{array}{ll} 0 & \text{con probabilità } 1 - r/b \\ b & \text{con probabilità } r/b \end{array}\right\} & r \neq 0. \end{cases}$$

In Tabella 3.14 sono indicate queste probabilità per l'arrotondamento in base 5.

**Tabella 3.14 Probabilità di arrotondamento superiore ed inferiore nell'arrotondamento stocastico non distorto controllato a zero, per b=5.**

| Resto | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Pr[Φ(r)=5] | 1/5 | 2/5 | 3/5 | 4/5 |
| Pr[Φ(r)=0] | 4/5 | 3/5 | 2/5 | 1/5 |

La giustificazione per l'adozione di questo metodo è proprio la non distorsione, anche se, in questo frangente, tale proprietà non sembra rivestire la stessa importanza che nella teoria della stima. Entrambi questi metodi prevedono che anche i totali marginali siano arrotondati e, quindi, non assicurano l'additività.

### 3.3.3 Arrotondamento stocastico controllato

I metodi di arrotondamento stocastico controllato sono i più importanti ed utilizzati, in quanto assicurano l'additività della tabella ottenuta. Come accennato sopra, l'individuazione di soluzioni che soddisfino questo vincolo richiedono l'impiego di metodi di soluzione alquanto complessi.

In Fellegi (1975) è presentato un metodo per tabelle unidimensionali ristretto a zero e non distorto, che andiamo ora a descrivere. Si consideri una tabella con $n$ celle con valori $z_i$ e totale T, che si vogliono arrotondare in base $b$. Siano $S_i = \sum_{j=1}^{i} r_i$ le somme cumulate dei resti, dove $S_0$=0, e $1 \le \mathrm{R1} \le b$ un intero casuale. Si definiscono le quantità $R_i = \mathrm{R1} + (i-1)b$ e l'arrotondamento si applica ponendo

$$\Phi(r_i) = \begin{cases} b & S_{i-1} < R_i \le S_i \\ 0 & \text{altrimenti;} \end{cases}$$

il totale marginale si ottiene come [T]=$\sum_{\mathrm{i}}[z_i]$, in modo da ottenere l'additività. Si può dimostrare che questo metodo è ristretto a zero e non distorto. in Tabella 3.15 è mostrata un'applicazione di questo metodo in base 5. Poiché le somme cumulate sono S=( 0, 0, 2, 6, 10, 14, 15, 19, 19, 20, 24) e avendo ottenuto R1=1, risulta R=(1, 6, 11, 16, 21, 26, 31, 36, 41, 46), e quindi tutti i valori saranno arrotondati al multiplo di cinque inferiore.

Per comparazione, nella Tabella 3.15 sono riportati i valori arrotondati convenzionalmente; le differenze tra i due metodi di arrotondamento sono evidenziate in neretto. Il totale del metodo di Fellegi differisce da quello vero molto di più di quello convenzionale, evidenziando come la protezione offerta dal primo metodo sia maggiore, così come sacrificio di informazione.

**Tabella 3.15 Tabella unidimensionale arrotondata con il metodo di Fellegi e con l'arrotondamento convenzionale deterministico. I valori diversi tra i due metodi sono evidenziati in neretto.**

| Celle | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Totale |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *z* | 25 | 37 | 4 | 14 | 49 | 26 | 39 | 50 | 36 | 24 | 304 |
| [*z*] Fellegi | 25 | 35 | **0** | **10** | **45** | 25 | **35** | 50 | 35 | **20** | **280** |
| [*z*] Deterministico | 25 | 35 | 5 | 15 | 50 | 25 | 40 | 50 | 35 | 25 | 305 |

Lo stesso tipo di arrotondamento per tabelle bidimensionali può essere ottenuto con il metodo descritto in Cox (1987), quello per le tabelle a maggiori dimensioni può essere ottenuto, in alcune circostanze, con metodi euristici basati sul metodo del simplesso per interi. Per maggiori dettagli su questi metodi si rimanda al testo di Willenborg e de Waal (2001) e alle referenze ivi citate.

# Capitolo 4. Tutela statistica della riservatezza per dati rilasciati da siti Web(*)

## 4.1 Introduzione

Internet sta diventando un canale standard attraverso cui le istituzioni comunicano con il pubblico: le istituzioni si aspettano che il pubblico cerchi informazioni sul Web ed il pubblico si aspetta di trovarcele. Molti istituti di statistica hanno cominciato, già da anni, a diffondere anche dati tramite siti Web. Infatti, quasi tutti gli istituti nazionali di statistica occidentali forniscono on line una serie di indicatori e di risultati di indagini. Ad esempio, l'Istat, nel sito http://www.istat.it, fino al settembre 2002 offriva le seguenti banche dati: 14° Censimento popolazione e abitazioni 2001; 8° Censimento industria e servizi 2001; Popolazione e statistiche demografiche; Commercio estero; Censimento intermedio industria e servizi 1996; Indicatori economici congiunturali; Handicap; Indicatori socio sanitari (regionali); Sistema di indicatori sociali (provinciali); Fmi - *National Summary Data Page* e Coltivazioni: dati congiunturali e Politiche di sviluppo regionali. L'*Office for National Statistics* (Ons) offre statistiche del Regno Unito al sito http://www.statistics.gov.uk. Gli Stati Uniti d'America offrono una gran quantità di dati on line, una lista di siti si può trovare a http://www.lib.umich.edu/govdocs/stats.html.

Le diffusioni stampate restano ancora il mezzo di diffusione ufficiale per la loro caratteristica di immutabilità, ma le pubblicazioni Web presentano dei grandi vantaggi rispetto a queste. Infatti, le pubblicazioni Web sono meno costose, sia per chi le offre che per chi ne fruisce, e sono molto più facilmente accessibili. Inoltre, consentono di rilasciare e aggiornare grandi quantità di dati tempestivamente. Poi, l'estrema flessibilità dei moderni linguaggi *Mark-Up*, consente di rilasciare i dati sia numericamente che graficamente. La prerogativa unica dei *Siti Web per la diffusione di dati* (da qui in avanti Swdd) è l'interattività, grazie alla quale gli utenti di possono scegliere tra varie soluzioni offerte; per esempio, quali dati richiedere. Fienberg (2000) nota che questa caratteristica deve comportare un cambiamento di mentalità nella diffusione.

Le caratteristiche positive del rilascio dei dati attraverso il Web, al tempo stesso, però, rendono la tutela statistica della riservatezza per i Swdd più complicata. Infatti, occorre proteggere in breve tempo un grande numero di dati, collegati e facilmente accessibili. In principio, le pubblicazioni Web potrebbero essere protette con i metodi standard visti nei capitoli precedenti. Però, quando il numero dei dati da proteggere è elevato, l'applicazione manuale di questi metodi risulta troppo onerosa. Quindi, nel trattare il problema della tutela statistica della riservatezza, daremo per scontato che i dati da proteggere sono in quantità tale da richiedere tecniche con applicazione automatica. Nella pratica, spesso, i metodi di protezione sono applicati seguendo tacitamente logiche euristiche, magari suggerite dalla conoscenza delle proprietà dei dati. L'applicazione automatica dei metodi di protezione, invece, necessita della formalizzazione dei criteri di queste procedure.

La ricerca sulla tutela statistica della riservatezza per dati rilasciati sul Web è ancora nella sua fase iniziale e non tutti i problemi sono stati ancora affrontati e risolti

(*) Capitolo redatto da Giovanni M. Merola

(Franconi e Merola, 2003). Data l'impellenza dell'argomento, il governo degli Stati Uniti ha costituito appositamente il progetto di ricerca, il *Digital Government*[8] (Dg), presso il *National Institute for Statistical Sciences* (Niss). A questo progetto partecipano professori e ricercatori di diverse università americane ed istituti di statistica, oltre che del Niss. L'orientamento metodologico per la protezione dei Swdd che sembra prevalere va verso la definizione di sistemi esperti completamente automatici (si veda, per esempio, Hoffman, 1977, Keller-McNulty e Unger, 1998, Fienberg *et al*. 1998, Malvestuto e Moscarini, 1999 e Fienberg, 2000).

La creazione di un sistema di protezione automatica di dati, richiede la formalizzazione del problema di tutela statistica della riservatezza; questo può essere fatto usando l'approccio *rischio-utilità*, Duncan *et al.* (2001), Trottini (2001), Trottini e Fienberg (2002). In questo approccio, gli effetti di una protezione dei dati vengono valutati in termini dell'utilità che i dati così protetti hanno per i fruitori. I parametri del metodo di protezione scelto vengono determinati massimizzando l'utilità dei dati rilasciati, sotto il vincolo che il rischio resti al di sotto di una data soglia. Siano: $D$ l'insieme dei dati eleggibili per essere rilasciati e $\mathcal{D}$ uno dei possibili risultati della protezione. I dati in $\mathcal{D}$ variano a seconda di come è applicato il metodo di protezione. Se $R(\mathcal{D})$ è la misura del rischio prescelta e $U(\mathcal{D})$ la misura di utilità, l'insieme ottimale, $\mathcal{D}^*$, viene determinato come soluzione del problema:

$$\begin{cases} \max_{\mathcal{D}} U(\mathcal{D}) \\ R(\mathcal{D}) \le \alpha, \end{cases}$$

dove $\alpha$ è la soglia di rischio accettabile. Il calcolo delle soppressioni secondarie ottimali è un esempio di approccio rischio-utilità, in cui l'informazione ha il ruolo di utilità per dati tabulari (vedi Paragrafo 3.1.2).

Le misure di rischio e di utilità possono essere scelte secondo diversi criteri, a seconda della natura dei dati. Duncan *et al*. (2001) considerano diverse misure di utilità per la perturbazione dei microdati, proponendo di valutare graficamente gli effetti delle scelte di diversi parametri rispetto all'utilità e il rischio ottenuti. Si noti che, nella definizione, ci si riferisce a misure di rischio ed utilità globali, cioè per l'insieme dei dati rilasciati e non per i singoli dati. Perciò, per calcolare queste misure è necessario considerare possibili dipendenze tra i dati rilasciati. Più avanti daremo degli esempi di misure di utilità. Quando i dati da rilasciare sono molto numerosi, il problema della scelta ottima dell'insieme $\mathcal{D}^*$ può essere computazionalmente tanto oneroso da non essere realisticamente risolvibile. Per cui, in molte applicazioni di sistemi automatici di protezione, si utilizzano algoritmi euristici per calcolare soluzioni sub-ottimali. Alcuni saranno illustrati nel prosieguo di questo capitolo.

Con i Swdd si possono rilasciare informazioni in diversi modi: i più comuni rilasciano tabelle ma, negli ultimi anni, sono nati Swdd più sofisticati, in cui si possono ottenere anche altre statistiche sulle variabili, come, per esempio, coefficienti di regressione

[8] Maggiori informazioni su questo progetto possono essere reperite al sito *Web* http://www.niss.org/dg/

o di correlazione. Per analizzare l'applicazione della tutela della riservatezza, distingueremo tre tipi di Swdd, per tipo di dati rilasciati e per approccio alla protezione:

**Swdd statici**: offrono solo tabelle; l'informazione massima ottenibile è predeterminata; si possono descrivere come una serie di collegamenti ipertestuali a dati già protetti;

**Swdd dinamici**: offrono solo tabelle; la protezione viene applicata in base all'informazione già rilasciata. In principio, consentono di richiedere qualsiasi tabella;

**Laboratori virtuali**: consentono di effettuare (anche) analisi statistiche sui dati disponibili; non distingueremo tra statici e dinamici.

## 4.2 Approcci alla protezione dei Siti Web per la diffusione di dati (Swdd)

I Swdd possono essere realizzati in molti modi diversi e rilasciano informazioni diverse in forme diverse. Perciò, la loro protezione deve essere personalizzata secondo la natura del sito e dei dati rilasciati. In molti casi, i Swdd contengono dati di cui sono già state pubblicate delle statistiche, per esempio dei totali; in altri casi, i dati vengono aggiornati periodicamente. Per cui occorre che la protezione dei Swdd sia coerente con le protezioni precedenti e con le informazioni già rilasciate. Questi problemi sono intrinseci ai siti dinamici, con la complicazione che le informazioni rilasciate possono essere molteplici e non conosciute in anticipo. Uno degli elementi che occorre considerare per definire le misure del rischio di violazione e di utilità, necessarie all'applicazione dei metodi statistici di tutela della riservatezza, è la natura degli utenti che possono accedere ai dati.

### 4.2.1 Limitazioni di ingresso e di uscita

Un modo per ridurre i rischi di violazione della privacy nei Swdd, è quello di restringere l'accesso solo ad utenti selezionati o di rilasciare le informazioni richieste per e-mail. In questo modo gli amministratori di sistema possono ridurre il rischio che utenti malintenzionati cerchino di fare un uso inappropriato dell'informazione messa a disposizione. Nei siti con accesso ristretto, gli utenti debbono registrarsi e la connessione è consentita solo a quelli che soddisfano date condizioni. Le condizioni richieste possono essere di vario tipo, per esempio: la più generica e comune è: "avere un indirizzo e-mail valido"; un'altra condizione comune, ma più restrittiva, è: "appartenere ad un'istituzione di ricerca". La registrazione degli utenti può richiedere o meno l'accettazione elettronica di un accordo di riservatezza. In tutti i modi, le restrizioni di accesso rigide limitano la pubblicità di un sito, mentre le restrizioni meno stringenti sono spesso poco efficaci, a causa della facilità con cui è possibile procurarsi un indirizzo e-mail in modo anonimo. Un'altra limitazione al rischio di intrusione è quello di rilasciare per e-mail le informazioni richieste. In questo modo l'utente deve fornire un indirizzo di e-mail valido e gli amministratori del sito possono tenere un registro di quello che è stato rilasciato a ciascun utente. Un altro modo di ottenere tale registro è quello di registrare gli IP che si connettono al sito e le attività degli utenti. Questo metodo, però, può essere molto laborioso ed è di difficile uso in sede legale, in

quanto nascosto all'utente. Le restrizioni di accesso e di uscita, comunque, sono pertinenti alla sicurezza informatica del sito e non alla tutela statistica della riservatezza e, pertanto, non saranno trattate oltre. Queste, però, in certi casi, sono utilizzate per giustificare ipotesi alla base dei metodi statistici e, quindi, costituiscono parte integrante del sistema di tutela della riservatezza per un Swdd. Analisi interessanti delle restrizioni di ingresso dal punto di vista della tutela statistica della riservatezza si possono trovare in Adam e Wortmann (1989), David (1998) e Blakemore (2001).

### 4.2.2 Metodi statistici

Duncan (2001) classifica i metodi statistici per la tutela della riservatezza in termini di *disclosure limiting masks* (Duncan e Pearson, 1991), noi considereremo due gruppi, più ampi e non mutuamente esclusivi:

**metodi perturbativi**: comprendono sia la perturbazione dei dati d'origine (soppressione di parte delle osservazioni, *swapping*, aggiunta di disturbi, riarrangiamento delle modalità, eccetera) che quella dei dati aggregati (arrotondamento, aggiunta di disturbi, simulazione, eccetera);

**metodi soppressivi**: comprende sia le soppressione di singole celle che la soppressione totale di tabelle, cioè il rifiuto di una richiesta.

Fienberg (2000) propone di considerare i diversi metodi statistici per la tutela della riservatezza alla luce di tre criteri: *usabilità* dei dati prodotti, *trasparenza* e *dualità*. La scelta del metodo, comunque, spesso deve tener conto anche della facilità di applicazione e delle consuetudini relative alla diffusione di ciascuna categoria di dati. Le tecniche perturbative, spesso, vengono applicate sulla base di una stima grossolana del rischio. Il principio di protezione su cui si basano è quello di ingenerare una incertezza diffusa a tutti i dati rilasciati, distorcendoli; quindi, vengono distorti anche i dati rilasciabili. I metodi perturbativi, sono facilmente implementabili ma la protezione reale che producono è difficilmente controllabile e, quindi, si tende ad implementarli proteggendo eccessivamente i dati, cioè introducendo una distorsione superiore a quella necessaria. I metodi soppressivi eliminano completamente i dati a rischio lasciando inalterati molti dei dati rilasciabili e la perdita di informazione che causano è controllabile. Per dirla con Dobra *et al.* (2002) (pag. 3): "… [*I metodi soppressivi*] dicono sempre la verità, anche se non tutta la verità." Dal punto di vista dell'applicazione, però, i metodi soppressivi sono quelli più complessi e, spesso, richiedono calcoli intensi.

## 4.3 Protezione dei Swdd che rilasciano tabelle

Il problema principale nella protezione dei Swdd che rilasciano tabelle è la presenza di molte celle dipendenti, per cui, in generale, il rischio e l'utilità devono essere completamente ricalcolati ogni volta che cambia l'insieme rilasciato. Quindi, il problema teorico di maggior rilievo è quello delle tabelle collegate (vedi Paragrafo 2.8);

la soluzione ottima di questo problema richiede algoritmi di ricerca che possono essere affrontati computazionalmente solo per basse dimensioni; quando il numero dei dati collegati offerto è elevato, il problema dell'ottimizzazione rischio-utilità diventa praticamente irrisolvibile. Ci sono due "scappatoie": perturbare i dati di origine o usare algoritmi sub-ottimali.

### 4.3.1 Metodi perturbativi: riarrangiamento delle modalità

Il riarrangiamento delle modalità di una variabile classificatrice, discussa nel Paragrafo 3.2, consiste nell'assegnare lo stesso valore a modalità diverse di una variabile classificatrice. Questo metodo pertiene alla protezione delle tabelle, in quanto il rischio è calcolato per dati tabellari, però può benissimo essere usato per la protezione di microdati.

In molti casi le modalità di una variabile classificatrice possono essere ordinate, diciamo, per affinità; come per affinità geografica (per esempio: comuni) o per affinità di classificazione (per esempio: classi Ateco di attività economica). In certi casi le etichette numeriche delle modalità sono *proxy* dell'ordine. Quindi, data una tabella con alcune celle a rischio, il metodo consiste nell'aggregare alcune modalità di una o più variabili in modo che le celle risultanti non siano a rischio. L'aggregazione di celle sicuramente riduce il rischio, se la sua misura è sub-additiva (vedi Paragrafo 2.4). Una buona aggregazione dovrebbe eliminare le celle a rischio minimizzando la perdita di informazione (vedi Paragrafo 3.1). Questi algoritmi devono essere costruiti *ad hoc* per ogni tipo di database. Per esempio, i dati del Nass[9] sono stati protetti con un sistema di riarrangiamento automatico (Karr *et al.*, 2002), che ora andiamo a descrivere.

#### I dati del Nass

Il Nass contiene 194410 osservazioni sull'uso di 322 prodotti chimici per 67 tipi di coltivazioni in 30500 aziende agricole statunitensi. I dati sono rilevazioni di un'indagine annuale, dal 1996 al 1998. Le informazioni contenute in ogni osservazione sono: *identificativo azienda*, *stato*, *contea, anno*, *superficie*, *coltivazione, tipo di prodotto chimico* e *quantità applicata*. Gli utenti possono richiedere celle fino al livello di contea. Sono state adottate due misure di rischio, la soglia di tre osservazioni e la dominanza $n=1$, $p=0.6$. Siccome più del cinquanta percento delle contee risulta a rischio, per la protezione è stato preferito il metodo dell'aggregazione delle modalità, all'alternativa di rifiutare le tabelle. Naturalmente, si tratta di aggregazione di dati di più contee. Cioè, i dati di ogni contea a rischio vengono diffusi aggregati in una *supercontea*, formata con una o più altre contee, in modo che non sia a rischio essa stessa.

La procedura di aggregazione delle contee si basa su due diversi algoritmi *greedy* (cioè, che non permettono di modificare scelte): il *pure* e lo *small*. Per ogni contea *a rischio,* gli algoritmi vagliano contee e supercontee confinanti, che aggregano o meno secondo delle *regole* di precedenza. Queste regole di precedenza sono analoghe a una

[9] National Agricultural Statistics Service, http://www.usda.gov/nass/

misura di utilità discreta. Nell'algoritmo *pure* si cerca di non aggregare contee rilasciabili, mentre nello *small* si favorisce la formazione di piccole supercontee. Il primo algoritmo tende a creare grandi supercontee formate da molte contee a rischio. Per esempio nel *pure*, una contea a rischio sarà aggregata ad una supercontea solo se non esistono altre contee a rischio. Le contee e supercontee confinanti sono visitate in ordine casuale. Un'aggregazione è effettuata alla prima occasione favorevole, fatta la quale, l'algoritmo prosegue con i valori aggiornati, sino ad esaurimento delle supercontee a rischio. Entrambi gli algoritmi tendono, in modo diverso, a formare aggregazioni troppo grandi. Per alleviare questo problema, per la protezione dei dati del Nass sono stati adottati entrambi, in due passi: prima viene fatto girare lo *small,* fino all'eliminazione delle singole contee a rischio, e poi viene fatto girare il *pure* per le supercontee. A quanto si può dedurre, le aggregazioni vengono eseguite solo se delle contee a rischio vengono richieste e sono poi registrate.

### 4.3.2 Metodi perturbativi casuali

Nei metodi perturbativi casuali rientrano tutte quelle tecniche di protezione che distorcono le osservazioni in maniera (pseudo-)casuale. Cioè, per esempio: assegnazione di pesi a ciascun record nel file di dati elementari di origine (Cuppen e Willenborg, 2003), il sottocampionamento, vale a dire la selezione di un campione delle osservazioni disponibili, l'aggiunta di disturbi, lo scambio di valori, il Pram, eccetera (vedi Paragrafo 8.3). I metodi casuali possono essere implementati facilmente e possono essere usati congiuntamente. Uno dei problemi principali nell'uso di questi dati per la protezione delle tabelle è che le tabelle protette possono non essere additive (cioè, in cui le somma dei valori delle celle non sono uguali ai totali marginali). L'uso delle protezioni casuali può essere giustificata quando il numero dei record individuali è molto elevato, in quanto si possono comunque ottenere stime corrette. Però, gli effetti della protezione e della perdita di informazione dovuti alle perturbazioni casuali sono difficili da valutare e, quindi, la scelta dei parametri, cioè, per esempio, la numerosità del campione, il tasso di *swapping* o la varianza dei disturbi, può essere problematica. In genere, questi vengono decisi euristicamente da esperti; Duncan *et al*. (2001) propongono di determinare i parametri graficamente attraverso la *Confidentiality Map*, costruita secondo l'approccio rischio-utilità. Evans *et al*. (1998) considerano l'aggiunta di disturbi ai dati individuali per la protezione delle tabelle.

Il sottocampionamento è spesso utilizzato per proteggere dati censuari (per esempio i dati del censimento della popolazione brasiliana del 1996)[10]; a volte non ci si accontenta di questo e si applicano protezioni aggiuntive, come nel caso dei dati del censimento degli Stati Uniti d'America, che andiamo ad illustrare brevemente di seguito.

#### Dati del censimento statunitense 2000 sull'American Fact Finder

Zayatz (2002) descrive, a grandi linee, la protezione dei dati del censimento della

[10] http://sda.berkeley.edu:7502/IBGE/

popolazione statunitense del 2000, che sono rilasciati sul Swdd statico *American Fact Finder*[11] (Aff). Il censimento degli Stati Uniti viene effettuato con due diversi questionari: la *short form*, distribuita al 100 per cento della popolazione, e la *long form*, che contiene molti più quesiti ed è distribuito ad un campione di un sesto dell'intera popolazione di famiglie. I dati dell'Aff sono protetti con perturbazione dei microdati e, inoltre, le tabelle sono controllate automaticamente prima di essere rilasciate.

Per la protezione dei dati del 2000 sono stati applicati più metodi perturbativi contemporaneamente, migliorando la protezione applicata ai dati del censimento del 1990. In particolare, i dati della long form sono stati protetti maggiormente, in quanto per il 1990 si era ritenuto che il campionamento garantisse, già di per se, una protezione sufficiente. I dati delle due indagini sono stati protetti con le stesse tecniche ma adottando parametri diversi. Per brevità, nel resoconto tralasceremo queste differenze.

Alcuni dei dati rilevati vengono rilasciati al livello di *isolato*, che ha dimensione media di 34 persone, altri sono rilasciati solo al livello di *gruppo di isolati,* che ha dimensione media di 1348 persone. Essendo, quindi, diffusi con ampio dettaglio, i dati richiedono una protezione elevata. Una prima forma di protezione è stata quella di scegliere un sottocampione dei dati disponibili. Di questi dati, alcuni sono stati sostituiti con valori imputati, come se fossero stati mancanti. I valori di alcune variabili continue sono stati *top* e *bottom coded*, mentre in alcune variabili sono state aggregate delle categorie (come per la variabile *razza*). Inoltre, i valori di alcuni variabili sensibili sono stati permutati con valori di altre unità, a parità di alcune variabili chiave. Le unità permutate non sono state rese note. Le probabilità di permutazione sono state definite per ogni isolato, in modo inversamente proporzionale al numero degli abitanti (e proporzionalmente al peso campionario per la *long form*). Gli isolati i cui dati erano stati tutti imputati sono stati esclusi dal processo di permutazione. Le tabelle vengono, quindi, prodotte con questi dati.

L'Aff offre agli utenti (anonimi) la possibilità di scegliere di visualizzare qualsiasi tabella. Per garantire una sicurezza assoluta, le tabelle prodotte dai dati perturbati sono, comunque, controllate automaticamente per problemi di riservatezza prima di essere rilasciate.

### 4.3.3 Metodi soppressivi

Dal punto di vista della tutela della riservatezza, i Swdd statici non differiscono dalle pubblicazioni cartacee, quindi le soppressioni potrebbero essere calcolate come visto nei capitoli precedenti; però, quando il numero delle variabili offerte nel database è elevato il calcolo delle soppressioni per tutte le tabelle possibili diventa un compito praticamente impossibile.[12] Per cui, in genere, si preferisce limitarsi a stabilire un sottoinsieme di tabelle che possano essere rilasciate garantendo la riservatezza, sopprimendo interamente le altre. Anche la scelta di questo sottoinsieme è molto onerosa in termini computazionali, principalmente per la presenza di tabelle collegate. Alcuni dei problemi di riservatezza legati al rilascio di tabelle collegate sono stati

[11] http://factfinder.census.gov/servlet/BasicFactsServlet
[12] Ricordiamo che se sono offerte $p$ variabili classificatrici e $q$ risposte il numero totale di tabelle ottenibili è $q2^p$.

illustrati nel Paragrafo 3.2. Di seguito daremo una descrizione di alcuni aspetti del rilascio di tabelle collegate gerarchiche.

#### 4.3.3.1 Tabelle collegate gerarchiche

Date $p$ variabili classificatrici categoriche in un database, l'insieme delle tabelle che è possibile costruire è detto spazio delle tabelle. La tabella massima che si può costruire è quella $p$-dimensionale, $T_p \in \boldsymbol{T}_p$. Chiaramente, $T_p$ contiene tutte le ($p$-1) tabelle ($p$-1)-dimensionali in $\boldsymbol{T}_{p-1}$, che sono le sue marginali, tutte le $p(p-1)/2$ tabelle ($p$-2)-dimensionali e così via. Ognuna delle $\binom{p}{k}$ tabelle $k$-dimensionali esistenti ha per marginali $k$ tabelle ($k$-1)-dimensionali, eccetera. Sarebbe a dire che, la Tabella 4.1

**Tabella 4.1 Tabella bidimensionale con variabili classificatrici C1 e C2**

| $C_1 \backslash C_2$ | $C_{21}$ | $C_{22}$ | Totale |
|---|---|---|---|
| $C_{11}$ | 3 | 4 | 7 |
| $C_{12}$ | 1 | 2 | 3 |
| Totale | 4 | 6 | 10 |

ha per marginali le due tabelle unidimensionali, mostrate in Tabella 4.2:

**Tabella 4.2 Tabelle unidimensionali marginali della Tabella 4.1**

| $C_{11}$ | $C_{12}$ | Totale |
|---|---|---|
| 7 | 3 | 10 |

e

| $C_{21}$ | $C_{22}$ | Totale |
|---|---|---|
| 4 | 6 | 10 |

Per convenzione la tabella zero-dimensionale, $\varnothing$, contiene il totale generale; nell'esempio sopra questa è il valore {10}.

Rilasciando una tabella $k$ dimensionale si rilasciano tutte le sue ($2^k - 1$) marginali, fino a $\varnothing$. Quindi esiste un ordinamento parziale *madri-figlie.* Le figlie sono ottenute dalle madri eliminando variabili classificatrici sommando rispetto alle loro categorie.
Per esempio, la tabella:

[età, sesso, stato civile, titolo di studio]

è madre della sua marginale

$$[\text{età,sesso,stato civile}] = \sum_{j=\text{titolo di studio}} [\text{età,sesso,stato civile,titolo di studio}_j] \ .$$

Viceversa, rilasciando una tabella marginale si rilasciano informazioni parziali sulle sue ($2^{p-k} - 1$) tabelle madri. Queste informazioni sono i limiti inferiori e superiori per ogni cella delle tabelle madri. Perciò, dato un insieme di tabelle rilasciate, in teoria, è possibile ricavare un intervallo di esistenza, [$LB$,$UB$], per ogni generica cella $C$ delle loro tabelle madri. Ad esempio, si supponga che vengano rilasciate le marginali in Tabella 4.2. La prima cosa che risulta evidente è che il totale {10} è ridondante, perché

date le marginali unidimensionali la tabella figlia zero-dimensionale è nota. Gli intervalli di esistenza per questa tabella bidimensionale possono essere calcolati come soluzioni di problemi di ottimizzazione lineare, analoghi a quelli per gli intervalli di esistenza per celle soppresse (vedi Paragrafo 3.1.4). Gli intervalli di esistenza per la Tabella 4.1 si possono facilmente ricavare e sono mostrati in Tabella 4.3.

**Tabella 4.3 Intervalli di esistenza desunti dalle marginali in Tabella 4.2**

| $C_1 \backslash C_2$ | $C_{21}$ | $C_{22}$ | Totale |
|---|---|---|---|
| $C_{11}$ | [1,4] | [3,6] | 7 |
| $C_{12}$ | [0,3] | [0,3] | 3 |
| Totale | 4 | 6 | 10 |

La valutazione del rischio derivante del rilascio di tabelle appartenenti all'insieme di tutte le possibili tabelle formabili con *p* variabili, deve tenere conto della dipendenza esistente tra le tabelle e delle informazioni sulle tabelle non rilasciate che è possibile inferire. Non è, cioè, sufficiente considerare il rischio individuale delle sole tabelle rilasciate. Se la misura del rischio per le celle adottata è sub-additiva (vedi Paragrafo 2.4), una misura di rischio globale può basarsi sul rischio delle celle della tabella massima. Essendo l'insieme considerato gerarchico e la misura di rischio sub-additiva, se le celle della tabella massima non sono a rischio, non lo saranno nemmeno le celle delle tabelle di ordine inferiore. Una misura di rischio globale che sarà utilizzata è la minima ampiezza degli intervalli di esistenza che è possibile ricostruire per le celle a rischio della tabella massima. Per un insieme di tabelle $\mathcal{D}$, detti $UB(C,\mathcal{D})$ e $LB(C,\mathcal{D})$ il limite inferiore e superiore calcolabili per una generica cella della tabella massima, *C*, questa misura del rischio è

$$R(\mathcal{D}) = -\min\{UB(C,\mathcal{D}) - UB(C,\mathcal{D}) : C \text{ è a rischio}\}.$$

Avendo adottato questa misura del rischio, il problema di calcolare efficientemente gli intervalli di esistenza diventa cruciale. Se la tabella massima ha *p* dimensioni e ogni variabile classificatrice ha $I_j$ modalità, occorre calcolare $2\prod_{j=1}^{p} I_j$ limiti. Questo problema di programmazione lineare è notoriamente NP-completo e, quindi, la sua soluzione richiede un grande sforzo computazionale. Già quando il numero delle variabili classificatrici è maggiore di 2, la complessità computazionale può essere critica, immaginarsi quando le variabili sono decine e le marginali non costituiscono l'intero insieme delle tabelle marginali di una data dimensione. Il metodo del simplesso, che è il metodo più naturale per risolvere problemi di programmazione lineare, calcola singolarmente ciascuno dei limiti cercati ed occorre eseguirne un ciclo completo per ognuno; quindi, non può, realisticamente, essere applicato. Inoltre, questo metodo difficilmente può essere adattato a tener conto di dipendenze tra le marginali rilasciate. Occorre, quindi, ricorrere ad altri metodi.

### 4.3.3.2 Metodi per il calcolo degli intervalli di esistenza delle celle di una tabella dato un insieme di sue marginali

Gli approcci euristici proposti per il calcolo degli intervalli di esistenza per le tabelle di intensità con marginali fissate presentano dei seri inconvenienti e, comunque, sono stati sviluppati per tabelle con non più di quattro dimensioni. Anche l'approccio al problema che utilizza i *network* risulta complesso e non sembra facilmente generalizzabile a tabelle con più di tre dimensioni.

Per le tabelle di frequenza sono stati proposti diversi approcci che possono essere generalizzati a tabelle multidimensionali. Il problema della determinazione di limiti per le celle di una tabella di frequenza *k*-dimensionale con marginali *m*-dimensionali note può essere riportato a quello inerente ai limiti di intersezioni di probabilità con marginali note. I limiti risultanti sono noti come limiti di Fréchet e di Bonferroni; una trattazione probabilistica completa di questi limiti si può trovare in Galambos e Simonelli (1996), mentre i risultati utili per il problema della tutela della riservatezza sono dati in Fienberg (1999) e discussi, per il caso tridimensionale, in Cox (2001).

#### Limiti di Fréchet

Si consideri una tabella di frequenze tridimensionale con generico elemento $n_{i,j,k}$ e con marginali bidimensionali date. Indicando con il simbolo "+" l'operazione di sommatoria rispetto all'indice sostituito, gli elementi delle marginali si indicano con $n_{+,j,k}, n_{i,+,k}, n_{i,i,+}, n_{i,+,+}, n_{+,j,+}, n_{+,+,k}, n_{+,+,+}$. I limiti di Fréchet, come è facile verificare, per l'elemento $n_{i,j,k}$ sono i seguenti:

$$\max\left\{0,(n_{i,j,+}+n_{i,+,k}-n_{i,+,+}),(n_{i,j,+}+n_{+,j,k}-n_{+,j,+}),(n_{i,+,k}+n_{+,j,k}-n_{+,+,k})\right\} \leq$$
$$n_{i,j,k} \leq \min\left\{n_{+,j,k}, n_{i,+,k}, n_{i,j,+}\right\}$$

Fienberg (1999) propone come esempio la tabella tridimensionale mostrata in Tabella 4.4. Le tabelle marginali bidimensionali della Tabella 4.4 sono mostrate in Tabella 4.5.

**Tabella 4.4 Esempio di tabella tridimensionale: reddito, sesso e razza. In parentesi sono indicati i simboli relativi alle modalità**

| | Uomini (U) | | | Donne (D) | | |
|---|---|---|---|---|---|---|
| | Alto(A) | Medio(M) | Basso(B) | Alto(A) | Medio(M) | Basso(B) |
| Bianchi (B) | 96 | 72 | 161 | 186 | 127 | 51 |
| Neri (N) | 10 | 7 | 6 | 11 | 7 | 3 |
| Cinesi (C) | 1 | 1 | 2 | 0 | 1 | 0 |

**Tabella 4.5 Tabelle marginali bidimensionali della Tabella 4.4**

| | U | D |
|---|---|---|
| B | 329 | 364 |
| N | 23 | 21 |
| C | 4 | 1 |

| | A | M | B |
|---|---|---|---|
| B | 282 | 199 | 212 |
| N | 21 | 14 | 9 |
| C | 1 | 2 | 2 |

| | U | D |
|---|---|---|
| A | 107 | 197 |
| M | 80 | 135 |
| B | 169 | 54 |

I limiti di Fréchet per l'elemento $n_{(B,A,U)}$=96, in Tabella 4.4, sono:

$$\max\{0,-82,85,80\} = 85 \le n_{(B,A,U)} \le 107 = \min\{282,329,107\}\ .$$

Se invece di una tabella tridimensionale si considera una tabella *k*-dimensionale, il calcolo dei limiti di esistenza per le celle $n_{i_1,i_2,\ldots,i_k}$ è più complesso. I limiti differiscono a seconda dell'ordine delle tabelle marginali che vengono rilasciate.

Se vengono rilasciate le marginali unidimensionali, i limiti di Fréchet sono:

$$\max\left\{0, n_{i_1,+,\ldots,+} + n_{+,i_2,+,\ldots,+} + \cdots + n_{+,+,\ldots,i_k} - (k-1)n_{+,+,+}\right\} \le$$
$$n_{i_1,i_2,\ldots,i_k} \le \min\left\{n_{i_1,+,\ldots,+}, n_{+,i_2,+,\ldots,+}, n_{+,+,\ldots,i_k}\right\}.$$

La somma di tutte le marginali unidimensionali che appare nel limite inferiore sarà utilizzata ancora e sarà indicata con $S_{1[i_1,i_2,\ldots,i_k]} = n_{i_1,+,\ldots,+} + \cdots + n_{+,+,\ldots,i_k}$. In maniera analoga si possono definire le somme delle marginali di ordine maggiore; siano $I_m = \{(i_1,i_2,\ldots,i_m), i_1 < i_2 < \cdots < i_m\}$ gli insiemi di indici, le somme parziali saranno

$$S_{m[i_1,i_2,\ldots,i_k]} = \sum_{I_m} n_{I_m,+,\cdots,+}$$

dove la sommatoria si estende a tutte le *m*-uple; $S_{0[i_1,i_2,\ldots,i_k]} = n_{+,+,\ldots,+}$ denota il totale generale.

Quando le marginali rilasciate sono tutte quelle di dimensione *m*, con $1 \le m \le (k-1)$, i limiti superiori si ottengono sostituendo le marginali *m*-dimensionali nella solita espressione, quindi risulta:

$$n_{i_1,i_2,\ldots,i_k} \le \min\left\{n_{i_1,i_2,\ldots,i_m,+,\ldots,+}, n_{i_1,i_2,\ldots,i_{m-1},+,i_{m+1},+,\ldots,+}, \cdots, n_{+,\ldots,+,i_{k-m+1},i_{k-m+2},\ldots,i_k}\right\}.$$

Il computo dei limiti superiori si basa sulla fusione delle modalità delle variabili indicatrici. Per il generico valore $n_{i_1,i_2,\ldots,i_k}$, si consideri la tabella $2^k$ in cui le variabili presentano le modalità della cella di interesse, $i_j$, o la modalità $\bar{i}_j$ ottenuta fondendo insieme le modalità diverse da $i_j$. I valori che presentano solo modalità aggregate saranno denotate con una barra in apice, per esempio la marginale $n_{\bar{i}_1,\bar{i}_2,+,\ldots,+,\bar{i}_k} = \bar{n}_{i_1,i_2,+,\ldots,+,i_k}$. Nella Tabella 4.6 sono riportate le tabelle così derivate per l'esempio di Fienberg.

Le tabelle così *dicotomizzate* permettono di esprimere il valore incognito $n_{i_1,i_2,\ldots,i_k}$ come:

$$\bar{n}_{i_1,i_2,\ldots,i_k} = S_{0[i_1,i_2,\ldots,i_k]} - S_{1[i_1,i_2,\ldots,i_k]} + S_{2[i_1,i_2,\ldots,i_k]} - \cdots (-1)^p n_{i_1,i_2,\ldots,i_k}\ .$$

Da questa equazione si possono ricavare i limiti superiori per il valore in questione, che saranno differenti a seconda che *p* sia pari o dispari. Se *p* è pari risulta:

$$n_{i_1,i_2,\ldots,i_k} = -S_{0[i_1,i_2,\ldots,i_k]} + S_{1[i_1,i_2,\ldots,i_k]} - S_{2[i_1,i_2,\ldots,i_k]} + \cdots + \bar{n}_{i_1,i_2,\ldots,i_k} \ge$$
$$-S_{0[i_1,i_2,\ldots,i_k]} + S_{1[i_1,i_2,\ldots,i_k]} - S_{2[i_1,i_2,\ldots,i_k]} + \cdots - S_{(p-1)[i_1,i_2,\ldots,i_k]}.$$

**Tabella 4.6 Tabelle derivate dall'esempio di Tabella 4.4**

| $U$ | | | |
|---|---|---|---|
| | $A$ | $\overline{A}$ | Totale |
| $B$ | $n_{BAU}$=96 | 233 | 329 |
| $\overline{B}$ | 11 | 16 | 27 |
| Totale | 107 | 249 | 356 |

| $\overline{U} = D$ | | | |
|---|---|---|---|
| | $A$ | $\overline{A}$ | Totale |
| $B$ | 186 | 178 | 364 |
| $\overline{B}$ | 11 | $\overline{n}_{BAU}$=11 | $\overline{n}_{B+U}$=22 |
| Totale | 197 | $\overline{n}_{+AU}$=189 | $\overline{n}_{++U}$=386 |

Marginali unidimensionali

| | | | | | |
|---|---|---|---|---|---|
| $B$ | 693 | $A$ | 304 | $U$ | 356 |
| $\overline{B}$ | 49 | $\overline{A}$ | 438 | $\overline{U}$ | 386 |
| Totale | 742 | Totale | 742 | Totale | 742 |

Se $p$ è dispari, considerando che $\overline{n}_{i_1,i_2,\ldots,i_p} \le \min_{x=1,2,\ldots,p}\left\{\overline{n}_{+,i_2,\ldots,i_p}, \overline{n}_{j_1,+,i_3,\ldots,i_p}, \overline{n}_{j_1,j_2,i_3,\ldots,+}\right\}$ risulta:

$$n_{i_1,i_2,\ldots,i_k} = S_{0[i_1,i_2,\ldots,i_k]} - S_{1[i_1,i_2,\ldots,i_k]} + S_{2[i_1,i_2,\ldots,i_k]} - \cdots - \overline{n}_{i_1,i_2,\ldots,i_k} \ge$$
$$S_{0[i_1,i_2,\ldots,i_k]} - S_{1[i_1,i_2,\ldots,i_k]} + S_{2[i_1,i_2,\ldots,i_k]} - \cdots + S_{(p-1)[i_1,i_2,\ldots,i_k]} - \min_{x=1,2,\ldots,p}\left\{\overline{n}_{+,i_2,\ldots,i_p}, \overline{n}_{j_1,+,i_3,\ldots,i_p}, \overline{n}_{j_1,j_2,i_3,\ldots,+}\right\}.$$

Conoscendo le marginali fino all'ordine $m$, i limiti superiori ed inferiori possono essere ricavati troncando le somme all'$m$-esimo termine, utilizzando una o l'altra formula a seconda che $m$ sia pari o dispari.

## Limiti di Bonferroni

I limiti di "Bonferroni" sono così chiamati perché derivati dalle note formule del Bonferroni per le probabilità di esclusione-inclusione. Si ricorda che, se $A_i$, $i=1,2,\ldots,k$ sono eventi indipendenti e $\overline{A}_i$ gli eventi complementari, la formula di base del Bonferroni è

$$P(\bigcap_{i=1}^{k}\overline{A}_i) = 1 - \sum_{i=1}^{k} P(A_i) + \sum_{i=1}^{k}\sum_{j\neq i} P(A_i \cap A_j) - \cdots$$

Dette le somme parziali delle probabilità dell'intersezione di $m$ eventi:

$$Q_m = \sum_{i_1 \neq i_2 \neq \cdots \neq i_m} P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}),$$

la formula del Bonferroni può essere scritta come:

$$P(\bigcap_{i=1}^{k}\overline{A}_i) = 1 - Q_1 + Q_2 - Q_3 + \cdots + (-1)^k Q_k \ .$$

Questa equazione può essere reinterpretata in termini di frequenze sostituendo alle somme $Q_m$ quelle delle frequenze marginali delle modalità $\overline{n}_{i_1,i_2,+,\ldots,+,i_k}$, definite sopra. Siano $\overline{S}_{m[i_1,i_2,\ldots,i_k]} = S_{m[\overline{i}_1,\overline{i}_2,\ldots,\overline{i}_k]}$ tali somme, si ha:

$$n_{i_1,i_2,\ldots,i_k} = \sum_{m=0}^{k} (-1)^m \overline{S}_{m[i_1,i_2,\ldots,i_k]} \; .$$

Essendo la sequenza degli $\overline{S}_{m[i_1,i_2,\ldots,i_k]}$ decrescente, troncando la somma ad un termine pari si ottiene un limite superiore mentre troncando la somma ad un termine dispari si ottiene un limite inferiore.

Si consideri l'esempio in Tabella 4.4, in cui vengono rilasciate le tabelle marginali bidimensionali della tabella tridimensionale. Poiché *k*=3 è dispari i limiti inferiori di Bonferroni coinvolgono solo le marginali unidimensionali e saranno ridondanti. Il limite superiore per $n_{BAU}$ può essere calcolato considerando che

$$\overline{n}_{BAU} = S_{0[BAU]} - S_{1[BAU]} + S_{2[BAU]} - n_{BAU}$$

e quindi

$$n_{BAU} \le \overline{n}_{BAU} + n_{BAU} = S_{0[BAU]} - S_{1[BAU]} + S_{2[BAU]} \; .$$

Sostituendo i valori in Tabella 4.4 nella formula si ha:

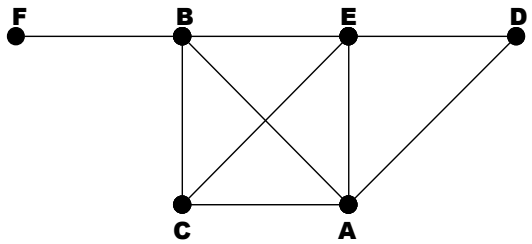
$$n_{BAU} \le 742 - 1353 + 718 = 107 \; .$$

Questo valore è lo stesso di quello ottenuto sopra come limite di Fréchet ma, come nota anche Cox (2001), nell'articolo di Fienberg (1999) ne è erroneamente indicato uno inferiore.

## Modelli log-lineari scomponibili

Il calcolo degli intervalli di esistenza di Fréchet è strettamente legato alla stima dei modelli log-lineari, Fienberg *et al.* (1998), si veda anche Fienberg (1999) e Dobra e Fienberg (2000), hanno generalizzato il calcolo di questi limiti al caso in cui l'insieme delle tabelle marginali rilasciate è associato ad un modello log-lineare scomponibile (più esattamente: l'insieme delle tabelle marginali rilasciate costituisce l'insieme minimale delle statistiche sufficienti per un modello log-lineare scomponibile) (ad esempio, si veda Bishop *et al.*, 1975). Un modello grafico è scomponibile quando due sottoinsiemi di esso sono indipendenti condizionatamente ad un altro sottoinsieme. Per esempio, si supponga che per una tabella a sei dimensioni [ABCDEF], vengano rilasciate le marginali [ABCE], [ADE] e [BF]. Il grafo di indipendenza indotto da queste marginali è riportato in Figura 4.1.

**Figura 4.1 Grafo di indipendenza indotto dalle marginali [ABCE], [ADE] e [BF].**



Da questo grafo risulta evidente che: le marginali [F] e [ADE] sono indipendenti condizionalmente alle marginali [BC] e le marginali [D] e [BCF] sono indipendenti condizionalmente a [AE]. Perciò, il modello è scomponibile nelle tre componenti [BCF], [ADE] e [BF] e nei *separatori* [BC] e [AE]. Una trattazione più approfondita dei modelli (grafici) scomponibili va oltre lo scopo di questo capitolo e si rimandano i lettori interessati agli articoli citati sopra e alle referenze in essi contenute. Per quanto riguarda il calcolo dei limiti, risulta che, se le tabelle marginali sono associate ad un modello grafico scomponibile, i limiti calcolabili per le celle della tabella massima si possono esprimere come funzioni esplicite delle tabelle marginali e non sono migliorabili (Dobra, 2000 e Dobra e Fienberg, 2001). Grazie a questo risultato, il calcolo dei limiti di Fréchet può essere esteso ai casi in cui l'insieme delle tabelle marginali rilasciate è composto da tabelle di diverse dimensioni, e non tutte e sole le tabelle marginali di una certa dimensione. I limiti ottenuti con questo metodo sono non migliorabili e sono uguali i limiti di Fréchet nei casi contemplati. Nei casi in cui i modelli corrispondenti all'insieme di tabelle rilasciate non siano scomponibili, i limiti così calcolati sono migliorabili. Per calcolare limiti migliori si può ricorrere ad algoritmi iterativi, come lo *shuttle algorithm* e le sue generalizzazioni.

## Lo *shuttle algorithm*

Buzzigoli e Giusti (1999) hanno proposto un algoritmo euristico, chiamato *shuttle algorithm*, per il calcolo dei limiti di una tabella *k*-dimensionale quando tutte le marginali (*k-1*)-dimensionali sono note. Lo *shuttle algorithm* si basa sul principio che se una cella assume il suo valore massimo (minimo) allora tutte le celle della riga e della colonna corrispondenti devono assumere il loro valore minimo (massimo). L'algoritmo procede iterativamente alternando il calcolo dei valori massimi e minimi, aggiornando i limiti con i valori calcolati ai passi precedenti. Per le tabelle tridimensionali, Cox (2001) raffronta i limiti di Fréchet e di Bonferroni in Fienberg (1999) con quelli calcolati con lo *shuttle algorithm*, concludendo che lo *shuttle algorithm* dà i limiti migliori, anche se non sempre minimali.

Si supponga che siano rilasciate tutte le tabelle marginali (*k*-1)-dimensionali di una tabella *k-dimensionale*. Per ogni valore $n_{i_1,i_2,\ldots,i_k}$ della tabella massima, ad ogni passo *s*, l'algoritmo calcola nuovi limiti inferiori e superiori, che indicheremo con $n^{L_s}_{i_1,i_2,\ldots,i_k}$ e $n^{U_s}_{i_1,i_2,\ldots,i_k}$, rispettivamente. I limiti vengono inizializzati semplicemente come

$$n^{L_0}_{i_1,i_2,\ldots,i_k} = 0 \le n_{i_1,i_2,\ldots,i_k} \le \min\left\{ n_{+,i_2,\ldots,i_k}, n_{i_1,+,i_3,\ldots,i_k}, \ldots, n_{i_1,i_2,\ldots,+} \right\} = n^{U_0}_{i_1,i_2,\ldots,i_k} .$$

Se $n_{i,i_2,\ldots,i_k} = n^{U_s}_{i,i_2,\ldots,i_k}$, per $i \neq i_1$, è evidente che

$$n_{i_1,i_2,\ldots,i_k} \ge n_{+,i_2,\ldots,i_k} - \sum_{i \neq i_1} n^{U_s}_{i,i_2,\ldots,i_k} .$$

Estendendo questa disuguaglianza a tutte le *k* dimensioni, i limiti inferiori al passo (*s*+1) vengono aggiornati come

$$n_{i_1,i_2,\ldots,i_k}^{L_{(s+1)}} = \min\left\{0, \left(n_{+,i_2,\ldots,i_k} - \sum_{i\neq i_1} n_{i,i_2,\ldots,i_k}^{U_s}\right), \left(n_{i_1,+,i_3,\ldots,i_k} - \sum_{i\neq i_2} n_{i_1,i,i_3,\ldots,i_k}^{U_s}\right), \ldots, \left(n_{i_1,i_2,\ldots,i_{(k-1)},+} - \sum_{i\neq i_k} n_{i,i_2,\ldots,i_{(k-1)},i}^{U_s}\right)\right\}.$$

Analogamente, i limiti superiori al passo ($s$+1) vengono aggiornati come

$$n_{i_1,\ldots,i_k}^{U_{(s+1)}} = \min\left\{n_{i_1,\ldots,i_k}^{U_s}, \left(n_{+,i_2,\ldots,i_k} - \sum_{i\neq i_1} n_{i,i_2,\ldots,i_k}^{L_{(s+1)}}\right), \left(n_{i_1,+,i_3,\ldots,i_k} - \sum_{i\neq i_2} n_{i_1,i,i_3,\ldots,i_k}^{L_{(s+1)}}\right), \ldots, \left(n_{i_1,\ldots,i_{(k-1)},+} - \sum_{i\neq i_k} n_{i_1,\ldots,i_{(k-1)},i}^{L_{(s+1)}}\right)\right\}.$$

Questo algoritmo è di facile implementazione e converge in un numero finito di passi (Buzzigoli e Giusti, 1999), Cox (2000) nota che questo algoritmo porta a limiti non peggiori di quelli di Fréchet e di Bonferroni e che, però, può dare intervalli che contengono valori non ammissibili.

Dobra e Fienberg (2001) propongono una generalizzazione dello *shuttle algorithm* per un insieme qualsiasi di tabelle marginali di una tabella $k$-dimensionale. Questo algoritmo si basa sulle stesse disuguaglianze sfruttate dallo *shuttle algorithm*, però applicate alle tabelle *dicotomizzate* fondendo le modalità, come fatto sopra per il calcolo dei limiti di Fréchet. Data una generica cella $n_1$, si consideri una marginale corrispondente $n_3$ e la cella unificata $n_2 = n_3 - n_1$. I limiti inferiori e superiori vengono aggiornati come

$$n_1^{L_s} = \max(n_1^{L_{(s-1)}}, n_2^{L_s} - n_3^{U_s})$$

e

$$n_1^{U_s} = \min(n_1^{U_{(s-1)}}, n_2^{U_s} - n_3^{L_s}).$$

Si noti che questa formulazione è estremamente generica in quanto si sono lasciati indeterminati tutti i valori considerati, mentre durante il corso delle iterazioni alcuni totali marginali possono divenire noti. L'algoritmo termina quando si ottiene una inconsistenza. Maggiori dettagli su questo algoritmo possono essere trovati in Dobra (2001).

## 4.4 Sistemi automatici per la protezione soppressiva di Swdd

Come detto sopra, la protezione soppressiva di una tabella con molte dimensioni necessita di procedure automatiche. Karr *et al*. (2002) e Dobra *et al*. (2002) ne descrivono due: uno per la protezione statica ed uno per la protezione dinamica. Entrambi i sistemi sono formulati in maniera abbastanza generica da consentire l'utilizzo di metodologie diverse al loro interno. In entrambi i sistemi i dati vengono protetti sopprimendo completamente le tabelle il cui rilascio comporta un rischio eccessivo. In questo modo vengono eliminati i problemi computazionali legati al calcolo delle soppressioni secondarie (vedi Paragrafo 3.1.2). Quindi, la protezione statica consiste nel determinare un sottoinsieme di tabelle marginali che possano essere rilasciate, mentre la protezione dinamica consiste nel valutare, considerando l'informazione già rilasciata, se una tabella richiesta possa essere rilasciata o meno.

### 4.4.1 Protezione soppressiva dei Swdd statici: selezione di un insieme di tabelle di frequenza rilasciabili

Dato un insieme di variabili classificatrici di cui si conoscono le frequenze e per cui la tabella di massima dimensione contiene celle a rischio, il problema relativo ai Swdd statici è quello di determinare un sottoinsieme di tabelle marginali che possano essere rilasciate senza compromettere la riservatezza. Quando il numero delle variabili è elevato non è pensabile di applicare gli algoritmi di soppressione complementare visti sopra: più realisticamente conviene sopprimere completamente alcune tabelle. La protezione soppressiva statica dell'Swdd si riduce, così, ad individuare un insieme di tabelle marginali che possano essere rilasciate senza essere protette.

La scelta di pubblicare solo un sottoinsieme di tabelle marginali è la più naturale e comunemente utilizzata. In genere, questo sottoinsieme viene selezionato manualmente, in base a criteri di opportunità e convenienza, da persone che conoscono i dati. Questa procedura, però, è arbitraria, non sempre replicabile e, spesso, iperprotettiva; l' *Optimal Tabular Release* (Otr) è un sistema automatico per la determinazione di questo sottoinsieme, sviluppato presso il *National Institute of Statistical Sciences* (Niss) nell'ambito del *Digital Government* (Dg) *project*, e sarà descritto brevemente di seguito.

#### *Optimal Tabular Releases*

Nell'*Optimal Tabular Release* (Dobra *et al.*, 2001) l'insieme ottimo da rilasciare è determinato secondo l'approccio rischio-utilità. Supponendo che almeno una cella della tabella massima sia a rischio, tra tutti i possibili sottoinsiemi di tabelle che garantiscono un rischio globale inferiore ad una soglia $\alpha$, viene scelto quello, $\mathcal{D}^*$, che ha massima utilità.

Come misura globale del rischio per un insieme di tabelle $\mathcal{D}$, viene presa l'ampiezza minima dell'intervallo di esistenza calcolabile per le celle a rischio della tabella massima. Detti $UB(C, \mathcal{D})$ e $LB(C, \mathcal{D})$ il limite inferiore e superiore calcolabili per una generica cella della tabella massima, $C$, la misura del rischio utilizzata è

$$R(\mathcal{D}) = -\min\{UB(C,\mathcal{D}) - LB(C,\mathcal{D}) : \#(C) < k\}$$

dove $\#(C)$ indica il numero di contributi della cella $C$ e $k$ la soglia di rischio accettabile, in genere si pone $k>2$. La scelta di questa misura globale di rischio assicura che la soglia venga rispettata per tutte le celle, però richiede il calcolo degli intervalli di esistenza per tutte le celle della tabella massima.

L'utilità di un insieme di tabelle può essere misurata in diversi modi. Una delle misure più semplici, che è quella utilizzata per esemplificare l'OTR, è il numero di tabelle rilasciate:

$$U(\mathcal{D}) = \#(\mathcal{D})$$

Come altre misure dell'utilità che si possono considerare, per esempio: il numero di celle indipendenti rilasciate (quantificabile con i gradi di libertà che esse rappresentano) oppure l'errore quadratico medio, ottenuto stimando le celle della tabella massima dai dati rilasciati, per esempio con un modello log-lineare. In quest'ultimo caso, però, è necessario adoperare il *fitting* proporzionale iterativo (Bishop *et al*., 1975),

che può essere dispendioso in termini computazionali. In certi casi può essere opportuno aggiungere alla funzione di utilità una qualche misura di differenziazione dell'informazione rilasciata, per esempio, che tenga in considerazione l'informazione rilasciata rispetto a ciascuna variabile. Questo potrebbe evitare di rilasciare sottoinsiemi che contengano poche tabelle con certune variabili.

Per l'Otr vengono suggeriti due approcci per il calcolo degli intervalli di esistenza. Il primo consiste nel limitare la ricerca dell'ottimo ai soli insiemi di tabelle marginali associate a modelli grafici scomponibili. Come visto, questa restrizione, oltre a ridurre il numero di sottoinsiemi che è necessario visitare, riduce drasticamente la complessità del calcolo del rischio. Poiché la restrizione degli insiemi rilasciabili alla classe delle statistiche sufficienti minimali di modelli scomponibili non è completamente giustificabile, viene anche considerato un accorgimento alternativo per il calcolo del rischio. Il secondo approccio che viene considerato è un algoritmo *greedy*, di tipo *bottom-up*, che aggiunge tabelle all'insieme $\mathcal{D}^*$ fino a che sia possibile senza violare il vincolo di massimo rischio. Le tabelle sono aggiunte secondo l'ordine determinato da una particolare misura euristica del rischio individuale.

Nel prototipo di Otr realizzato presso il Niss, l'ottimo viene ricercato nella classe dei modelli grafici scomponibili. Il prototipo è stato applicato a dati con 13 variabili classificatrici. Il numero dei modelli scomponibili esistenti per tredici variabili è, comunque, troppo elevato per poter essere affrontato, quindi la soluzione ottima è stata determinata per mezzo del *simulated annealing*, un metodo iterativo di tipo Monte Carlo per la determinazione di un massimo. Maggiori dettagli sull'Otr e sull'implementazione di questi accorgimenti possono essere trovati in Karr *et al.* (2002).

### 4.4.2 Protezione soppressiva dei Swdd dinamici: valutazione *al volo* del rischio di rilasciare tabelle aggiuntive

L'approccio alla tutela della riservatezza per i Swdd dinamici, è necessariamente basato sulla conoscenza accumulata. In questo approccio, i dati sono protetti sequenzialmente in risposta ad una serie di richieste (Hoffman, 1977, Keller-McNulty e Unger, 1998 e Fienberg *et al.*, 1998). L'applicazione della protezione dinamica richiede la definizione di regole per la restrizione dell'insieme delle richieste ammissibili (Hoffman, 1977). In queste restrizioni, il rischio che comporta il rilascio di una nuova tabella richiesta è calcolato sulla base dell'informazione già rilasciata. Più avanti si discuterà sulle implicazioni di considerare l'informazione totale rilasciata o solo quella rilasciata a ciascun utente.

Di seguito andiamo ad illustrare un sistema per la protezione dinamica di Swdd realizzato dai ricercatori del Dg *project*, chiamato *Table Server*.

#### *Table Server*

Un *Table Server* è un sistema che applica la soppressione totale di tabelle dinamicamente, cioè decidendo se rilasciare o meno una tabella marginale richiesta sulla base del rischio calcolato tenendo conto anche delle tabelle marginali precedentemente rilasciate.

All'istante $t$, l'insieme delle tabelle rilasciate, $\mathrm{T}(t)$, contiene tutte le tabelle rilasciate direttamente ed indirettamente (cioè anche le marginali non rilasciate ma figlie di marginali rilasciate). L'insieme $\mathcal{T}(t)$ è completamente individuato dalla *frontiera rilasciata*, $\mathcal{FT}(t)$, che contiene gli elementi massimi di $\mathcal{T}(t)$ (cioè le tabelle marginali rilasciate di cui non siano state rilasciate marginali madri). Avendo adottato una misura di rischio globale, R, ed una soglia di rischio accettabile, $\alpha$, l'insieme delle tabelle rilasciate deve sempre soddisfare $R(\mathrm{T}(t)) \le \alpha$. Quando giunge una richiesta per una tabella **T** non contenuta in $\mathcal{T}(t)$, essa verrà rilasciata se

$$R(\mathrm{T}(t) \cup \mathbf{T}) \le \alpha .$$

Questa formulazione del rischio prevede che gli utenti cooperino tra di loro, quindi il rischio è calcolato su l'informazione rilasciata a tutti gli utenti. Anche per i *Table Server* una misura del rischio globale che può essere adottata è quella dell'ampiezza minima degli intervalli di esistenza calcolabile per le celle a rischio della tabella massima. Usando la quale, una tabella verrà rilasciata se

$$-\min\left\{UB(C,(\mathrm{T}(t) \cup \mathbf{T})) - LB(C,(\mathrm{T}(t) \cup \mathbf{T})) : \#(\mathrm{C}) < \mathrm{k}\right\} \le \alpha .$$

Quando il numero delle variabili nel database è grande, lo sforzo computazionale per il calcolo di questo rischio può essere tale da richiedere tempi troppo lunghi per il rilascio *al volo*. Per rendere più spediti i calcoli è possibile ricorrere allo *shuttle algorithm* e alla sua generalizzazione. Altrimenti, si può pensare di associare ad ogni nuova richiesta anche altre tabelle in modo da formare l'insieme minimale delle statistiche sufficienti di un modello grafico scomponibile che permetta il computo diretto dei limiti di Fréchet con l'approccio di Dobra e Fienberg (*ibidem*).

Il rilascio di una tabella marginale può rendere altre tabelle non più rilasciabili. Da questo punto di vista, la misura del rischio globale additiva vista sopra può essere insufficiente a proteggere il database dall'essere, casualmente o dolosamente, deviato rilasciando informazione di interesse ristretto, impedendo il rilascio di altra informazione di maggior interesse generale. Come osservano anche Karr *et al.* (2002), che chiamano la misura di rischio vista sopra *miope*, può allora essere opportuno aggiungere criteri di differenziazione dell'informazione rilasciata. Un'altra possibilità è quella di attendere che siano sottoposto più interrogazioni per applicare simultaneamente la protezione all'insieme richiesto, rilasciando le tabelle richieste con ritardo.

## 4.5 Laboratori virtuali

Per laboratorio virtuale si intende un sito Web in cui gli utenti abbiano la possibilità di effettuare analisi statistiche su dati. Questi, quindi, sono sistemi applicativi che su richiesta accedono in remoto ad un collezione di record fornendo informazioni sintetiche. I laboratori virtuali interattivi permettono agli utenti di richiedere analisi statistiche on line per mezzo di procedure Html o simili. In risposta ad ogni

interrogazione il server Web chiama alcuni programmi che eseguono le analisi e preparano i dati ottenuti per la visualizzazione o il *download*.

In Internet si possono trovare diversi laboratori virtuali interattivi, per esempio quelli che utilizzano il programma Sda.[13] Tra questi vi è il sito dell'istituto Cattaneo dell'Università di Bologna, "*Cattaneo Web Archive of Italian Election Results and Surveys*",[14] che mette a disposizione degli utenti campioni di dati relativi alle elezioni politiche in Italia negli anni 1990, 1992, 1994 e 1996.

Sempre a titolo esemplificativo, di seguito forniamo l'elenco delle funzioni offerte nei laboratori virtuali da Sda:

- **Esplorazione della documentazione dei dati o del questionario**
  - lista delle variabili
  - descrizione di ogni variabile
- **Analisi dei dati**
  - frequenze e tabulazioni
  - confronto di medie (con errore standard complesso)
  - matrici di correlazione
  - confronto di correlazioni
  - regressione (minimi quadrati ordinari)
  - regressione logit e probit
  - lista dei valori individuali
- **Creazione di nuove variabili**
  - ricodifica di una o più variabili esistenti
  - creazione di nuove variabili come funzioni di quelle esistenti
- **Creazione e scaricamento di un sottoinsieme di variabili o osservazioni**
  - file di dati in formato Ascii
  - definizioni dei dati per Sas, Spss o Stata
  - documentazione per il sottoinsieme

Naturalmente, i laboratori virtuali possono offrire opzioni e funzioni diverse da quelle fornite da Sda. Le funzionalità offerte dovrebbero dipendere dal tipo di dati a disposizione per l'analisi. I laboratori virtuali che necessitano di essere protetti contro violazioni della riservatezza sono quelli che offrono analisi di dati riservati. Ovviamente, in questi siti non deve essere possibile consultare direttamente i dati e le tabelle possono essere protette come nei Swdd.

### 4.5.1 Laboratori virtuali server di programmi degli utenti

I laboratori virtuali interattivi permettono agli utenti di eseguire predeterminate analisi sui dati disponibili. Un altro tipo di laboratorio virtuale è quello in cui gli utenti possono sottoporre loro programmi che vengono fatti girare in remoto. Un laboratorio di

[13] Informazioni sul programma Survey Documentation and Analysis (Sda) e sui laboratori virtuali che lo utilizzano sono reperibili sul *Web* all'indirizzo http://sda.berkeley.edu.
[14] http://sda.berkeley.edu:7502/cattaneo.html

questo tipo è, per esempio, implementato per i dati del *Luxembourg Income Study* (Lis) e del *Luxembourg Employment Study* (Les). Un ulteriore esempio è l'utilizzo dei dati dell'indagine sulla struttura dei salari (*Structure of Earning Survey*) nell'ambito del progetto Piep (si rimanda a Schouten e Cigrang, 2003 per dettagli dell'implementazione). Questi laboratori, in genere, sono predisposti per ricercatori, cioè utenti che hanno necessità di effettuare analisi particolari, quindi l'accesso è riservato ad appartenenti a istituzioni di ricerca e senza fini di lucro. Per accedere ai dati del Lis e del Les, infatti, occorre registrarsi e sottoscrivere una promessa di riservatezza.

L'idea alla base di questo tipo di laboratori è quella di mettere a disposizione degli utenti uno o più pacchetti statistici che effettuino le analisi richieste e di restituire solo i risultati. Gli archivi Lis e Les accettano programmi in Sas, Stata e Spss. Naturalmente, non tutte le funzioni di questi pacchetti saranno richiamabili. Le funzioni che richiamano i valori individuali e quelle che modificano permanentemente i dati in memoria dovranno, evidentemente, essere inibite e, sempre per tutelare la riservatezza dei dati, si possono inibire altre funzioni. Poiché spesso solo in alcuni casi i risultati di alcune analisi statistiche pongono a rischio la riservatezza, è opportuno predisporre dei controlli, automatici e, in seconda battuta, manuali, sui risultati, prima di rilasciare l'*output*.

La teoria della tutela statistica della riservatezza per i laboratori virtuali non è stata ancora affrontata in modo sistematico: mentre in molti casi è vero che la conoscenza di pochi parametri di una popolazione non consente di conoscere i valori delle singole unità, ci sono situazioni in cui questo pericolo esiste. Per esempio, rivelando i coefficienti di una regressione di una variabile sensibile su variabili note quando il coefficiente di correlazione parziale è molto alto. Anche la possibilità di creare un numero di variabili artificiali illimitato potrebbe consentire di regredire una risposta su di un numero di variabili pari al numero delle osservazioni, ottenendo un fitting perfetto. Il problema diventa molto più complesso se si considera che i laboratori virtuali possono fornire molte altre statistiche. In genere, essendo il rischio di violazione totalmente sconosciuto, i laboratori virtuali con dati sensibili vengono protetti perturbando fortemente i dati di origine, per esempio mettendo a disposizione solo un piccolo campione.

## 4.6 Strategie per l'applicazione della tutela statistica della riservatezza ai Swdd

Le metodologie per la tutela statistica della riservatezza nei siti Web possono essere raggruppate in tre classi generali:

1. protezione dei dati di origine (sottocampionamento, scambio dei valori, aggiunta disturbi, eccetera);
2. protezione dei dati richiesti (aggiunta disturbi, soppressione di valori, arrotondamento, eccetera);
3. restrizione delle interrogazioni permesse (per esempio: rifiuto).

I metodi appartenenti a classi diverse non sono necessariamente esclusivi, nel

senso che metodi diversi possono essere applicati simultaneamente. Dei metodi appartenenti alle prime due classi ci siamo già occupati. La restrizione dell'insieme delle richieste ammissibili (Hoffman, 1977 e Keller-McNulty e Unger, 1998) consiste nel definire un insieme di regole per rispondere alle richieste degli utenti. La restrizione più semplice è quella che vieta o consente il rilascio di dati richiesti. Una forma più sofisticata offre anche una terza opzione: *proteggere i dati da rilasciare* (Fienberg *et al.* (1998)). La restrizione dell'insieme delle richieste ammissibili è utile per ridurre il rischi di violazione connessi con le tabelle collegate e l'accumulazione di conoscenza.

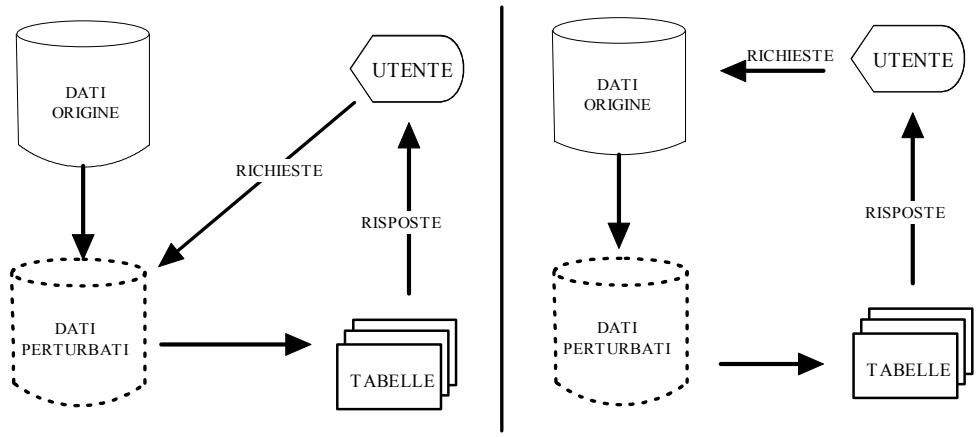
Si può distinguere tra restrizioni *generiche* e restrizioni *specifiche*. Le prime sono definite per tutti gli utenti e, per esempio, limitano la dimensione massima delle tabelle rilasciabili, oppure non consentono di rilasciare tabelle con determinate combinazioni di variabili, eccetera Le restrizioni *specifiche* sono restrizioni più elaborate, basate sull'informazione già rilasciata. Per esempio, possono limitare il numero di tabelle rilasciate ad ogni utente, oppure il rilascio di una certa informazione se ne è già stata richiesta un'altra e così via. I sistemi Otr e *Table Server*, discussi sopra, possono essere visti come metodi per la definizione di restrizioni, rispettivamente generiche e specifiche.

Nelle pubblicazioni cartacee i dati devono necessariamente essere protetti prima della pubblicazione, invece nei Swdd la tutela statistica della riservatezza può essere applicata anche dopo che un'informazione è stata richiesta. Chiameremo PRE la tutela statistica della riservatezza applicata prima che i dati siano richiesti (o messi on line) e POST quella applicata dopo. La protezione PRE si effettua necessariamente off line, la protezione POST può essere applicata on line ma si può applicare off line, rimandando il rilascio dei dati richiesti. Il vantaggio della protezione POST è che può essere applicata adattivamente alle richieste di ogni utente o dell'informazione già rilasciata. Nel primo caso, essa richiede l'identificazione degli utenti e la registrazione delle loro richieste.

La perturbazione dei dati di origine viene generalmente applicata PRE: i dati vengono perturbati e poi messi in linea, cosicché le tabelle vengono costruite su questi dati perturbati. Questo approccio viene in genere adottato per grandi popolazioni che ammettono una sostanziale riduzione dei record disponibili o a cui è possibile applicare la legge dei grandi numeri. In principio, la perturbazione dei dati di origine può anche essere effettuata POST, per esempio estraendo un nuovo sotto-campione ad ogni richiesta. Questa procedura, da una parte, permette la protezione adattiva, per esempio scegliendo la numerosità del campione o la varianza dei disturbi, a seconda del rischio dei dati richiesti; d'altra parte, però, può dare risultati diversi per le stesse richieste, indebolendo l'efficacia della protezione perché un utente potrebbe ottenere una stima abbastanza precisa del valore vero ripetendo la richiesta più volte. In Figura 4.2 vengono mostrati gli schemi di applicazione della perturbazione dei dati di origine PRE e POST.

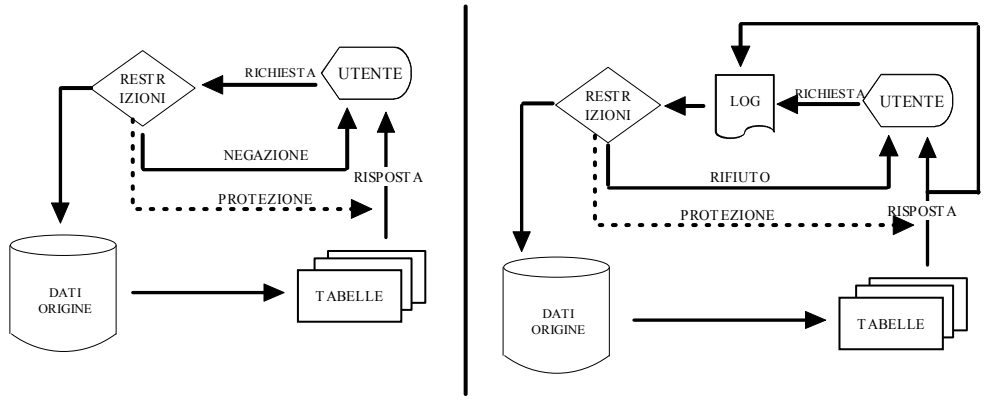
L'applicazione delle restrizioni *generiche* dovrebbe essere considerata PRE anche se viene applicata dopo che i dati vengono richiesti, in quanto le restrizioni sono definite prima che i dati vengano richiesti. Le restrizioni specifiche devono essere applicate POST e richiedono la registrazione dell'utente e la registrazione della sua attività.

**Figura 4.2 Applicazione della perturbazione dei dati di origine ad un Swdd. Applicazione PRE a sinistra e POST a destra.**




Le restrizioni specifiche possono essere poco efficaci contro coalizioni di intrusi, però, se implementate in modo opportuno, esse possono ridurre la necessità di perturbare i dati. Gli schemi di applicazione PRE e POST delle restrizioni delle richieste sono mostrati in Figura 4.3.

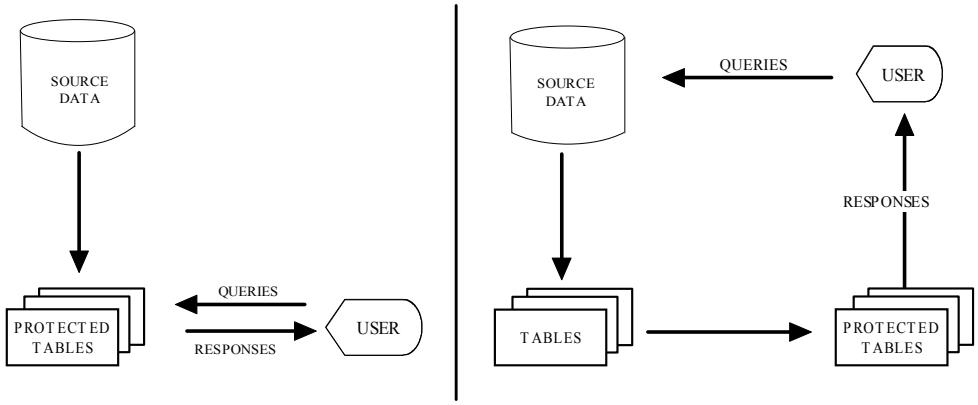
**Figura 4.3 Applicazione della restrizione delle richieste. Applicazione PRE a sinistra e POST a destra**




La protezione dei dati richiesti può essere applicata sia PRE (siti statici) che POST (siti dinamici). L'aggiunta di disturbi ai valori delle celle POST può presentare il problema di richieste ripetute, come nel caso dell'aggiunta di disturbo ai dati di origine. La soppressione di celle POST deve essere registrata e ripetuta coerentemente nelle successive soppressioni; per esempio, se il valore di una cella è stato rilasciato, non può essere soppresso successivamente, viceversa, una volta soppresso deve essere sempre soppresso nei rilasci successivi. Come accennato sopra, la soppressione di singole celle *al volo* non sembra praticabile per la complessità e lunghezza dei calcoli necessari all'individuazione delle soppressioni complementari. L'arrotondamento dei valori delle celle con base costante non sembra presentare particolari controindicazioni, però nella

maggioranza dei casi questa tecnica non garantisce una buona protezione. Le tecniche di arrotondamento più efficaci, come l'arrotondamento controllato, presentano il problema di non essere consistenti per tabelle diverse, quindi perdono di efficacia se applicate a tabelle sovrapposte. Per quanto si sa, la simulazione dei dati viene applicata solo per dati elementari, anche se il suo impiego per dati tabellari potrebbe essere di qualche vantaggio (si veda, Fienberg e Makov, 2001). La simulazione è ancora sotto studio è non sembra essere stata ancora adottata in via ufficiale ma sembra una metodologia promettente, specialmente se applicata per la protezione dei laboratori virtuali. In Figura 4.4 sono mostrati gli schemi di applicazione PRE e POST della protezione dei dati richiesti.

**Figura 4.4 Applicazione della protezione dei dati richiesti, PRE a sinistra e POST a destra**



# PARTE TERZA

## La tutela statistica della riservatezza per dati individuali

# Capitolo 5. Il rilascio dei dati individuali(*)

## 5.1 Introduzione

In questa terza parte del manuale ci occupiamo del rilascio di dati individuali (dati elementari o microdati) intendendo con ciò un tipico file sequenziale in cui ogni record è riferito ad una singola unità statistica e contiene il valore delle variabili rilevate da un'indagine. In pratica, un file di dati individuali può essere considerato il prodotto finale di una rilevazione statistica prima della fase della diffusione, ossia l'archivio di dati dal quale vengono prodotti i risultati da rilasciare all'esterno (del sistema statistico nazionale), siano esse stime di indici, tabelle sintetiche o altro, compreso un file di dati individuali.

Anche nel caso dei dati individuali assumeremo che il concetto di "violazione della riservatezza" sia associato a quello di "rischio di identificazione" (*re-identification*) legato al rilascio di informazione statistica. L'interpretazione che in questo modo viene data dei dettami legislativi, ci consente di trattare la materia dal punto di vista statistico (*statistical disclosure*) distinguendo per tipologia di rilascio (dati aggregati o dati individuali), di dati (sociali o economici) e di utenti.

Dedicheremo il Capitolo 6 al caso in cui gli interessati da tutelare, ossia i soggetti cui sono riferite le informazioni raccolte dalla rilevazione statistica, sono le persone fisiche e il Capitolo 7 al caso delle imprese. Dal punto di vista statistico, le peculiarità di questa distinzione verranno specificate nei rispettivi capitoli. Osserviamo qui che la legislazione specifica in materia di trattamento di dati personali non discrimina, per quanto riguarda il rilascio di informazione statistica, fra le persone fisiche e le imprese ma nella pratica le cautele nei due casi sono estremamente diversificate. Inoltre, nel caso delle imprese, occorre tenere conto delle norme sulla concorrenza e sul segreto industriale che impongono alla statistica ufficiale dei comportamenti di neutralità. In pratica, si deve porre attenzione a che i dati rilasciati non mettano un'impresa in condizioni di privilegio informativo rispetto a una sua concorrente. I metodi di protezione verranno, invece, descritti nel Capitolo 8.

In questo capitolo introduttivo daremo una descrizione generale dell'ambito in cui è inquadrato il problema dal punto di vista statistico, cominciando con il definire più dettagliatamente cosa si intende per dati individuali nella fase di diffusione di informazione statistica (Paragrafo 5.2). Successivamente, definiremo il problema della violazione della tutela della riservatezza per tipologia dei dati individuali (Paragrafo 5.3). Quindi, verranno distinti i diversi possibili destinatari di questa forma di rilascio ipotizzando anche possibili scenari in cui possono essere tentate violazioni del segreto statistico (Paragrafo 5.4).

## 5.2 I dati individuali

Consideriamo l'insieme dei dati che si intende rilasciare come una matrice

(*) Capitolo redatto da Giovanni Seri eccetto i paragrafi 5.4 e 5.5 redatti rispettivamente da Luisa Franconi e da Alessandra Capobianchi

$\mathbf{A} = \{a_{ij}\} = (\mathbf{X}_{(n,k)}, \mathbf{C}_{(n,q)})$ dove il generico elemento $a_{ij}$ è il valore della $j-ma$ variabile assunta dalla $i-ma$ unità statistica. Fino a questo momento i dati contenuti nella matrice $\mathbf{A}$ non hanno subito nessun trattamento ai fini della tutela statistica della riservatezza.

Le righe della matrice (o i record del file) corrispondono, pertanto alle $n$ unità statistiche campionarie rilevate da un'indagine su una popolazione $U$ ampiezza $N$ ($N > n$) individui. Le colonne della matrice $\mathbf{A}$ rappresentano, invece, le variabili da rilasciare. Ai fini della tutela della riservatezza statistica la matrice $\mathbf{A}$ è partizionata in due sottomatrici che raccolgono due differenti gruppi di variabili:

matrice $\mathbf{X}$: $k$ "variabili chiave" o "identificativi indiretti";

matrice C: $q$ "variabili confidenziali" o "riservate".

Per tale distinzione la matrice $\mathbf{A}$ si presenta nel seguente modo:

$$\mathbf{A} = \begin{pmatrix} x_{11} & \dots & x_{1k} & c_{11} & \dots & c_{1q} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ x_{i1} & \dots & x_{ik} & c_{i1} & \dots & c_{iq} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{nk} & c_{n1} & \dots & c_{nq} \end{pmatrix}.$$

Prima di descrivere la matrice precisiamo che, per quanto riguarda gli aspetti statistici trattati nel seguito, solo le variabili chiave hanno rilevanza; pertanto, considereremo equivalente riferirsi alla sola matrice $\mathbf{X}$ piuttosto che alla matrice $\mathbf{A}$.

Le unità statistiche possono essere persone, famiglie, imprese o altre entità. Senza perdere in generalità possiamo considerare le unità statistiche come i soggetti il cui diritto alla riservatezza deve essere tutelato o anche, più sinteticamente, come "rispondenti". Può capitare, infatti, che l'unità statistica non sia un soggetto da tutelare o che le notizie non siano fornite direttamente dagli interessati, ma che il contenuto dei dati sia comunque riservato. Ad esempio, si consideri un file in cui le unità statistiche sono le merci esportate da una nazione e che, quindi, i dati siano rilevati dai registri doganali. Una merce, in generale, non è portatrice del diritto alla tutela della riservatezza, ma, se la classificazione dei prodotti è sufficientemente dettagliata, il produttore di certi articoli può facilmente essere unico. Di conseguenza, tutte le informazioni riferite a quel prodotto sono automaticamente associate al produttore che è invece un soggetto da tutelare. Ancora, il caso di un file in cui le unità statistiche sono le famiglie dove i soggetti da tutelare sono i singoli componenti più che la famiglia stessa.

Va osservato, inoltre, che possono sussistere relazioni di dipendenza fra le unità statistiche di cui si deve tenere conto. Un esempio è stato appena citato e riguarda la relazione gerarchica fra un individuo e la famiglia a cui appartiene. Un'informazione sulla famiglia può contribuire a identificare un suo componente, come pure un'informazione su un individuo può contribuire all'identificazione di un altro componente della stessa famiglia. Analogamente per le imprese può essere identificata una struttura gerarchica che mette in relazione imprese appartenenti ad uno stesso gruppo proprietario.

Infine, sempre relativamente alle unità statistiche, il problema del rilascio di un

file di microdati viene normalmente trattato dal punto di vista statistico con riferimento al rilascio di "collezioni campionarie" di dati elementari in quanto il rischio di violazione associato al rilascio di archivi censuari è generalmente ritenuto troppo elevato. D'altro canto la normativa è conforme a questa impostazione considerando la possibilità di rilascio di dati elementari solo sotto la forma di "collezioni campionarie" a prescindere che la rilevazione abbia avuto carattere censuario o meno. Le stesse considerazioni valgono anche per quelle indagini, più frequenti in ambito economico, che possono essere considerate parzialmente censuarie poiché alcune unità statistiche (ad esempio tutte le imprese con più di 50 addetti nel settore dei servizi) vengono incluse nel campione con probabilità 1.

Ovviamente, nell'ambito di un manuale metodologico, come vuole essere questo, non hanno interesse quei casi di rilascio di informazioni regolati da norme specifiche e che quindi non richiedono interventi dal punto di vista statistico come ad esempio il trasferimento di dati all'ufficio statistica della Comunità Europea in adempimento ai Regolamenti ufficiali che governano talune indagini.

Per quanto riguarda le colonne della matrice $\mathbf{A}$, assumiamo, come è già stato fatto in precedenza, che gli *identificativi diretti* (o *identificativi*), ossia quelle caratteristiche quali il nome o la ragione sociale, l'indirizzo, il codice fiscale, eccetera che rendono l'unità statistica unica e quindi riconoscibile nella popolazione, siano sempre esclusi dal file di microdati che si intende rilasciare. Questa operazione a volte viene indicata come *anonimizzazione* di un file di microdati, ma vedremo in seguito che il processo di anonimizzazione è più complesso e consiste nell'eliminazione dei possibili riferimenti individuali, sia diretti che indiretti.

Per "variabili chiave" o "identificativi indiretti" intendiamo quelle variabili le cui modalità, da sole o in combinazione con le modalità di altre variabili chiave associate ad una unità statistica, contengono elementi che contribuiscono alla possibile identificazione dell'unità statistica stessa. In questa categoria rientrano quelle variabili che sono facilmente disponibili a chi volesse tentare l'identificazione di una unità statistica, ad esempio: l'età, la professione o il luogo di residenza per gli individui, oppure l'attività economica e la collocazione geografica per le imprese. Un laureato in statistica in un piccolo comune di 300 abitanti della Toscana è probabilmente un caso unico nella popolazione $U$ e, quindi, riconoscibile da chiunque sia in grado di associare le due informazioni sul titolo di studio e il luogo di residenza a nome e cognome di quella unità statistica (ad esempio un suo conoscente impiegato nell'ufficio statistico dello stesso comune). In tale modo sarebbero rese note a chi effettua l'identificazione le informazioni contenute nelle q variabili "riservate" dell'indagine relative a quella persona. D'altro canto, riducendo il dettaglio dell'informazione geografica, l'informazione rilasciata sulla stessa unità potrebbe essere: "un laureato in statistica che vive in Toscana" e la sua identificazione risulterebbe certamente operazione ben più complessa. Nella stessa categoria di variabili chiave vanno considerate anche quelle variabili che, pur non essendo facilmente disponibili al destinatario del file rilasciato, consentono di acquisire elementi per l'identificazione. Un esempio tipico è il fatturato di un'impresa, che è rappresentativo (variabile *proxy*) della dimensione dell'impresa stessa. Dati il tipo di attività economica e la collocazione geografica di un'impresa, la sua dimensione può spesso consentire l'identificazione, specie per le grandi imprese. Si

pensi ad esempio alla Benetton nel settore tessile in Veneto o alla Fiat nel settore fabbricazione autovetture in Piemonte (o perfino in Italia). Analogo ragionamento è valido per le variabili quantitative correlate con la dimensione dell'impresa (numero di addetti, costi, eccetera). Anche per gli individui una variabile come il reddito può costituire un identificativo indiretto molto efficace (soprattutto se associato alla professione). Per questo, in genere, variabili particolarmente sensibili come il reddito degli individui vengono rilasciate solo previa adeguata riduzione in classi o addirittura escluse dal rilascio.

Le variabili "riservate", invece, sono quelle reperibili esclusivamente nel file rilasciato e che non contengono informazioni utilizzabili per identificare le unità statistiche. Ad esempio, per un'impresa può essere considerata riservata una variabile come la "Quota di fatturato destinata a ricerca e sviluppo", il "Paese di destinazione delle esportazioni" o anche il "Numero di presenze stagionali" per gli esercizi turistici. Per le persone gli esempi vanno da variabili come "Spese per l'acquisto di elettrodomestici" ad altre che toccano argomenti molto delicati quali la salute come "Causa del ricovero".

Nell'ambito delle indagini sociali, dove le unità statistiche sono generalmente persone fisiche, le variabili chiave sono state talvolta definite "pubbliche" per il fatto che gli identificativi indiretti sono perlopiù informazioni contenute in registri pubblici come l'Anagrafe (data di nascita, residenza, eccetera).

Con l'introduzione della Legge 675/1996 il concetto di *dato pubblico* ha assunto il significato di "dati contenuti o provenienti da pubblici registri, elenchi, atti o documenti conoscibili da chiunque" ma per i nostri fini crediamo sia meglio definito il concetto di variabile chiave.

Un aspetto da considerare è che, in alcuni casi, informazioni su un singolo soggetto possono essere acquisite in forma nominativa dietro richiesta specifica agli opportuni organismi amministrativi, mentre le stesse informazioni eventualmente rilasciate da un Istituto nazionale di statistica sarebbero al più rintracciabili in un archivio non nominativo. Si pensi alle variabili contenute nei bilanci delle imprese. Queste sono disponibili mediante visura presso i Tribunali, tuttavia non possono essere diffuse da un Istituto nazionale di statistica che le ha rilevate direttamente presso l'impresa durante una rilevazione proprio perché la finalità statistica non contempla questo tipo di esigenza ed è, pertanto, soggetta al segreto statistico. Questo contribuisce ad ipotizzare che un intruso intenzionato a scoprire informazioni su un individuo o un'impresa preferisca richiederle direttamente se disponibili presso qualche registro, piuttosto che tentare di rintracciarle in un archivio non nominativo costituito per motivi statistici.

## 5.3 Violazione della riservatezza nel caso del rilascio di dati individuali

In questo paragrafo formalizziamo il concetto di violazione della riservatezza nel caso di rilascio di dati individuali in base al quale viene definita una misura del rischio che si verifichino delle violazioni e conseguentemente adottate misure statistiche a protezione dei dati.

Rispetto a ogni altra forma di rilascio di informazione statistica, il rilascio di un file di dati elementari fornisce il maggior contenuto informativo sia dal punto di vista dell'analisi statistica che da quello della violazione della riservatezza. Il file che stiamo considerando, almeno fino a questo punto, non ha subito ancora nessuna operazione, pertanto mantiene intatte le proprie peculiarità.

Alcuni esempi di cosa intendiamo per violazione della riservatezza sono stati presentati nel paragrafo precedente, tuttavia, in letteratura esistono diversi approcci per formalizzare il concetto di violazione della riservatezza che qui presentiamo sinteticamente (Duncan e Lambert, 1986 e 1989).

Un approccio molto generale è basato sul principio introdotto in Dalenius (1977) principalmente per i dati tabellari: una violazione si verifica se il rilascio di una statistica S permette di conoscere un'informazione riservata più accuratamente di quanto non sia possibile fare senza conoscere S. Nel caso di rilascio di microdati l'approccio prende il nome di *inferential disclosure* e si enuncia: una violazione si verifica quando un utente può inferire nuove informazioni su un rispondente dai dati rilasciatigli, anche se nessun record nel file è associato a quel rispondente e se le informazioni inferite non sono esatte.

Un approccio anch'esso di tipo inferenziale è quello basato sul concetto di "violazione tramite un modello" (Palley e Simonoff, 1986), per cui si può acquisire un'informazione riservata stimando un valore riferito ad un individuo a partire da un modello statistico costruito sui dati rilasciati. La violazione viene, in questo caso, misurata tramite differenza fra il modello stimato sui dati originali e quello stimato sui dati rilasciati. Le difficoltà di questo approccio sono notevoli e ben descritte in Willenborg e de Waal (1996, 2001) che lo introducono come *predictive disclosure*. Supponiamo che $c$ sia una variabile riservata e denotiamo con $\mathbf{x}_i$ e $c_i$ rispettivamente i valori del vettore di variabili chiave e della variabile $c$ assunti dalla unità i-ma nel file. Assumiamo che un utente sia intenzionato a conoscere il valore $c_i$ avendo a disposizione in un proprio archivio $\mathbf{x}'_l = \mathbf{x}_i$ di un soggetto identificato. Sulla base dei dati rilasciati, l'utente è in grado di stimare, ad esempio, un modello:

$$c = f(\mathbf{x},\mathbf{b}) + \varepsilon \qquad\qquad \varepsilon \sim N(0,\sigma^2)$$

dove $b$ è un vettore di parametri, e, quindi, ottenere una predizione di $c_i$.

Se la politica di diffusione di un Istituto nazionale di statistica prevede il rilascio di un file di dati elementari di un'indagine, non è possibile escludere l'eventualità che vengano condotte analisi statistiche di questo tipo. D'altro canto la legislazione è interpretata normalmente nel senso di tutelare il diritto individuale del rispondente relativamente alle informazioni fornite, piuttosto che rispetto all'aumento generico di conoscenze che una terza persona può acquisire in base a un'analisi statistica. L'informazione così ottenuta può riguardare una sottopopolazione (di $U$) di individui rispondenti a certe caratteristiche a prescindere dal fatto che i singoli componenti abbiano partecipato alla rilevazione o meno. Rispetto a questi ultimi, comunque, non è stato sottoscritto il vincolo del rispetto del segreto statistico e, quindi, sarebbe ulteriormente complicato stabilire se si profili o meno una violazione della riservatezza.

Posto che si intenda limitare la possibilità di ottenere un'informazione riservata

mediante la stima di un modello predittivo, resta comunque da stabilire in che misura questa può considerarsi una violazione. L'incertezza dell'utente sul valore $c_i$ dipende da $\sigma^2$. Si può ipotizzare, di conseguenza, di misurare il rischio di violazione con un valore inversamente proporzionale a $\sigma^2$. Tuttavia, risulta problematico definire una soglia che non dipenda dalla variabile $c$ e che sia quindi valida per l'intero file. Ad esempio, se fosse possibile adattare un modello ai dati per stimare il livello di esportazioni di una certa impresa per area geografica, il livello di incertezza accettabile non potrebbe essere lo stesso di quello accettabile per la variabile dicotomica esporta/non esporta, per l'evidente divario di interesse che le due informazioni riservate potrebbero suscitare presso i competitori dell'impresa stessa. Inoltre, considerato il numero elevato di variabili che un file di dati elementari può contenere, anche l'ipotesi di definire una soglia di incertezza accettabile per ogni variabile in gran parte dei casi non è facilmente percorribile. Un altro modo di misurare l'eventuale rischio di violazione potrebbe essere, sempre secondo l'approccio predittivo, quello di ipotizzare che l'utente abbia *a priori* informazioni su $c$ sintetizzate con una distribuzione di probabilità. La valutazione del rischio, in tal caso, può essere misurata tramite la riduzione di incertezza conseguente al rilascio dei dati (Duncan e Lambert, 1986 e 1989). Ma in questo caso si pone un problema contraddittorio. Infatti, un utente con scarse conoscenze *a priori* capitalizzerebbe un maggior guadagno di informazioni e, quindi, una violazione maggiore, rispetto a un utente con una distribuzione *a priori* molto simile a quella predittiva stimata *a posteriori* sui dati.

In generale, comunque, l'obiezione principale a questo tipo di approccio sta nel fatto che, se l'utente non dispone dell'informazione che il soggetto di cui vuole conoscere informazioni riservate appartiene al file rilasciato allora la stima di una distribuzione predittiva non può essere considerata una violazione, poiché, come detto in precedenza, la stessa violazione si avrebbe per tutti i soggetti della popolazione con caratteristiche simili tra loro rispetto alle variabili chiave. D'altro canto, se all'utente è noto che il soggetto a cui è interessato appartiene al file, il discorso cambia perché le informazioni riservate possono essere associate a un individuo che le ha fornite con la promessa del rispetto della sua riservatezza. Questo ci porta a considerare un differente approccio per definire la violazione della riservatezza, basato sul rischio di identificazione del rispondente. Questo concetto risponde alla cosiddetta "violazione di identità", *identity disclosure* (Paass, 1988), consistente appunto nell'identificazione di un rispondente nel file rilasciato. In questo approccio la violazione da prevenire è associata al semplice riconoscimento di un individuo tra le unità statistiche del file.

Conseguente alla violazione d'identità è la violazione di attributo, *attribute disclosure* (Cox e Sande, 1979), dovuta all'acquisizione di informazioni riservate non disponibili altrimenti. In pratica, si verifica una violazione se è possibile associare correttamente il record a un rispondente identificato e se le informazioni contenute in un record del file rilasciato sono sufficientemente dettagliate ed accurate anche relativamente alle informazioni riservate. Questo è, in sostanza, l'approccio utilizzato generalmente dagli Istituti Nazionali di Statistica e va sotto il nome di *re-identification disclosure* (de Waal e Willenborg, 1996, Fienberg e Makov, 1998, Skinner e Holmes, 1998). Da notare che la violazione consiste nel permettere l'acquisizione di

informazioni riservate (o confidenziali) relative a un soggetto. Non è rilevante, al momento, che queste informazioni siano veritiere o un'approssimazione più o meno accurata della verità o di quanto è stato rilevato.

In base a questo approccio i record che intuitivamente presentano un più alto rischio di violazione sono quelli che rappresentano degli *outlier* rispetto alle variabili chiave in quanto riferiti a soggetti più facilmente identificabili. Si noti la differenza rispetto all'approccio predittivo in cui si può considerare un file dati privo di *outlier* più informativo per l'acquisizione di informazioni riservate nel senso che un "buon" modello che permetta di stimare informazioni confidenziali su un individuo "tipico" sarà più facile da ottenere se il file non presenta *outlier*.

Assumiamo che un ipotetico intruso (*intruder*) sia intenzionato ad ottenere informazioni riservate su una più unità statistiche (*target*) utilizzando i dati rilasciati. Si possono fare diverse ipotesi su quali siano le intenzioni di un intruso e su quali informazioni abbia a disposizione. Ad esempio può trattarsi di un'impresa che vuole ottenere informazioni relative a una sua concorrente oppure un malintenzionato che vuole solo dimostrare che il sistema consente violazioni. Diverse ipotesi possono farsi anche riguardo le informazioni *a priori* a disposizione dell'utente. A questo proposito assumiamo che riguardo alle unità statistiche l'utente abbia a disposizione un archivio nominativo in cui a ogni nome e cognome sono associate alcune variabili che almeno in parte coincidono con l'insieme delle variabili chiave.

Per archivio nominativo intendiamo un insieme di soggetti identificati, ad esempio:

| ID | Nome | Indirizzo | Provincia |
|---|---|---|---|
| 1 | Mario Rossi | Via Tirreno,1 | Pisa |
| 2 | Luigi Bianchi | Viale Ionio,2 | Roma |
| 3 | Gino Verdi | Viale Adriatico,34 | Roma |
| .. | …. | …. | ….. |

Per semplicità indichiamo l'identificativo di un soggetto unicamente con ID. I dati a disposizione dell'utente possono, allora, essere rappresentati con una matrice dati:

$$\mathbf{A}' = (\mathbf{ID}, \mathbf{X}')$$

dove l'insieme delle variabili in $\mathbf{X}'$ è un sottoinsieme delle variabili chiave in $\mathbf{X}$.

Diciamo allora che:

- un'unità statistica è identificabile se è associabile ad un identificativo, ossia a un soggetto identificato;
- definiamo re-identificazione la possibilità di stabilire una relazione biunivoca tra un record del file di dati elementari rilasciato, $\mathbf{A}$ e il suo identificativo nel file $\mathbf{A}'$;
- il contenuto della violazione è la conseguenza della re-identificazione cioè la conoscenza illecita di informazioni riservate;
- una re-identificazione è una relazione biunivoca stabilita tra un record di $\mathbf{A}$ e un record di $\mathbf{A}'$ basata esclusivamente sulle informazioni comuni ai due file.

Oltre alle informazioni contenute nell'archivio nominativo si considerano anche tutte le informazioni che in qualche modo possono aumentare il rischio di identificazione. Queste possono riguardare, ad esempio, la conoscenza del processo di produzione dei dati relativamente a disegno campionario, procedure di imputazione e correzione nonché le tecniche di protezione a tutela della riservatezza.

Quello appena descritto è lo schema generalmente utilizzato dagli Istituti nazionali di statistica per definire la violazione della riservatezza. Il processo di anonimizzazione di un file di dati elementari consiste allora nel rendere improbabile (al di sotto di una soglia ragionevole) la re-identificazione delle unità statistiche nel file rilasciato.

Definiamo, allora, il rischio di re-identificazione per un'unità statistica come la probabilità che il record $i$ nel file rilasciato appartenga al soggetto identificato ID date le informazioni a disposizione dell'intruso, ossia la matrice delle variabili chiave rilasciata $\mathbf{X}$, e i valori delle variabili chiave $\mathbf{x}'_{ID}$ presenti nell'archivio esterno assunti dal soggetto identificato ID:

$$\Pr\left[ID = ID(i) / \mathbf{X}, \mathbf{x}'_{ID}\right].$$

Nel valutare questa probabilità si può tenere conto di alcuni fattori che influenzano il rischio di re-identificazione come il fatto che le variabili chiave note all'utente non necessariamente sono classificate allo stesso modo delle corrispondenti nel file rilasciato, o che per entrambi gli archivi le variabili possono essere affette da errori, o ancora che l'archivio esterno a disposizione dell'utente non contiene necessariamente le stesse unità statistiche del file rilasciato: il target non è nel file rilasciato (si veda il Paragrafo 5.5 per dettagli).

Rispetto agli approcci descritti in precedenza, quello basato sul rischio di re-identificazione è più aderente al concetto di violazione della riservatezza stabilito dalla legge, che impone che i dati vengano rilasciati a condizione che non sia possibile trarne "riferimenti individuali". Rispetto all'approccio predittivo, è più facilmente interpretabile l'idea di un livello di rischio accettabile definito come probabilità. Ad esempio, "la probabilità che un rispondente venga re-identificato è pari a uno su diecimila" è immediatamente valutabile rispetto a "la variabilità implicita al modello predittivo stimato per la variabile $c$ è pari a 0.4". Inoltre, il livello così definito non è dipendente dalla variabile cui si riferisce e può essere unico per ogni $c$ in $\mathbf{C}$ e anche per ogni unità statistica $i$ nel file.

E' chiaro a questo punto che il processo di anonimizzazione messo in opera dagli Istituti nazionali di statistica consiste nell'applicare una serie di metodi statistici, ma anche tecnici e amministrativi, per contenere al di sotto di una soglia fissata il rischio di re-identificazione.

## 5.4 Tipi di rilascio per utente

Analizziamo in questo paragrafo e nel successivo alcuni fattori che influenzano il rischio di re-identificazione con particolare riferimento alle scelte operate dall'Istat. Ci riferiamo più precisamente al principio di "ragionevolezza" che solo recentemente è

stato introdotto anche sotto il profilo normativo (in Italia con la legge n.675/1996 e, a livello europeo, con il Regolamento Ce 322/1997) e della tipologia di utenti cui è destinato il file da rilasciare.

In teoria, infatti, ogni rilascio di un file di microdati comporta la possibilità che vi possano essere violazioni della riservatezza e l'unico modo per evitare ogni rischio è decidere di non rilasciare dati. L'atteggiamento che viene, invece, suggerito è quello di tenere in considerazione i mezzi che "ragionevolmente" un utente malintenzionato può impiegare per tentare l'identificazione di qualche unità statistica senza perdere di vista la dovuta prudenza. In tal senso non si deve dimenticare che il diritto individuale alla riservatezza deve prevalere nei confronti di quello collettivo all'informazione statistica, perciò non è escluso, e come spesso di fatto accade, si rinunci al rilascio dei dati se il rischio è troppo elevato.

Le risorse che possono essere impegnate per tentare la re-identificazione di un rispondente comportano un "costo" (disponibilità di archivi, tempo, disponibilità di hardware e software, professionalità, eccetera) che va misurato in funzione sia della possibilità di reperire le stesse informazioni presso altre fonti con un costo inferiore, sia in funzione della natura e dell'"età" dei dati. Tutto ciò influisce sulle motivazioni che possono spingere l'utente a tentare di violare la riservatezza dei dati. Un file relativo a dati generici sulla popolazione di dieci anni fa può ragionevolmente destare un minore interesse rispetto all'ultima rilevazione mensile sulle esportazioni da parte delle imprese.

Anche la procedura con cui si intende rilasciare il file può influire sul rischio di re-identificazione. Ad esempio, l'imposizione di un vincolo contrattuale è certo una garanzia che non può essere fatta valere se i dati vengono resi disponibili su internet.

Le misure amministrative dipendono dal tipo di utente cui vengono rilasciati i dati. Distinguiamo tre tipi di utenti:

1. utenti privilegiati (*special contractors*): sono ricercatori e analisti che hanno accesso ai dati per motivi istituzionali. In questo caso il fine di ricerca è comune con quello dell'Istituto che produce i dati e i soggetti ammessi alla comunicazione sono equiparati, sia dal punto di vista dell'accesso ai dati che da quello dei vincoli al rispetto delle norme sulla riservatezza, ai ricercatori dipendenti dell'Istituto stesso. L'accesso ai dati, perciò, non è limitato e nessuna misura di protezione o definizione di una soglia di rischio è necessaria.

2. ricercatori (*researchers*): sono utenti che richiedono di poter analizzare dati a livello individuale esclusivamente per fini di ricerca scientifica o statistica. Per questa categoria di utenti vengono predisposti, se possibile, dei file di dati elementari corrispondenti a un campione di rispondenti ai quali sono applicate delle misure di protezione statistica ai fini della tutela della riservatezza. Tali file prendono il nome di *collezioni campionarie di dati elementari* o *file standard* (*microdata file for research*, Mfr). Ai ricercatori che accedono a questa forma di comunicazione viene richiesta la sottoscrizione di un contratto che li impegna a non tentare la re-identificazione dei rispondenti e al rispetto delle norme sul segreto statistico. La soglia di rischio per questi file è relativamente "moderata" in considerazione delle finalità di ricerca e soprattutto per le garanzie derivanti

dalla firma di un accordo che ha valore contrattuale. L'Istat rilascia ogni anno *file standard* sulle principali indagini sulla popolazione (indagine sulle forze di lavoro, sui consumi delle famiglie, eccetera). Su richiesta possono essere elaborati *file standard* "*ad hoc*" secondo le stesse procedure ma con un rimborso spese più elevato data la specificità del prodotto.

3. pubblico (*general public*): in alcuni casi un Istituto nazionale di statistica può voler rendere disponibile un insieme di dati a livello individuale a chiunque ne faccia richiesta, ossia al pubblico. Solitamente si tratta di file dati esemplificativi che trovano collocazione su internet per essere acceduti liberamente o, più frequentemente, tramite una password che viene rilasciata previa identificazione dell'utente. In questi casi, considerata l'assenza di un accordo scritto e l'impossibilità di verificare l'uso che viene fatto dei dati, le misure di protezione prese in considerazione sono estremamente severe tanto da escludere con ogni probabilità che possano verificarsi re-identificazioni. Questo tipo di file viene denominato file "ad uso pubblico" (*public use file*, Puf) e presuppone che venga fissata una soglia di rischio accettabile estremamente bassa. L'Istat non ha prodotto esperienze significative per questo tipo di rilascio se non per dati molto dettagliati ma che rientrano a pieno titolo nell'ambito dei dati aggregati.

Da alcuni anni, ed esclusivamente per motivi di ricerca, la comunità scientifica ha la possibilità di accedere al Laboratorio per l'analisi dei dati elementari (Adele) presso la sede centrale dell'Istat. L'accesso al Laboratorio Adele non si configura come uno strumento per il rilascio di dati ma solo come una possibilità di condurre analisi statistiche. Le misure di protezione sono di tipo amministrativo e tecnico/pratico. Un controllo statistico per verificare la rispondenza alle norme sul segreto statistico viene condotto sui risultati prodotti. Il Laboratorio Adele viene descritto approfonditamente nel Paragrafo 9.3.

## 5.5 Fattori che influenzano il rischio di re-identificazione

In questo paragrafo vengono descritti alcuni fattori che tradizionalmente sono considerati elementi in grado di influenzare la percezione o il calcolo del rischio di identificazione da parte degli Istituti nazionali di statistica.

Il primo e principale fattore di rischio è la presenza e il numero dei *casi unici nella popolazione* rispetto alle variabili identificative indirette presenti nel file da rilasciare. In generale si definisce caso unico rispetto a determinate variabili un record che presenta una combinazione unica di modalità di tali variabili. L'importanza di tale fattore è piuttosto evidente se si considera che le unità che risultano essere casi unici rispetto alle variabili chiave sono più facilmente identificabili nella popolazione in quanto, appunto, univocamente individuati. Bisogna comunque fare una distinzione tra casi unici nella popolazione e casi unici campionari. In effetti nel file da rilasciare, soprattutto in ambito sociale, molti casi unici possono derivare da casi che nella popolazione (e nell'archivio esterno) sono doppi o tripli eccetera e per i quali è quindi difficile l'identificazione.

Per i dati riferiti alle imprese il numero dei casi unici riveste un aspetto particolare in quanto le variabili chiave sono spesso di tipo quantitativo e, conseguentemente, quasi ogni unità statistica rappresenta un caso unico. D'altro canto i dati riferiti alle imprese presentano delle difficoltà pratiche per la tutela della riservatezza statistica non solo sotto questo aspetto. Tratteremo queste peculiarità nel Capitolo 7.

In ambito sociale, invece, il numero dei casi unici può entrare direttamente nel calcolo del rischio di re-identificazione e in tal caso bisogna distinguere due situazioni: file derivanti da indagini censuarie; file derivanti da indagini campionarie. Nel primo caso il calcolo è fatto sulla base di dati disponibili mentre nel secondo caso occorre utilizzare delle procedure di stima che, sulla base dei dati campionari, permettano di fare inferenza sul numero dei casi unici nella popolazione. Nel Capitolo 6 verranno presentati i diversi modelli probabilistici per la stima di tale quantità presenti in letteratura.

Il secondo fattore che influenza il rischio di violazione è il *tasso di campionamento* della collezione campionaria. Per capire l'influenza di tale fattore sul rischio basta considerare il legame che intercorre tra tasso di campionamento e casi unici nella popolazione. Infatti più è piccolo tale fattore, minore sarà la numerosità di casi unici della popolazione presenti nel campione e viceversa.

Ricordiamo che nel caso in cui l'indagine di riferimento è un'indagine censuaria, la collezione campionaria sarà costituita da un campione estratto dall'indagine stessa, mentre nel caso in cui si considera un indagine campionaria la collezione campionaria potrà essere lo stesso campione o un sottocampione da esso estratto.

Il terzo fattore riguarda invece la *quantità* e la *concentrazione delle informazioni esterne*.

Logicamente tanto più concentrata e consistente è l'informazione in possesso dell'intruso tanto maggiore sarà il suo effetto sul rischio di identificazione.

Il quarto fattore riguarda invece la probabilità che le variabili identificative indirette, presenti nell'archivio dei dati a disposizione dell'intruso, siano presentati con la *medesima codifica* di quelle presenti nel file dei dati rilasciati. In effetti ci potrebbero essere delle differenze dovute all'utilizzazione di definizioni o classificazioni differenti, ad un intervallo temporale tra la costruzione dell'archivio e quella di rilascio del file o più semplicemente alla presenza di errori di registrazione in uno o su entrambi i registri.

L'ultimo fattore che consideriamo è la *propensione dell'utente ad effettuare un tentativo di identificazione*. Tale fattore viene generalmente considerato molto contenuto considerato che abitualmente i file di dati elementari vengono rilasciati esclusivamente per motivi di ricerca e dietro la sottoscrizione di un accordo-contratto che vincola l'utente a non tentare di violare la riservatezza dei dati. Anche in questo caso la differenza tra l'ambito sociale e quello economico è percettibile in quanto è ragionevole pensare che ci possa essere un maggiore interesse a recuperare informazioni riservate su un'impresa (ad esempio da parte di una impresa concorrente) rispetto alle informazioni rilevate sulle persone fisiche.

# Capitolo 6. Rischio di violazione di dati individuali in ambito sociale(*)

## 6.1 Introduzione

L'ambito delle indagini sociali, quelle cioè dove le unità statistiche sono le persone fisiche, è quello tradizionalmente più interessato dalle problematiche per il rilascio di file di dati elementari da parte degli Istituti nazionali di statistica. In particolare, l'Istat rilascia ormai da diversi anni i cosiddetti *file standard* sulle principali indagini campionarie sulle famiglie (Forze di lavoro, Consumi delle famiglie, eccetera) mentre a tutt'oggi non si registrano esperienze analoghe per le imprese. E' evidente, infatti, che la materia trova una ragionevole applicazione considerata la natura dei dati sociali mentre è ostacolata decisamente per quanto riguarda i dati economici.

Anche la ricerca metodologica e le diverse definizioni di "violazione della riservatezza" nascono inizialmente per i dati in ambito sociale proprio per la maggiore applicabilità dei metodi a questo tipo di dati (Biggeri e Zannella, 1991, Coccia, 1993). L'approccio, comunque, si è evoluto nel tempo e da un approccio globale basato sulla misura di protezione generale per un file di dati, si è passati ad un approccio individuale che permette, in maniera più analitica, la misurazione del rischio di identificazione per ogni singolo record nel file.

Nei successivi paragrafi verranno analizzati i due diversi approcci nella definizione di tale funzione.

Nel Paragrafo 6.2 la funzione che misura il rischio di re-identificazione è definita come una misura relativa all'intero file dei dati da rilasciare, parleremo quindi di funzioni del *rischio globale*, mentre nel Paragrafo 6.3 viene definita una misura del rischio a livello di singolo record; si parlerà così di funzioni del *rischio individuale*.

## 6.2 Funzioni di Rischio globale

In questo paragrafo si considera la funzione del rischio di violazione come una misura del rischio associato all'intero file dei dati. In quest'ottica un ufficio di statistica che è interessato al rilascio di dati individuali dovrà stabilire una soglia di massimo livello di rischio che si è disposti ad accettare e ogni qualvolta un file presenti un rischio superiore a tale soglia, questo verrà modificato secondo alcune tecniche di protezione. L'applicazione di tali tecniche, dette anche di "contenimento del rischio", comporta una diminuzione dell'informazione contenuta nel file e di conseguenza una riduzione del rischio di violazione ad esso associato. Questi metodi sono discussi nel Paragrafo 8.3 che è dedicato a una trattazione unitaria delle tecniche di protezione per i dati elementari (si veda in particolare i Paragrafi 8.3.1 e 8.3.2).

Tenendo conto di quanto detto nel Paragrafo 5.5 risulta che il rischio di identificazione globale è funzione di diversi fattori ed è quindi esprimibile come:

(*) Capitolo redatto da Alessandra Capobianchi

$$R_g = f(f_u, f_c, f_a, f_I, f_T)\,.$$

dove con $f_u$ indichiamo la frequenza dei casi unici nel campione, con $f_c$ il tasso di campionamento della collezione campionaria, $f_a$ la frequenza relativa delle unità presenti nell'archivio esterno, $f_I$ la probabilità che le variabili siano codificate identicamente nei due file ed infine con $f_T$ la propensione dell'utente all'identificazione.

Per poter ottenere una effettiva stima del valore del rischio globale associato ad un determinato file bisognerà specificare la forma della funzione *f* e fissare o stimare i valori dei fattori di cui è funzione il rischio stesso.

Nel Paragrafo 6.2.1 si analizza la stima del numero di casi unici nella popolazione. Nel Paragrafo 6.2.2 è presentata una breve rassegna dei modelli di rischio.

### 6.2.1 Stima del numero dei casi unici

Il fattore che maggiormente influenza il rischio di violazione è sicuramente il numero dei casi unici nella popolazione. Come precedentemente notato tale valore è direttamente osservabile nel caso in cui il file dei microdati da rilasciare proviene da un'indagine censuaria, mentre nel caso di un'indagine campionaria tale valore dovrà essere stimato.

A tale scopo sono stati proposti diversi modelli. Il più noto tra questi è quello proposto da Bethlehem *et al*. nel 1990 detto anche *modello Poisson-Gamma*.

Consideriamo la tabella di contingenza associata all'insieme delle variabili identificative indirette; (ad esempio se nel file sono presenti sesso, età, stato civile e regione di residenza allora la tabella associata sarà sesso× età× stato civile× regione di residenza) ed indichiamo con:

*N*= numero delle unità della popolazione;

*K*= numero di combinazioni possibili di modalità di variabili identificative indirette ovvero il numero delle celle della tabella di contingenza associata;

$F_i$= numero di unità della popolazione che presentano la *i*-ma combinazione di modalità di variabili identificative indirette (*i*=1,…,*K*) ovvero frequenza della cella *i*-ma;

$U_p$= numero dei casi unici nella popolazione ovvero il numero delle celle che presentano frequenza unitaria ($F_i$=1).

Nel modello teorico di Bethlehem si considera la tabella di contingenza associata alla popolazione di *N* unità come la realizzazione di *K* variabili aleatorie $Y_i$ (*i*=1,…,*K*) con distribuzione di Poisson di media $\mu_i = N\pi_i$, dove $\pi_i$ rappresenta la probabilità di una singola unità della popolazione di appartenere alla cella i-ma. La variabile $Y_i$ descrive quindi il numero delle unità della popolazione aventi combinazione *i*-ma e le sue realizzazioni *y* (*y*=0,1,…,*N*) corrispondono alle frequenze $F_i$ ; in particolare abbiamo:

$$\Pr(Y_i = y) = \frac{\mu_i^y \exp(-\mu_i)}{y!}\,.$$

Per poter ottenere la stima del numero dei casi unici occorre stimare i valori attesi $\mu_i$ per $i$=1,…,$K$. Essendo $K$ generalmente piuttosto elevato, per risolvere il problema di stima si assume che le $\pi_i$ siano realizzazioni di $K$ variabili aleatorie indipendenti di tipo Gamma con parametri $\alpha$ e $\beta$ (con $\alpha\beta$=1/$K$) ed inoltre si assume che tali parametri siano indipendenti dalle caratteristiche delle celle e quindi uguali per tutte. Sotto queste ipotesi (modello Poisson-Gamma) otteniamo che la distribuzione marginale di ogni $Y$ è una binomiale negativa ovvero:

$$\Pr(Y_i = y) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\Gamma(y+1)} \frac{(N\beta)^y}{(1+N\beta)^{\alpha+y}} \tag{6.1}$$

dove $\Gamma(\cdot)$ è la funzione gamma.

In base alla (6.1) la probabilità di avere un caso unico nella cella i-ma è dato da:

$$\Pr(Y_i = 1) = \frac{N}{K}(1+N\beta)^{-(\alpha+1)}$$

e il numero atteso dei casi unici nella popolazione sarà dato da:

$$\mathrm{E}(U_p) = K \Pr(Y_i = 1) = N(1+N\beta)^{-(\alpha+1)}$$

e dipende dai parametri incogniti $\alpha$ e $\beta$. Per ottenere una stima di tale valore bisogna stimare i parametri $\alpha$ e $\beta$ i quali, a partire dai dati campionari disponibili, possono essere stimati con diversi metodi, ad esempio come il metodo dei momenti o quello della massima verosimiglianza. Nell'ottica di questo approccio otteniamo: $\hat{U}_p = \hat{E}(U_p)$, ossia una stima del numero atteso di casi unici nella popolazione.

Diverse applicazioni a dati reali o simulati hanno dimostrato che l'uso del modello Poisson-Gamma per la stima del numero dei casi unici porta a risultati instabili. In particolare ricordiamo lo studio condotto da Biggeri e Zannella (1991) i quali hanno utilizzato due diversi insiemi di microdati relativi ad indagini condotte in Italia e hanno stimato il numero di casi unici fissando diversi insiemi di variabili identificative. La stima è risultata accettabile solo nel caso di un numero ridotto di variabili identificative, con una distorsione crescente al crescere del numero delle variabili stesse; inoltre l'applicazione di tale metodo ha messo in evidenza una crescente sottostima dei casi unici al diminuire della numerosità campionaria e ad una instabilità nel caso in cui il numero dei casi unici da stimare è relativo a delle sottopopolazioni.

A partire dalle considerazioni appena fatte Skinner e Holmes (1993) propongono un diverso modello nel quale, come nel modello proposto da Bethlehem, si assumono le $Y_i$ distribuite come una Poisson con parametri $\mu_i$ ma, a differenza del modello precedentemente analizzato, tali parametri vengono considerati come realizzazioni di variabili di tipo Lognormale.

Un netto miglioramento è stato ottenuto da Crescenzi (1993) considerando un modello che si basa sulla combinazione di una Binomiale Negativa con una Gamma. Consideriamo $\alpha$ e $\gamma$ i parametri di una tale distribuzione e due insiemi di dati di dimensione $N_1$ ed $N_2$. Conoscendo il numero dei casi unici contenuti nel primo insieme dei dati è possibile prevedere il numero dei casi unici contenuti nel secondo insieme

mediante una funzione, del tipo:

$$U_2 = U_1\left(1+\gamma\log\left(\frac{N_2}{N_1}\right)\right)^{-(\alpha+1)} = U_1\ PF(\alpha,\gamma)\,.$$

La funzione $PF(\cdot)$ è detta funzione di previsione e stabilisce una relazione fra la dimensione del file di dati ed il numero dei casi unici in esso presenti. Se con $U_2$ indichiamo $U_p$, il numero dei casi unici nella popolazione, e con $U_1$ il numero dei casi unici campionari, la funzione di previsione indica la relazione intercorrente tra i casi unici campionari e quelli nella popolazione. La funzione di previsione dipende dai parametri incogniti $\alpha$ e $\gamma$ quindi per ottenere la stima $\hat{U}_p$ bisogna stimare detti parametri. A tale scopo Crescenzi (1993) propone un metodo iterativo che fornisce una soluzione numerica per la stima di $\alpha$ e $\gamma$.

### 6.2.2 Modelli di rischio

Abbiamo precedentemente sottolineato come per ottenere un'effettiva stima del rischio di un file di microdati sia necessario, oltre alla determinazione o stima dei diversi fattori che lo influenzano, specificare la forma della funzione *f*.

In letteratura è possibile trovare diverse proposte. Il primo tentativo di quantificazione del rischio di violazione associato ad un file di microdati lo troviamo in Bethlehem *et al*. (1990).

Il modello proposto si basa esclusivamente sulla stima del numero atteso di casi che sono unici nella popolazione e che sono stati inclusi nel campione, $U_{pc}$. Secondo questo criterio, detto *criterio assoluto,* il rischio globale $R_g$ associato al file dei microdati viene posto uguale a tale quantità, la cui stima sarà data da:

$$\hat{R}_g = \hat{U}_{pc} = \hat{U}_p\frac{n}{N}$$

dove $\hat{U}_p$ è la stima del numero atteso dei casi unici nella popolazione ottenuta con il modello Poisson-Gamma proposto dallo stesso Bethlehem e precedentemente descritto.

Nella letteratura internazionale vengono prese in considerazione due particolari funzioni del rischio (*RV*1 e *RV*2) le quali, a differenza della funzione proposta da Bethlehem, non dipendono solo dal numero dei casi unici nella popolazione ma sono funzioni anche di altri fattori (descritti nel Paragrafo 5.5), che influenzano il rischio stesso come: il tasso di campionamento della collezione campionaria, la frequenza relativa delle unità presenti nell'archivio esterno, la probabilità che le variabili siano codificate identicamente nei due file e la propensione dell'utente all'identificazione.

Indicando con $X$ la variabile casuale che rappresenta il numero di identificazioni nel file da rilasciare, il numero medio atteso può essere scritto come:

$$\mu = E(X) = N\ f_u f_c f_a f_I f_T\ .$$

Assumendo inoltre per la variabile casuale $X$ una distribuzione di tipo Poisson la

probabilità che si abbia almeno una identificazione è data da:

$$\Pr(X > 0) = 1 - \Pr(X = 0) = 1 - \exp\{-\mu\} .$$

Una prima misura del rischio di violazione può essere data dal numero atteso di unità del file che sono identificabili rapportato alla numerosità del campione, ottenendo così:

$$RV1 = \frac{\mu}{n} .$$

In generale il valore di $RV1$ è minore di 1. Per tale motivo si usa spesso esprimere il valore di $RV1$ in termini di numero di unità campione necessarie per avere un'identificazione.

Alternativamente la probabilità che, estratto un campione, esso contenga almeno un caso identificabile, può essere considerata come misura del rischio di violazione ovvero:

$$RV2 = 1 - \exp(-\mu) = 1 - \exp(-N\, f_u f_c f_a f_1 f_T) .$$

La funzione del rischio $RV2$ è un'estensione della funzione proposta da Mokken *et al*. (1992) in cui, come fattori che influenzano il rischio, vengono considerati esclusivamente il tasso di campionamento, la frequenza nel campione di casi unici della popolazione e la proporzione di unità presenti anche nell'archivio esterno.

In particolare ricordiamo che la procedura di controllo di violazione della riservatezza di file di microdati che veniva applicata in precedenza in Istat prevedeva l'utilizzo della funzione $RV1$ come funzione del rischio, mentre come procedura per la stima del numero dei casi unici della popolazione veniva applicata la procedura proposta da Crescenzi (1993) descritta nel paragrafo precedente.

## 6.3 Rischio Individuale

I modelli di rischio presentati nel precedente capitolo definiscono il rischio di violazione come una misura globale associata all'intero file dei dati da rilasciare. Tale misura, pur consentendo di valutare se il file, così com'è, può essere rilasciato, non fornisce nessuna informazione su come ogni singola unità contribuisca alla determinazione del rischio di violazione e quindi non permette un suo utilizzo come strumento per l'individuazione di unità a "rischio" su cui intervenire prima del rilascio del file dei microdati (Duncan e Lambert 1989).

L'esigenza di passare da un rischio globale per l'intero file ad un rischio individuale per ciascuna unità del file è ampiamente riconosciuta a livello internazionale.

In effetti l'utilizzo di una funzione del rischio definita a livello individuale comporta un duplice vantaggio in quanto l'intervento sulle sole unità che risultano maggiormente individuabili permette da una parte di ottenere un file maggiormente protetto e dall'altra di aumentare il contenuto informativo del file stesso, in quanto non verrebbero alterate le informazioni relative a record che non necessitano di protezione.

Proprio per questi motivi, negli ultimi anni, si sono sviluppate nuove metodologie volte alla determinazione di un rischio individuale. Nel Paragrafo 6.3.1 si analizzano le proposte di Skinner e Holmes (1998) e quella di Benedetti *et al*. (2003); in entrambi i casi si definisce una misura del rischio di violazione individuale. Nell'approccio descritto da Skinner e Holmes il rischio, seppur pensato a livello individuale, viene definito e stimato per i soli casi unici presenti nel campione tramite un modello log-lineare che dipende dalle variabili chiave. L'approccio proposto da Benedetti e Franconi si basa invece sull'utilizzo dei pesi campionari per la stima del rischio individuale che viene definito per tutti gli individui contenuti nel file da rilasciare. Nel Paragrafo 6.3.2 viene descritta la metodologia sviluppata all'interno dell'Istituto nazionale di statistica olandese (Cbs) (si veda Willenborg e de Waal, 1996, e 2001). Tale metodologia si basa sulla definizione del concetto di rarità e, a differenza delle due precedenti metodologie, il rischio di violazione non è definito in termini probabilistici.

### 6.3.1 Rischio Individuale probabilistico

Indichiamo con $r$ il record contenuto nel file da rilasciare per il quale si vuole definire una misura del rischio di violazione della riservatezza indichiamo con $r^*$ l'unità della popolazione da cui proviene tale record.

Come misura base del rischio di violazione del record $r$ si può considerare la quantità:

$e_r$= *evidenza* disponibile ad un intruso in supporto ad un legame tra $r$ ed $r^*$

dove con il termine *evidenza* si indica un insieme di informazioni disponibili; per poter esplicitare una funzione del rischio tale concetto deve essere ulteriormente specificato. Uno dei possibili approcci è quello di misurare l'*evidenza* in termini di probabilità. Indicando con:

$$p_{rs^*} = \Pr(\textit{record } r \textit{ corrisponda all' unità } s^*)$$

una misura dell'evidenza si ottiene ponendo $e_r = p_{rr^*}$ dove $r^*$ è l'unità della popolazione da cui proviene il record $r$.

Nell'approccio proposto da Skinner e Holmes (1998) la misura di probabilità sopra considerata è definita rispetto ad un modello per le variabili identificative indirette relative sia al file dei microdati che ai dati esterni e rispetto allo schema di campionamento.

Prima di procedere introduciamo alcune notazioni. Supponiamo che l'intruso abbia a disposizione $m$ variabili identificative indirette discrete, $X_1, X_2, \ldots, X_m$ ed indichiamo con:

$x=(x_1,\ldots,x_K)$ la generica combinazione di valori di variabili identificative indirette;

$X$ la variabile aleatoria che assume valori $x=(x_1,\ldots,x_K)$;

$F_x$ il numero delle unità della popolazione che presentano la combinazione $x$;

$f_x$ il numero delle unità del file dei microdati che presentano la combinazione $x$;

$x(r)$ la combinazione di modalità associata al record $r$.

Ipotizziamo che siano noti i valori di $F_x$, che la variabile $X$ non presenti errori di rilevazione e che le unità presenti nel file dei microdati siano state selezionate con uguale probabilità dalla popolazione di riferimento. In questo caso otteniamo:

$$\Pr(\textit{record } r \textit{ provenga da unità } s^* / F_{x(r)}) = 1/F_{x(r)}$$

in quanto per l'intruso il generico record $r^*$ può provenire da una qualsiasi delle $F_{x(r)}$ unità della popolazione che presentano la stessa combinazione di variabili identificative indirette (ovvero lo stesso valore della variabile aleatoria $X$).

La probabilità non condizionata sarà data da:

$$\Pr(\textit{record } r \textit{ provenga da unità } s^*) = \Pr(F_{x(r)} = 1) + \Pr(F_{x(r)} = 2)/2 + \Pr(F_{x(r)} = 3)/3 + \cdots$$

In realtà i valori delle $F_x$ sono incogniti. Se l'intruso è in grado di associare una distribuzione di probabilità $\Pr(F_x)$ ad $F_x$ sarà possibile valutare la precedente espressione.

In ogni caso le probabilità sopra considerate devono essere sempre valutate in base alle informazioni disponibili all'intruso e in particolare in relazione ai valori di $f_x$. Skinner e Holmes (1998) ipotizzano che per i record che non sono unici nel campione ovvero record per cui $f_{x(r)} \geq 2$, il rischio di violazione sia sufficientemente basso. Sulla base di queste considerazioni gli autori focalizzano la loro attenzione sulla quantità $\Pr(F_{x(r)} = 1 | f_{x(r)} = 1)$ come misura di evidenza, ovvero:

$$e_r = \Pr(F_{x(r)} = 1 | f_{x(r)} = 1).$$

Per calcolare questa quantità gli autori suppongono che le $F_x$ siano indipendenti e generate da una distribuzione di Poisson di media $\lambda_x$ e che le $\lambda_x$ siano, a loro volta, generate da una mistura tra una distribuzione comune $g(\lambda_x)$ e una massa di probabilità in 0 per valori non possibili della variabile $X$. Per quanto riguarda la distribuzione $g(\lambda_x)$ gli autori prendono in considerazione la seguente generalizzazione di un modello lognormale:

$$\begin{aligned} &\log \lambda_x = \eta_x + \varepsilon_x, \quad \varepsilon_x \approx N(0, \sigma^2) \\ &\eta_x = \mu + u_{x_1}^{X_1} + u_{x_2}^{X_2} + \ldots + u_{x_k}^{X_k} \end{aligned} \tag{6.2}$$

dove gli $u_{x_i}^{X_{1i}}$ rappresentano gli effetti principali delle modalità delle variabili identificative indirette.

Il modello (6.2) potrebbe essere esteso includendo termini che riflettono le interazioni tra le variabili chiave. Se venissero incluse le interazioni fino ad un ordine sufficientemente elevato il termine $\varepsilon_x$ potrebbe essere eliminato. D'altra parte però, l'inclusione di un numero alto di interazioni potrebbe comportare: una instabilità nella stima di $\Pr(F_{x(r)} = 1 | f_{x(r)} = 1)$, maggiori complicazioni computazionali ed inoltre potrebbero sorgere problemi relativi alla selezione del modello stesso.

Sotto le ipotesi del modello (6.2) la distribuzione $g(\lambda_x)$ è data da:

$$g(\lambda_x) = (2\pi\sigma^2)^{-\frac{1}{2}} \lambda_x^{-1} \exp[-(\log \lambda_x - \eta_x)^2 / 2\sigma^2].$$

Trattando il campione dei microdati come un campione ottenuto da un'estrazione di tipo bernoulliano con parametro $p = n/N$, dove con *n* e *N* indichiamo rispettivamente la numerosità del campione e della popolazione, si può assumere che la variabile aleatoria $f_x \mid \lambda_x$ sia distribuita come una Poisson di media $p\lambda$ mentre $F_x - f_x \mid \lambda_x$ abbia una distribuzione di Poisson con media $(1-p)\lambda_x$ per $x$=1,2,…,$K$ e che tali variabili siano tra loro indipendenti dato $\lambda_x$. Otteniamo così:

$$e_r = P(F_x = 1 \mid f_x = 1) = \int \exp[-(1-p)\lambda]\ g(\lambda \mid f = 1) d\lambda$$

dove

$$g(\lambda_x \mid f_x = 1) = \lambda_x e^{-p\lambda_x} g(\lambda_x) / \int \lambda e^{-p\lambda} g(\lambda) d\lambda .$$

Per ottenere una misura che possa essere calcolata a partire dai dati campionari è necessario stimare i parametri incogniti $\eta_x$ e $\sigma^2$. A questo scopo gli autori considerano il modello lognormale precedentemente introdotto come un modello sovradisperso e la stima della media $\mu_x = pE(\lambda_x) = p\exp(\eta_x + \sigma^2/2)$ viene ottenuta utilizzando i classici stimatori dei parametri di un modello loglineare. Un problema si incontra nella stima di $\sigma^2$ in quanto non è nota la proporzione delle celle per le quali $\lambda_x = 0$. Per ovviare a tale problema, gli autori, per la stima di tale parametro, considerano esclusivamente i dati relativi a celle per le quali $f_x \geq 1$ e considerando che:

$$\frac{E(f_x^2 - f_x \mid f_x \geq 1)/\mu_x^2}{E(f_x \mid f_x \geq 1)/\mu_x} = \exp(\sigma^2)$$

si ottiene la stima:

$$\hat{\sigma}^2 = \log\left\{\left[\sum_x (f_x^2 - f_x)/\hat{\mu}_x^2\right] \Big/ \left[\sum_x f_x/\hat{\mu}_x\right]\right\},$$

mentre la stima di $\eta_x$ è data da:

$$\hat{\eta}_x = \log[\hat{\mu}_x / p(\exp(\hat{\sigma}^2/2))] .$$

Per quanto sopra otteniamo la seguente stima del rischio individuale:

$$\hat{e}_r = \hat{P}(F_x = 1 \mid f_x = 1) = \frac{\int \exp[-\lambda - (\log\lambda - \hat{\eta}_x)^2/2\hat{\sigma}^2] d\lambda}{\int \exp[-p\lambda - (\log\lambda - \hat{\eta}_x)^2/2\hat{\sigma}^2] d\lambda} .$$

L'approccio fin qui seguito può portare ad ottenere delle stime di segno negativo per il parametro $\sigma^2$. In questo caso gli autori suggeriscono di porre uguale a zero il valore di $\sigma^2$, e a partire dalla 6.1 ottengono la seguente versione semplificata della funzione del rischio:

$$\Pr(F_{x(r)} = 1 \mid f_{x(r)} = 1) = \Pr(F_x - f_x = 0) = \exp[1 - (1-p)\lambda_x] .$$

Stimando $\lambda_x = \exp(\eta_x)$ con $\hat{\mu}_x / p$ si ottiene:

$$\hat{e}_r = \hat{P}(F_x = 1 \mid f_x = 1) = \exp[-(1-p)\hat{\mu}_x / p].$$

Questa formula del rischio può essere usata anche per valori positivi di $\sigma^2$ in quanto risulta notevolmente più semplice da calcolare rispetto alla precedente.

Un approccio in parte simile al precedente è quello dovuto a Benedetti e Franconi (1998) e Benedetti *et al*., 2003. Come prima indichiamo con $F_x$ ed $f_x$ rispettivamente il numero delle unità della popolazione e del file dei microdati che presentano la combinazione $x$ della variabile aleatoria $X$. Si può notare che la combinazione delle modalità delle $m$ variabili identificative comporta la determinazione di $K$ domini sia nella popolazione di riferimento che nel file dei microdati, tali domini saranno quindi identificati dalla variabile aleatoria $X$. Nel file dei microdati da rilasciare solo un sottoinsieme di tali domini presenterà una dimensione non nulla e solo questi saranno di interesse da un punto di vista della riservatezza.

Prima di sviluppare un modello quantitativo per la definizione di un rischio di violazione individuale bisogna fare delle assunzioni sulla natura delle possibili strategie che l'intruso può utilizzare per violare la riservatezza.

Per esaminare la condizione più sfavorevole possibile si assume che:

- l'intruso abbia a disposizione un archivio o un pubblico registro contenente tutte le variabili identificative dirette e indirette relative all'intera popolazione di riferimento;
- sia il file dei microdati che il registro a disposizione dell'intruso non presentino errori di misurazione nella rilevazione delle variabili identificative indirette;
- l'intruso cerchi di identificare il record $r$ confrontando le combinazioni di modalità delle variabili identificative indirette osservate nel file dei microdati con quelle relative alle unità presenti nei registri in suo possesso.

L'identificazione del record $r$ (definita come l'individuazione corretta delle variabili identificative dirette associate a tale record; evento indicato con $I_r$) avviene quando l'intruso effettua un corretto collegamento tra il record r e l'individuo $r^*$, contenuto nel registro esterno della popolazione, da cui proviene.

Se nel registro esterno esiste una sola unità che presenta la stessa combinazione del record $r$ allora il collegamento sarà univocamente determinato. Per i record per cui sono presenti più unità che presentano la medesima combinazione, il collegamento non sarà deterministico ma la scelta sarà basata su un meccanismo di tipo probabilistico. In questo approccio il rischio di violazione associato al record $r$ viene definito come la probabilità di effettuare un'identificazione dato il campione osservato $s$, ovvero:

$$e_r = P(I_r \mid s).$$

Da notare che per l'intruso i record contenuti nel file dei microdati che presentano la stessa combinazione di modalità di variabili identificative indirette sono identiche in termini di incertezza nell'effettuare una identificazione. Quindi per ogni record $r$' contenuto nel dominio $x(r)$ abbiamo:

$$e_{r'} = e_r \quad \forall \; r' \in x(r).$$

Nel seguito indicheremo con il termine $e_{x(r)}$ il rischio associato a tutti i record contenuti nel dominio $x(r)$. L'informazione contenuta nel campione in termini di identificazione $r$ consiste nella frequenza osservata $f_{x(r)}$ quindi otteniamo:

$$e_r = e_{x(r)} = P(I_r \mid s) = P(I_r \mid f_{x(r)}) .$$

Tale funzione può essere espressa come:

$$\begin{aligned} e_{x(r)} = P(I_r \mid f_{x(r)}) = P(I_r \mid F_{x(r)} = 1, f_{x(r)}) P(F_{x(r)} = 1 \mid f_{x(r)}) + \\ + P(I_r \mid F_{x(r)} = 2, f_{x(r)}) P(F_{x(r)} = 2 \mid f_{x(r)}) + ... \end{aligned}$$

Ricordiamo che una delle ipotesi fatte relativamente allo scenario dell'intruso è che il registro a sua disposizione contenga le informazioni relative a tutta la popolazione. Si ha quindi:

$$P(I_r \mid F_{x(r)} = h, f_{x(r)}) = \frac{1}{h} \qquad h = 1,\ 2,...,\ K ,$$

per cui:

$$e_{x(r)} = \sum_{h \geq f_{x(r)}} \frac{1}{h} P(F_{x(r)} = h \mid f_{x(r)}) . \tag{6.3}$$

Per poter valutare la precedente espressione è necessario fare delle ipotesi distribuzionali sui parametri incogniti $F_{x(r)}$ dati i valori osservati $f_{x(r)}$. Per semplificare la notazione nel prosieguo ometteremo l'indicazione dell'unità $r$ di riferimento ovvero poniamo $x(r)=x$.

L'idea di base è quella di considerare la variabile casuale $F_x \mid f_x$ distribuita come una binomiale negativa con $f_x$ successi e probabilità di successo pari a $p_x$ (Skinner *et al*., 1994; Bethlehem *et al*., 1990):

$$P(F_x = h \mid f_x) = \binom{h-1}{f_x - 1} p_x^{f_x} (1 - p_x)^{h - f_x} \qquad h = f_x, f_{x+1}, ...;\ f_x > 0 . \tag{6.4}$$

Questo approccio si basa sull'assunzione che, avendo osservato $f_x$ successi, dove un successo viene considerato come un'estrazione con probabilità $p_x$ dal $x$-mo dominio della popolazione di dimensione $F_x$, lo schema di estrazione può essere visto come un campionamento di tipo binomiale inverso. Sulla base di queste considerazioni la funzione del rischio di violazione (6.3) non è altro che l'espressione del momento di ordine –1 di una distribuzione binomiale negativa ovvero $e_x = E(F_x^{-1} / f_x)$. Utilizzando la funzione generatrice dei momenti di una distribuzione binomiale negativa nell'espressione $e_x = E(F_x^{-1} / f_x)$ e dopo alcune sostituzioni si ottiene:

$$e_x = \left( \frac{p_x}{q_x} \right)^{f_x} \int_1^{\frac{1}{f_x}} \frac{(y-1)^{f_x - 1}}{y} dy \qquad \text{dove} \quad q_x = 1 - p_x . \tag{6.5}$$

Per poter valutare l'espressione precedente bisogna stimare il parametro $p_x$ per $x$=1,…,$K$. Lo stimatore di massima verosimiglianza di $p_x$ è dato da $\hat{p}_x^{\text{MLE}} = f_x / F_x$. Per poter valutare tale stimatore viene usata l'informazione contenuta nel campione ottenendo:

$$\hat{p}_x = \frac{f_x}{\sum w_r} \qquad \text{dove} \qquad \hat{F}_x = \sum_{r:x(r)=x} w_r$$

dove con $w_r$ indichiamo i pesi di riporto all'universo e quindi $\hat{F}_x$ è il classico stimatore di Horvitz-Thompson del parametro $F_x$.

Nella procedura fin qui descritta i record vengono considerati indipendenti e nessuna informazione relativa all'appartenenza degli stessi a dei gruppi viene utilizzata nel calcolo del rischio di violazione. In realtà in molti file di microdati i record rilasciati non possono essere considerati come indipendenti in quanto è presente l'informazione sulla loro appartenenza a gruppi, informazione che può essere sfruttata dall'intruso per violare la riservatezza. Ovviamente affinché tale informazione sia utilizzabile dall'intruso esso deve disporre di registri o data-base che contengano lo stesso tipo di informazione contenuta nel file dei dati rilasciati. Una tale situazione è molto comune per diverse collezioni di dati raccolte dagli Istituti nazionali di statistica, in particolare ricordiamo che collezioni campionarie relative alla popolazione hanno come unità primarie le famiglie. La famiglia è un tipico esempio di variabile di "gruppo" che rende gerarchici i record di molti file di microdati.

Il problema è quindi quello di definire una funzione del rischio di violazione individuale che tenga conto dell'incremento del rischio stesso causato dai legami intercorrenti tra gli individui.

Prima di definire una funzione del rischio individuale gerarchico bisogna fare delle ipotesi sul come l'intruso può sfruttare l'informazione di appartenenza ad un gruppo per effettuare un'identificazione. Indicheremo con $\delta(r)$ il gruppo di appartenenza del record $r$. Una possibile strategia che l'intruso può utilizzare per violare la riservatezza è la seguente:

**passo 1** l'intruso cerca di collegare il singolo record $r$ all'unità $r^*$ che presenta la stessa combinazione di modalità di variabili identificative indirette;

**passo 2** l'intruso cerca di sfruttare l'informazione relativa al gruppo ovvero cerca di individuare un legame tra ogni singolo record $j$ appartenente al gruppo $\delta(r)$ e le unità $j^*$ della popolazione (*interazioni del primo ordine)*;

**passo 3** cercando di sfruttare ulteriormente le informazioni relative al gruppo $\delta(r)$, ovvero l'informazione relativa alle *interazioni del secondo ordine,* l'intruso cerca di individuare un collegamento tra tutte le coppie ($j$, $v$) appartenenti al gruppo $\delta(r)$ con le possibili coppie ($j^*$, $v^*$) contenute nel registro.

Si può quindi procedere a considerare interazioni di terzo ordine, quarto ordine etc. Tutti i tentativi sopra descritti sono ipotizzati indipendenti uno dall'altro.

Per fare un esempio della strategia sopra descritta consideriamo una famiglia di

tre individui; padre madre e figlio e supponiamo inoltre che l'intruso sia interessato all'individuazione del "padre".

In accordo con quanto sopra descritto, in primo luogo l'intruso cercherà di individuare direttamente il "padre" e quindi tenterà di effettuare un corretto collegamento tra la combinazione di variabili identificative indirette associate al padre nel file dei dati da rilasciare e le combinazioni delle stesse variabili contenute nel registro in suo possesso.

D'altronde l'intruso può sfruttare l'informazione relativa al gruppo di appartenenza del soggetto di interesse e cercare di individuare la madre e/o il figlio per poi, tramite la relazione di gruppo, individuare il padre stesso. L'informazione relativa al gruppo di appartenenza può essere sfruttata ulteriormente, ovvero l'intruso può cercare di individuare la coppia (padre, madre) e così via.

In realtà il rischio associato alle interazioni di ordine uguale o superiore a due viene considerato non rilevante e quindi trascurabile in quanto la strategia sopra descritta risulta essere sempre più onerosa all'aumentare della dimensione dell'ordine di interazione.

Per quanto riguarda la definizione di una funzione del rischio di violazione che tenga conto della struttura gerarchica del file e quindi delle interazioni di primo ordine, si considera la formula di Boole per eventi mutuamente esclusivi. Indicando con $e_x^{ind}$ il rischio o evidenza associata al record $r$ calcolata nel caso di indipendenza (6.5) otteniamo:

$$e_r^{ger} = (1 - e_r^{ind}) \sum_{\substack{j:h(j)=\delta(r) \\ j \neq r}} e_j^{ind} \prod_{\substack{v:h(v)=\delta(r) \\ v<j \\ v \neq j}} (1 - e_v^{ind})$$

dove $\prod_{\substack{v:h(v)=\delta(r) \\ v<s \\ v \neq s}} (1 - e_v^{ind}) = 1$ quando non ci sono valori di $v$ che soddisfano la relazione $v<j$.

Possiamo notare che il rischio per i record appartenenti allo stesso gruppo è costante. Ciò è legato al fatto che, grazie all'informazione relativa al gruppo di appartenenza, l'intruso, una volta individuato uno dei componenti del gruppo, ha contemporaneamente individuato anche gli altri componenti.

Anche se entrambi i metodi sopra descritti prevedono la costruzione di un rischio individuale le differenze sono comunque sostanziali

Infatti mentre Skinner e Holmes definiscono un rischio esclusivamente per i record che risultano essere casi unici nel file da rilasciare, Franconi e Benedetti definiscono un rischio per tutti i record contenuti in detto file, ciò fa si che in quest'ultimo approccio è possibile tener conto anche della eventuale struttura gerarchica del file stesso.

### 6.3.2 Metodologia Cbs

La metodologia di controllo del rischio di violazione che analizzeremo in questo paragrafo è stata sviluppata all'interno dell'Istituto nazionale di statistica olandese (Cbs) come riportata in Willenborg e de Waal, 1996, e Willenborg e de Waal, 2001.

Un importante concetto nell'ambito di questa teoria è la cosiddetta *"chiave"*, che indicheremo con *k*. Definiamo con il termine *chiave* una generica combinazione di modalità di variabili identificative indirette. Ad esempio supponiamo di avere tre variabili identificative indirette (sesso, stato civile, regione), le combinazioni delle modalità "maschio×celibe", "maschio×Lazio" o anche "maschio×celibe×Lazio", sono delle chiavi.

Da notare che ogni singolo record contenuto nel file dei microdati, così come le unità contenute nel registro esterno, è caratterizzato da più di una chiave in quanto nella stessa definizione non è fissato il numero delle variabili identificative indirette coinvolte nella combinazione.

La chiave è utilizzata per l'identificazione di un generico record contenuto nel file dei microdati e, in particolare, ciò può avvenire nel caso in cui l'unità da cui proviene il record è unica nella popolazione rispetto alla chiave considerata. Lo scopo della metodologia di Willenborg e de Waal è quello di evitare che tale situazione si verifichi. A tale scopo si applicano misure protettive ai record che provengono da unità che risultano essere casi unici nella popolazione rispetto ad una generica chiave. In pratica però non è detto che sia corretto proteggere esclusivamente tali record. Per questo si introduce il concetto di *rarità* e si considera la strategia che evita la presenza di record che presentano chiavi *rare* piuttosto che controllare esclusivamente i casi unici.

Per poter definire come *rara* una chiave è necessario fissare un valore soglia $D_k$, valore che può dipendere dalla chiave *k* considerata. Definiremo quindi una combinazione di modalità di variabili identificative indirette rara se essa si verifica nella popolazione per non più di $D_k$ volte.

Ovviamente non sono note le frequenze delle chiavi nella popolazione ($F_k$) ed inoltre in molti casi si dispone esclusivamente del file dei microdati da rilasciare. Quindi, per poter stabilire se una chiave è rara o meno, è necessaria una stima della frequenza nella popolazione della chiave stessa. Verrà considerata rara la chiave per la quale la frequenza stimata $\hat{F}_k$ risulta minore del valore soglia $D_k$. Successivamente verranno descritti gli stimatori maggiormente utilizzati per ottenere la stima $\hat{F}_k$.

Una volta definito il concetto di rarità per una combinazione chiave bisogna individuare quali combinazioni devono essere controllate. La determinazione di tale insieme di combinazioni dipende dal livello di rischio che l'istituto che rilascia i dati è disposto ad accettare. In pratica, per semplificare la definizione di tale insieme gli autori suddividono le variabili identificative indirette in gruppi sulla base dei quali è poi possibile definire le combinazioni da controllare.

Ad esempio la metodologia suddivide le variabili identificative indirette associando ad esse un particolare livello di "identificabilità"; ovvero *variabili estremamente identificative*, *variabili molto identificative, variabili identificative.* I criteri da utilizzare per una tale suddivisione in classi possono essere diversi, ad esempio si può far riferimento alla rarità, alla visibilità o alla rintracciabilità di particolari valori delle variabili stesse. Nelle intenzioni degli autori esiste una struttura gerarchica nella suddivisione in classi secondo i livelli di identificabilità sopra considerati, ovvero una variabile definita estremamente identificativa è anche molto identificativa per definizione e così via. Una volta associato un livello di identificabilità

a ciascuna variabile identificativa indiretta, è possibile decidere quali chiavi devono essere controllate. Per esempio è possibile richiedere un controllo solo per le chiavi del tipo *variabile estremamente identificativa* × *variabile molto identificativa* × *variabile identificativa* (nella definizione delle combinazioni bisogna ricordare la struttura gerarchica dei livelli di identificabilità). Possiamo notare che in una situazione come quella appena descritta verranno controllate chiavi che coinvolgono solo tre variabili identificative indirette.

La procedura appena descritta, unitamente ad alcune procedure di protezione, è stata implementata in un software statistico, μ-Argus, sviluppato dallo stesso Istituto di statistica olandese all'interno del progetto europeo Casc (*Computational Aspects of Statistics Confidentiality*) sulla tutela della riservatezza dei dati (per maggiori informazioni si rimanda al sito http://neon.vb.cbs.nl/casc/default.htm).

#### 6.3.2.1 Stimatori delle frequenze nella popolazione nella metodologia Cbs

Nel seguito descriveremo molto brevemente gli stimatori più frequentemente utilizzati per la stima delle frequenze nella popolazione delle combinazioni di modalità di variabili identificative indirette $F_k$. In particolare l'interesse è stato rivolto all'individuazione di un buon stimatore per la quantità $F_{i,k}$, la frequenza della chiave $k$ nella regione $i$, e definiscono una notazione che si accordi con quella utilizzata nella letteratura per la stima per piccole aree.

#### 6.3.2.2 Stimatore indiretto

Sia $f_{i,k}$ la frequenza campionaria della chiave $k$ nella regione $i$ ed $n_i$ il numero degli individui campionati nella regione $i$. La frazione delle persone che presentano la combinazione di modalità $k$ nella regione $i$, indicata con $\upsilon_{i,k}$, può essere stimata con $\hat{\upsilon}_{i,k} = f_{i,k} / n_i$. Lo stimatore *diretto* del numero delle persone nella regione $i$ che presentano la chiave $k$ è dato da:

$$\hat{F}_{i,k} = \hat{\upsilon}_{i,k} N_i$$

dove $N_i$ è il numero delle persone nella popolazione che vivono nella regione $i$.

Tale stimatore presenta dei problemi nel caso in cui i valori di $f_i$ siano piccoli. Per limitare tali problemi gli autori ipotizzano che gli individui nella popolazione che presentano la chiave $k$ siano distribuiti in maniera omogenea su tutta la popolazione. Sotto queste ipotesi gli autori considerano la stima della frequenza relativa $\hat{\upsilon}_k = \sum_i f_{i,k} / \sum_i n_i$ e definiscono lo *stimatore sintetico* come:

$$\hat{F}_{i,k} = \hat{\upsilon}_k N_i .$$

Tale stimatore risulta avere una varianza minore del precedente, ma nel caso in

cui non sia soddisfatta la condizione di omogeneità di distribuzione degli individui che presentano la combinazione *k*, tale stimatore risulta essere fortemente distorto. Per mediare le proprietà dei due stimatori appena descritti gli autori considerano uno stimatore ulteriore detto *stimatore di compromesso* dato dalla combinazione dei due stimatori considerati in precedenza $\hat{\upsilon}_{i,k}$ , $\hat{\upsilon}_k$ ovvero:

$$\hat{\hat{\upsilon}}_{i,k} = W_i \hat{\upsilon}_{i,k} + (1 - W_i)\hat{\upsilon}_k$$

dove $0 \leq W_i \leq 1$ è scelto in modo tale da minimizzare l'errore quadratico medio.

# Capitolo 7. Rischio di violazione di dati elementari di impresa(*)

## 7.1 Introduzione

Nella pratica, trattare dati sulle persone fisiche è sostanzialmente diverso dal trattare dati sulle imprese, e questo si riflette necessariamente anche sui metodi di protezione. Per i dati sociali sono generalmente ritenute sufficienti tecniche di protezione non perturbative, che comportano una riduzione, ma non un'alterazione, del contenuto informativo dell'informazione rilevata.

Nel caso delle imprese, invece, a causa dell'elevato rischio di identificazione connesso a questo tipo di dati per le ragioni che si vedranno in seguito, i metodi proposti sono prevalentemente di tipo "perturbativo": il file di microdati viene prodotto a partire dai dati originali modificandoli, secondo qualche criterio, con l'obiettivo primario di rendere più difficile e incerta l'eventuale identificazione e disincentivare i tentativi di violazione. In caso di re-identificazione l'intruso acquisisce informazioni "alterate" e quindi potenzialmente meno utili perché difformi rispetto alle originali, in misura non nota all'utente stesso. E' chiaro che l'efficacia di questi metodi dipende dall'entità delle modifiche apportate ai dati e che la stessa incide sulla qualità dell'informazione rilasciata. Questi due aspetti sono, generalmente, in contrasto tra di loro. Obiettivo delle tecniche di perturbazione è quello di trovare il miglior equilibrio che garantisca la protezione del dato e riduca al massimo la perdita di informazione.

Un'osservazione si rende necessaria riguardo alle tecniche di protezione dei dati di impresa e riguarda il rischio di identificazione. Nel caso dei dati sociali il rischio di identificazione è generalmente misurato *a priori* della protezione dei dati, come visto nel Capitolo 6 e ridotto con opportuni provvedimenti che garantiscono il rispetto di una soglia di rischio fissata. Nel caso dei metodi di protezione proposti per i dati di impresa questo controllo può essere fatto solo *a posteriori*, verificando che i dati prodotti non consentano l'identificazione di qualche unità statistica, anche se in alcuni casi si potrebbe ragionevolmente assumere che tale eventualità sarebbe molto remota. Le proposte per effettuare questa operazione afferiscono alle tecniche di *record linkage* e di abbinamento statistico (*statistical matching*), ma una soluzione soddisfacente non è ancora stata raggiunta sia per la complessità computazionale del problema che per la difficoltà di rappresentare i comportamenti dei possibili intruder e le informazioni a loro disposizione, operazione quest'ultima che introduce inevitabilmente degli elementi di arbitrarietà.

## 7.2 Fattori di rischio nei dati di impresa

La comunicazione della presidenza Istat prot. n.SP/250.94 del 23.3.94 permette il rilascio ad utenti esterni al Sistan di collezioni campionarie di dati elementari resi anonimi per quanto concerne indagini su individui e famiglie, evidenziando che il

(*) Capitolo redatto da Giovanni Seri

rischio di identificazione delle unità statistiche in ambito economico risulta troppo elevato soprattutto per le imprese grandi e/o quelle che vengono incluse nel campione con probabilità di inclusione uno. D'altro canto le esperienze di rilascio di microdati di impresa riportate in letteratura sono rare e molti tentativi di produrre file di questo tipo non sono andati a buon fine o perché non si è riusciti a ridurre sufficientemente il rischio di identificazione o perché non si è riusciti a riprodurre con qualità accettabile le peculiarità dei dati originali (Mc Guckin e Nguyen, 1990). Infatti, i metodi statistici utilizzati per limitare il rischio di identificazione nel caso di file di microdati per gli individui non sono efficaci nel caso delle imprese e quelli specifici proposti per la protezione dei dati economici hanno un impatto più consistente sul contenuto informativo del file. Tale limitazione è dovuta alla natura stessa dell'impresa come unità di rilevazione e alle sue caratteristiche in termini di distribuzione per classi dimensionali. In particolare, le principali caratteristiche che rendono difficoltosa la tutela della riservatezza di tale tipologia di dati possono essere così sintetizzate:

- *la popolazione:* l'identificabilità dei dati di impresa è facilitata dal tipo di popolazione. Le popolazioni di imprese sono sparse e con distribuzioni fortemente asimmetriche: ciò significa che alcune aziende, soprattutto quelle di elevate dimensioni, possono essere riconosciute facilmente (Cox, 1995b);
- *la distribuzione territoriale:* esistono aree ad alta concentrazione di imprese e di conseguenza, al di fuori di queste realtà locali, le imprese possono essere facilmente identificabili;
- *i legami gerarchici:* le imprese sono caratterizzate da legami aventi struttura gerarchica - ciascuna impresa è suddivisa in unità locali e talvolta può appartenere ad un gruppo o cartello di imprese. Ciò facilita i collegamenti tra dati di impresa e, di conseguenza, facilita la violazione della riservatezza: è infatti possibile analizzare i dati collegando le unità locali all'impresa e le imprese al gruppo;
- *il disegno campionario:* in alcune indagini, per ottenere una visione non distorta e rappresentativa del fenomeno in studio, è necessario includere (con probabilità uno) le grandi imprese, che sono altamente riconoscibili e per le quali è indispensabile sia l'accuratezza che la completezza delle informazioni. Il disegno campionario deve prevedere che queste imprese vengano sempre inserite nel campione, fornendo così ulteriori informazioni all'utente esterno che tenti l'identificazione (Cox, 1995b). Per gli stessi motivi di completezza e accuratezza delle informazioni non è sempre possibile rimuovere i record relativi a queste imprese dal file prodotto;
- *la "motivazione":* esiste un interesse particolare nel violare la riservatezza delle imprese. In genere, la riservatezza di un individuo rappresenta un problema sociale che riguarda la sua privacy; per quanto concerne l'impresa, invece, l'interesse si rivolge essenzialmente ad informazioni di tipo economico e quindi a delicati meccanismi di concorrenza di mercato.

Il rischio di violazione, interpretato come rischio di identificazione, porta intuitivamente a considerare maggiormente identificabili le unità statistiche che presentano caratteristiche uniche, o comunque rare, nella popolazione rispetto alle

variabili che vengono rilevate sulle imprese. In particolare, rispetto a variabili strutturali come la classificazione delle unità statistiche secondo l'attività economica prevalente e la collocazione geografica. In Italia la classificazione delle attività economiche (Ateco) ha una struttura gerarchica su sei livelli e consente un dettaglio di informazione notevole, tant'è che raramente si riesce ad utilizzare l'ultimo livello per la diffusione. Analogamente il dettaglio geografico che è normalmente basato sulle suddivisioni amministrative del territorio. In molti casi basta avere queste due informazioni a un livello sufficientemente dettagliato per individuare un'impresa. Per questo, in genere, la popolazione che viene presa in considerazione per l'applicazione di metodi di protezione è definita proprio dalla combinazione di queste due variabili tenendo conto della numerosità e delle caratteristiche delle imprese presenti.

La dimensione dell'impresa è un aspetto critico da considerare. Infatti, molte variabili rilevate sulle imprese sono di natura quantitativa e rappresentano direttamente o indirettamente (variabili *proxy*) la dimensione di impresa. Anche per questo uno schema della valutazione del rischio di identificazione basato sulla "rarità" di certe caratteristiche nella popolazione rispetto a un insieme di variabili chiave mal si adatta al caso delle imprese. Infatti, poiché le variabili quantitative *proxy* della dimensione di impresa si prestano naturalmente come variabili chiave (identificativi indiretti) il rischio di re-identificazione per le imprese è generalmente molto elevato perché le stesse presentano caratteristiche uniche (o rare) nella popolazione rispetto a un (limitato) insieme di variabili chiave. Per fare un esempio, supponiamo di aver limitato la popolazione alla ripartizione geografica del Nord-ovest e al settore di attività economica "Produzione di mezzi di trasporto". E' sufficiente che venga rilasciato una caratteristica rappresentativa della dimensione dell'impresa come il "fatturato", il "numero di addetti" o il "costo per acquisto di materie prime", perché un'impresa grande e ben nota come la Fiat venga quasi certamente riconosciuta.

# Capitolo 8. Le tecniche di protezione per i dati individuali[(*)]

## 8.1 Metodi di protezione: una visione d'insieme.

In questa sezione si presenterà una visione generale dei metodi per la tutela della riservatezza dei dati statistici, secondo un approccio unificante che permetterà di introdurre e discutere in un'ottica globale le diverse proposte.

In linea del tutto generale, per il rilascio di dati protetti gli Istituti nazionali di statistica possono ricorrere a diverse soluzioni, che vanno dall'accorpamento di due o più classi di modalità adiacenti, alla perturbazione dei valori numerici secondo metodi più o meno sofisticati, all'introduzione di valori mancanti nel *data set* originale, al rilascio di valori artificiali, non rappresentanti cioè alcuno degli individui effettivamente rilevati.

Le strategie sopra menzionate possono essere ricondotte a due grandi categorie:

- *Coarsening* dei dati, che consiste nell'uso di trasformazioni che possono essere applicate alla matrice dei dati originali tanto nel senso delle unità, quanto nel senso delle variabili. Tra i vari esempi di *coarsening* troviamo la ricodifica globale, la microaggregazione, la perturbazione. L'introduzione di valori mancanti (soppressione locale), che include come caso particolare l'estrazione di sottocampioni, rappresenta una versione estrema del paradigma della trasformazione.
- Simulazione di record o di interi campioni di dati artificiali.

Il *coarsening* consiste nel trasformare i dati tramite l'applicazione di operatori deterministici o aleatori, siano essi invertibili o singolari.

In questa prima categoria si può collocare la proposta di Little (1993) che suggerisce di rilasciare una sintesi dei dati, ad esempio un insieme di statistiche sufficienti per il modello statistico di base. Un esempio di questa strategia è la pubblicazione, per dati categorici, di tabelle marginali appartenenti alla configurazione sufficiente minimale per un dato modello loglineare non saturato. L'uso di statistiche sufficienti costituisce al tempo stesso un esempio di trasformazione non invertibile, a meno naturalmente che la statistica sufficiente utilizzata non consegua alcuna riduzione dei dati. Come estremizzazione di tale approccio possiamo citare i metodi di soppressione locale, che riducono il rischio di violazione della riservatezza introducendo artificialmente valori mancanti.

In entrambi i casi sopra menzionati, Little (*ibidem*) discute le inferenze ottenibili dai dati protetti sulla base di verosimiglianze esatte o approssimate (pseudo-verosimiglianze) avvalendosi dell'algoritmo EM (Dempster *et al.*, 1977; Little e Rubin, 1987).

L'imputazione completa, ossia la generazione di un intero campione di unità artificiali, è un'alternativa percorribile dagli istituti statistici nazionali. L'idea di rilasciare un campione sintetico offre infatti l'opportunità di evitare qualunque rischio di violazione della riservatezza, dal momento che lo Stato tutela esclusivamente la

(*) Capitolo redatto da Silvia Polettini eccetto il paragrafo 8.4 redatto da Giovanni Seri

riservatezza degli individui *reali*. In quest'ottica, Rubin (1993) propone l'uso dell'imputazione multipla ed afferma la superiorità di quest'approccio rispetto ad altre tecniche di protezione dei dati. Difficoltà nell'applicazione di tale metodologia sono documentate in Kennickell (1999); un'esperienza applicativa complessa che fa ricorso allo stesso approccio è riportata in Abowd e Woodcock (2001); per quanto riguarda gli sviluppi più strettamente metodologici, l'applicazione delle tecniche di imputazione multipla ai problemi di tutela della riservatezza e le strategie di stima sono discusse in Ragunathan *et al (2003).* e in diversi altri lavori (si veda ad esempio Reiter, 2002 e 2003).

L'idea della simulazione è strettamente collegata al concetto che l'informazione statistica fornita dai dati è racchiusa nella verosimiglianza piuttosto che nei singoli valori individuali. Di conseguenza, dal punto di vista inferenziale, un modello che ben rappresenti i dati può sostituire l'intero *data set* (si veda in proposito Kooiman, 1998). In alternativa, un campione (o un insieme di campioni, secondo l'ottica dell'imputazione multipla) simulato da tale modello può rappresentare una strategia più vicina alle aspettative degli utilizzatori. Di fatto, a partire dalla proposta di Rubin (1993), diversi autori hanno esplorato la possibilità di rilasciare campioni di dati artificiali: si veda in proposito Fienberg *et al.*(1998) e, più recentemente, Grim *et al.* (2001) e Dandekar *et al.* (2001). La scelta del modello che genera i dati è in ogni caso guidata dall'obiettivo principale, che consiste nel riprodurre in media alcune prefissate caratteristiche del campione. Analogo obiettivo è perseguito nel lavoro di Burridge (2003), che propone una strategia di simulazione atta a riprodurre in modo esatto le statistiche sufficienti del modello che si suppone sia di interesse per l'utente. Diversamente dagli approcci sopra discussi, quest'ultimo ha la peculiarità di restituire, sulla base di dati artificiali, gli stessi risultati inferenziali che si otterrebbero utilizzando i dati reali.

## 8.2 Modelli di protezione

Si noti che tutte le procedure menzionate nel paragrafo precedente possono essere classificate come strategie di *imputazione,* consistendo tutte nella formalizzazione di un *modello di protezione* e nel rilascio, in sostituzione dei dati originali, di *valori generati* dal modello in questione.

A nostro avviso, l'ingrediente basilare di qualunque tecnica di tutela della riservatezza è il modello di protezione. Come evidenziato dagli esempi visti finora, per modello di protezione si intende una relazione che lega i valori protetti ai valori osservati tramite una qualche trasformazione.

È opportuno a questo punto introdurre la notazione che verrà adottata in seguito: si considererà una matrice $\mathbf{X}$ di dati osservati di dimensione $n$ per $k$. Al solito, le righe di $\mathbf{X}$ corrispondono alle unità rilevate, le colonne alle variabili osservate sugli $n$ individui; le singole variabili verranno indicate con il simbolo $X_l$, $l = 1,\ldots,k$. Con una tilde si indicheranno le corrispondenti quantità rilasciate, di modo che $\tilde{\mathbf{X}}$ denoterà la matrice rilasciata, $\tilde{X}_l$ la variabile rilasciata $l$-ma e così via.

Usando la notazione appena introdotta, si deduce che il *modello di protezione* sopra richiamato, che può essere espresso formalmente tramite la relazione

$$\tilde{\mathbf{X}} = m(\mathbf{X})$$

permette di specificare, direttamente o attraverso assunzioni sulla famiglia di leggi che governano la matrice $\mathbf{X}$, una classe di distribuzioni per i dati da rilasciare $\tilde{\mathbf{X}}$. Il grado di specificazione della componente distribuzionale del modello di protezione varia da modello a modello: alcuni metodi non utilizzano alcuna assunzione in merito alla distribuzione da cui provengono i dati, altri invece specificano una classe parametrica per la legge di probabilità dei dati rilasciati, di solito tramite assunzioni sulla matrice dei dati originali. Talvolta, inoltre, si assegna una distribuzione prefissata soltanto ad una componente del modello di protezione. In alcuni casi, poi, vengono specificate soltanto alcune caratteristiche del modello distribuzionale, ad esempio le medie condizionate.

In quest'ottica si distingueranno i *modelli di protezione* in non parametrici, semi parametrici e completamente parametrici; a partire da tali modelli, si classificheranno i *metodi* proposti in letteratura per la tutela della riservatezza come non parametrici, semi parametrici e parametrici. Questa classificazione si basa sull'assunto che è il grado di formalizzazione del modello a rendere le strategie intimamente differenti tra loro.

### 8.2.1 Metodi non parametrici per la tutela della riservatezza

Supponiamo che la distribuzione di $\tilde{\mathbf{X}}$ sia completamente generale e che il modello per la matrice rilasciata $\tilde{\mathbf{X}}$ abbia la forma di un mascheramento matriciale (*matrix masking*),

$$\tilde{\mathbf{X}} = \mathbf{XB}. \tag{8.1}$$

Come si vedrà più nel dettaglio nel Paragrafo 8.3.11, l'ultima espressione rappresenta una notazione compatta che comprende diverse procedure di protezione, come discusso in Little (1993) e formalizzato in Cox (1994). Cox ha dimostrato che, a seconda della forma che assume la matrice $\mathbf{B}$, detta matrice di trasformazione delle modalità, questo modello di protezione produce, tra l'altro, dati protetti secondo soppressione locale, microaggregazione, *data swapping*.

L'introduzione di una componente additiva nel modello di mascheramento (8.1) lo generalizza ulteriormente a comprendere altri tipi di trasformazione, quali ad esempio la censura (*topcoding*); in questo caso il modello assume la forma $\tilde{\mathbf{X}} = \mathbf{XB} + \mathbf{C}$.

L'esclusione di unità selezionate è poi conseguita nel modello (8.1) tramite l'introduzione di una nuova matrice di mascheramento $\mathbf{A}$, detta matrice di trasformazione delle unità: $\tilde{\mathbf{X}} = \mathbf{AX}$.

Infine, l'esclusione di alcune unità selezionate (sottocampionamento) seguita dalla soppressione di alcune modalità o valori di record predeterminati è ottenuta tramite il modello più generale $\tilde{\mathbf{X}} = \mathbf{AXB}$; in effetti, Cox (*ibidem*) utilizza la formalizzazione $\tilde{\mathbf{X}} = \mathbf{AXB} + \mathbf{C}$, che racchiude tutti i casi precedenti.

Quanto all'uso di metodi di simulazione per la tutela della riservatezza, strategie di protezione di tipo non parametrico possono essere generate facendo ricorso a procedure quali il bootstrap, o versioni modificate di esso. Un esempio è fornito dal metodo proposto in Dandekar *et al.* (2001), basato sul campionamento da ipercubi

latini, in cui la funzione di ripartizione empirica è usata per creare intervalli equiprobabili che permettono di utilizzare per l'estrazione di unità artificiali una procedura di campionamento stratificato. Il lavoro di Fienberg *et al.* (1998) discute strategie analoghe che classifichiamo tra i metodi di protezione non parametrici.

### 8.2.2 Metodi semi parametrici per la tutela della riservatezza

Nel paragrafo precedente, il modello di protezione contiene, per ciò che riguarda la componente distribuzionale, null'altro che la funzione di ripartizione empirica, più costanti note, quali le matrici di trasformazione richiamate al paragrafo precedente.

Attraverso assunzioni relative alle matrici di mascheramento $\mathbf{A},\mathbf{B},\mathbf{C}$ e/o alla matrice dei dati osservati $\mathbf{X}$, è possibile introdurre nel modello di protezione una struttura semi parametrica.

Introduciamo in particolare una matrice casuale $\mathbf{C}$ con distribuzione nota; la matrice protetta $\tilde{\mathbf{X}}$ ottenuta aggiungendo a $\mathbf{X}$ o a una sua trasformazione $\mathbf{AXB}$ una realizzazione di $\mathbf{C}$ rappresenta allora una *perturbazione* dei dati originali. Naturalmente anche a $\mathbf{C}$ è possibile applicare un mascheramento matriciale $\mathbf{D}$ operante sulle variabili, di modo che sia possibile aggiungere disturbo casuale solo alle variabili che necessitano di tale protezione. Quanto alla perturbazione, Duncan e Mukherjee (2000) studiano fino a che punto essa possa essere applicata se si vogliono ottenere inferenze valide, ossia precise sui parametri di interesse; il problema è studiato nel contesto della protezione di database; in particolare, gli autori ottengono limiti da imporre alla varianza della distribuzione di un disturbo normale a media nulla.

Per una disamina approfondita della perturbazione tramite aggiunta di disturbo casuale si consulti Brand (2002) e le citazioni riportate nel lavoro.

Un caso particolarmente semplice di mascheramento semiparametrico è il modello discusso in Little (1993), che sostituisce i dati osservati con la media campionaria più un disturbo casuale, il che è ottenuto ponendo $\mathbf{A} = \mathbf{1}_{n\times k}$.

I modelli del tipo appena richiamato in generale prescrivono per i dati da rilasciare una convoluzione della distribuzione dei dati, eventualmente trasformati in modo opportuno, con la distribuzione del disturbo. La distribuzione dei dati può essere lasciata non specificata, nel qual caso si otterrà una convoluzione della distribuzione del disturbo con la funzione di ripartizione empirica.

Possiamo classificare come semi parametrico anche il modello di imputazione basato sul rilascio di stime di particolari modelli di regressione ottenute con il metodo dei minimi quadrati e perturbate con l'aggiunta di un disturbo casuale; la versione più elementare di questa procedura consiste nel rilascio delle medie campionarie; introducendo variabili di tipo qualitativo tra i regressori e rilasciando quindi le medie condizionate "di strato" opportunamente perturbate si ottiene una procedura che può essere considerata una generalizzazione della microaggregazione.

### 8.2.3 Metodi parametrici per la tutela della riservatezza

Un gradino ulteriore nel processo di specificazione del modello è rappresentato dalla introduzione di una classe di distribuzioni per i dati da rilasciare. Se le variabili sono quantitative continue, molto spesso la scelta ricade sulla normale multivariata, eventualmente in congiunzione con l'uso di una trasformazione dei dati che renda più verosimile tale assunzione.

Un'opzione nella protezione basata su modello è la pubblicazione dei valori stimati a partire da un modello di regressione normale per una o più variabili $X_l$ della matrice dei dati $\mathbf{X}$. Tale principio è alla base della proposta in Franconi e Stander (2002). Una variante del metodo basato su regressioni consiste nel rilasciare i valori previsti dal modello più un disturbo casuale, estratto dalla distribuzione dell'errore stimata (si veda in proposito Little, 1993). Tale strategia è tesa a compensare la riduzione della variabilità dei valori stimati rispetto ai valori originali.

Naturalmente la procedura di protezione basata su modelli di regressione può essere applicata soltanto ad un sottogruppo delle unità osservate. Nella notazione dei paragrafi precedenti, ciò può essere ottenuto introducendo nel modello di protezione una matrice $\mathbf{A}$ di selezione delle unità.

Tra i metodi che definiamo parametrici di tutela della riservatezza un ulteriore esempio è costituito dal rilascio di intervalli di predizione per le variabili da proteggere. Tali intervalli possono essere derivati in base ad assunzioni distribuzionali, utilizzando o meno un modello di tipo regressivo. Per una strategia analoga, si veda Franconi e Stander (2000).

Infine, la protezione per simulazione di individui artificiali può essere basata su un modello parametrico completamente specificato. A questa classe di procedure appartiene il modello proposto in Grim *et al.* (2001); si tratta di un modello mistura i cui parametri vengono stimati col metodo della massima verosimiglianza, utilizzando l'algoritmo EM.

Per variabili categoriali, Fienberg *et al.* (1998) propongono di rilasciare *data set* sintetici estratti da un modello loglineare non saturato "che catturi le caratteristiche essenziali dei dati". In pratica, si tratta di individuare un modello loglineare non saturato che abbia un buon accostamento ai dati (misurato, nella fattispecie, con il test del rapporto delle verosimiglianze) e che inoltre contenga i parametri di interazione considerati di interesse per gli utenti. A partire da tale modello, è possibile generare una o più tabelle artificiali, nell'ottica dell'imputazione multipla.

Ancora nel contesto della protezione per simulazione, Franconi e Stander (2000) e (2003), nell'ambito di una formulazione Bayesiana, utilizzano modelli spaziali gerarchici, avvalendosi di metodi Markov Chain Monte Carlo per la stima dei parametri. Il medesimo tipo di simulazione permette di rilasciare, per le variabili da proteggere, intervalli estratti dalla distribuzione predittiva dei dati anziché singoli valori individuali.

Per una breve rassegna dei metodi Bayesiani per la tutela della riservatezza, si può consultare il lavoro di Cox (2000).

## 8.3 Una rassegna dei metodi di protezione per microdati

In questa sezione viene presentata una panoramica generale dei metodi di protezione per il rilascio di microdati, senza operare distinzioni sulla natura dei dati da proteggere. Per l'esposizione dei diversi metodi si è seguito un criterio che tiene conto delle caratteristiche metodologiche delle tecniche di protezione piuttosto che del campo applicativo delle singole tecniche.

Per quanto riguarda l'aspetto applicativo, si sottolinea che non tutte le tecniche che si esporranno hanno diffusione presso gli Istituti nazionali di statistica, sia perché molte di esse sono apparse in letteratura soltanto di recente, sia per la complessità tecnica e computazionale di molti dei metodi proposti, specie per quanto riguarda le tecniche disegnate per la protezione di microdati d'impresa. Le più significative esperienze pratiche di rilascio di microdati avutesi sinora in Istat sono documentate nel Capitolo 9.

### 8.3.1 Ricodifica globale, *topcoding* e arrotondamento

Si tratta dei trattamenti dei dati più elementari; la ricodifica globale consiste nell'accorpamento di classi adiacenti o similari per variabili qualitative, ad esempio l'area geografica o quella di attività economica e nella riduzione in classi (censura per intervallo) per variabili continue, quali ad esempio il fatturato, il numero di addetti e così via. Si tratta ovviamente di una procedura di protezione basata sull'uso di funzioni non invertibili dei dati osservati, ed è altrettanto chiaro che tale metodo implica generalmente una perdita di informazione che può essere notevole. Infine, dal punto di vista dell'utente, la presenza di variabili ridotte in classi crea problemi nell'utilizzo di tecniche disegnate per dati quantitativi.

Un caso particolare di ricodifica, da effettuarsi su un sottogruppo dei dati osservati o congiuntamente alla censura per intervallo, si applica ai valori che si collocano sulle code della distribuzione della variabile da proteggere. Si tratta del cosiddetto *topcoding* (censura), in cui al dato osservato viene sostituito un intervallo aperto a destra o a sinistra a seconda della coda in cui si colloca il dato da proteggere.

L'arrotondamento o *rounding* si applica a variabili quantitative e consiste nell'esprimere i valori osservati modulo una base prescelta. Sostanzialmente, si tratta anche in questo caso di una riclassificazione dei dati.

Nel caso in cui la variabile da proteggere sia quantitativa, e si possa formalizzare un modello in base al quale derivare una verosimiglianza per un parametro di interesse, è possibile esprimere tale verosimiglianza in termini dei dati originali non raggruppati e,come descritto da Little, utilizzare per la stima del parametro l'algoritmo EM o tecniche di imputazione multipla che estraggano campioni da ogni intervallo, dove la distribuzione può, ad esempio, essere assunta uniforme. Tuttavia, se la distribuzione delle variabili originali è particolarmente asimmetrica, come spesso avviene nelle applicazioni economiche, l'utilizzo di sottodistribuzioni uniformi risulta inadeguato, e inoltre le stime basate sull'imputazione multipla risultano in questo caso piuttosto sensibili alle assunzioni relative alla coda destra della distribuzione.

### 8.3.2 Soppressione locale

Anche questa tecnica rappresenta un trattamento piuttosto elementare dei dati elementari, consistendo nella sostituzione, per alcuni individui, del valore osservato con il codice "valore mancante" in una o più variabili. Prioritariamente all'applicazione della tecnica è necessario disporre di un *criterio di selezione delle unità* a cui applicare la soppressione locale, e in seguito di un *criterio di selezione delle variabili* di cui rimuovere l'informazione registrata sui record precedentemente individuati.

Per quanto riguarda la selezione delle unità, si tratta di individuare i record maggiormente a rischio. A questo scopo, sono particolarmente indicate le tecniche di stima del rischio per dati sociali descritte nel Capitolo 6. In altri casi, si può pensare di utilizzare un criterio di distanza (unidimensionale o multidimensionale) in base al quale gli outlier sono considerati maggiormente a rischio; per i record così individuati, i valori di alcune variabili (ad esempio variabili di tipo monetario, o l'informazione sull'appartenenza geografica delle unità) sono sostituiti dal codice "valore mancante". Chiaramente la tecnica assicura protezione inducendo una perdita di informazione che può essere molto elevata; inoltre, essa introduce valori mancanti con un meccanismo che non è del tipo *missing at random* (Mar) o *missing completely at random* (Mcar), producendo insiemi di dati che non possono essere analizzati dall'utente con le tecniche tradizionali, nemmeno a discapito di perdite di efficienza (si noti che in questo caso l'eliminazione dei valori mancanti provoca distorsione).

Per queste ragioni di fatto la tecnica viene utilizzata nella pratica insieme ad altre misure di protezione, in modo da mitigare gli effetti di cui si è parlato.

Quanto alla selezione delle variabili da sopprimere sui singoli record selezionati, è necessario definire un criterio di ottimalità; naturalmente è opportuno che esso tenga conto della perdita di informazione associata all'applicazione del metodo; a questo proposito, si veda de Waal e Willenborg (1998).

### 8.3.3 *Data Swapping*

Sebbene sia molto utilizzata per dati categorici che quindi conducono alla creazione di tabelle, il *data swapping* (Dalenius e Reiss, 1982) è una tecnica di protezione che opera al livello dei microdati. La protezione in questo caso è ottenuta modificando una frazione dei record; la procedura consiste nello scambiando, in un insieme selezionato di coppie di record (*swap pairs*), un sottoinsieme delle variabili. Naturalmente questo scambio di valori delle variabili tra coppie di record rende difficile l'identificazione in quanto non viene reso noto l'insieme delle coppie scambiate né l'insieme delle variabili su cui è applicato lo *swapping*.

La tecnica necessita di un criterio per scegliere i record da scambiare, le variabili da scambiare, il tasso di *swapping*, fattori che concorrono tutti alla determinazione della *performance* del metodo in termini di protezione del file di microdati. Per dati categorici, una scelta frequente è di applicare lo *swapping* a unici o quasi unici campionari secondo una data classificazione. Infatti si tratta dei record potenzialmente più a rischio. In genere per questo tipo di dati si utilizza una versione vincolata di

*swapping*, in modo da preservare le distribuzioni di frequenza, obiettivo particolarmente importante quando i microdati vengano utilizzati per la creazione di tabelle. Si noti tuttavia che pur essendo in grado di mantenere le frequenze invariate, la tecnica non preserva statistiche ausiliare, quali le intensità: lo *swapping* può preservare ad esempio il numero di imprese di un certo tipo, ma non il *turnover* di tale gruppo di imprese.

Per dati quantitativi o almeno ordinabili è possibile utilizzare il cosiddetto *rank swapping* o *rank based proximity swapping* che, al fine di minimizzare le distorsioni, sfrutta il campione ordinato e scambia il valore osservato $X_i$ con un suo "vicino", un record cioè il cui rango differisca da quello di $X_i$ meno di una data soglia.

In generale, l'effetto della tecnica è di deteriorare notevolmente la qualità del dato; in particolare, le associazioni tra variabili vengono usualmente modificate (si veda in proposito il lavoro di Carlson e Salabasis, 1998).

Gomatam, Karr, e Sanil (2003) riportano valutazioni relative a rischio e utilità nell'uso del *record swapping* per dati categorici. In particolare gli autori, confrontando misure di distanza tra le distribuzioni prima e dopo la protezione, osservano, come è naturale attendersi, che la distorsione tende a crescere e il rischio a decrescere all'aumentare del tasso di *swapping*; Inoltre, scambi su una singola variabile sono meno protettivi di scambi effettuati su più variabili.

### 8.3.4 Aggiunta di disturbo

Una delle prime proposte per la protezione di dati quantitativi consiste nell'aggiunta di disturbo ai valori originali. Il metodo è stato formalizzato da Kim (1986), Sullivan e Fuller (1989); Fuller (1993). Per una rassegna integrata, si veda Brand (2002).

Supponiamo per semplicità di proteggere una sola variabile $X^j$; il modello di protezione può essere formalizzato come $\tilde{X}_i^j = X_i^j + \varepsilon_i$ .In esso il disturbo è applicato ad ogni unità in modo indipendente; se la componente $\varepsilon$ è una variabile aleatoria a media zero e varianza $\sigma^2$ fissata, incorrelata con $X^j$, si deduce facilmente che la variabile protetta $\tilde{X}^j$ ha la stessa media di $X^j$ e varianza pari a $Var(X^j) + \sigma^2$. La procedura appena descritta può essere considerata di tipo semi-parametrico, in quanto la distribuzione da cui sono generati i dati osservati non è specificata. La distribuzione da cui sono generati i disturbi può essere scelta in vario modo; si usa parlare di *white noise* nel caso in cui si utilizzi la distribuzione normale.

Passando alla perturbazione di un'intera matrice di variabili quantitative, definiamo il vettore *k*-dimensionale $\boldsymbol{\varepsilon} \sim F(0, \Sigma_\varepsilon)$, con componenti incorrelate con le variabili di $\mathbf{X}$. La scrittura $F(0, \Sigma_\varepsilon)$ indica una distribuzione generica di vettore delle medie 0 e matrice di varianza e covarianza $\Sigma_\varepsilon$. Supponiamo che la distribuzione empirica delle variabili di $\mathbf{X}$ sia caratterizzata da un vettore di medie $\mu$ e una matrice di varianze e covarianze $\Sigma$. Il modello:

$$\tilde{\mathbf{X}} = \mathbf{X} + \boldsymbol{\varepsilon} \tag{8.2}$$

induce sulle variabili perturbate una media campionaria invariata rispetto ai dati

originali e una matrice di varianza pari a $\tilde{\Sigma} = \Sigma + \Sigma_\varepsilon$. Da quanto sopra si deduce che la perturbazione induce una modifica nella struttura di varianza dei dati osservati nel senso di un generale incremento dei vari indici (anche se alcuni potrebbero rimanere invariati).

Nel caso in cui, come appena descritto, la componente di disturbo non sia correlata con le variabili da proteggere, si possono utilizzare strutture specifiche della matrice di varianza e covarianza $\Sigma_\varepsilon$ per indurre particolari effetti sulla matrice di varianza e covarianza $\tilde{\Sigma}$. Supponendo di conoscere la varianza delle variabili di interesse, si può pensare di generare un disturbo la cui varianza è proporzionale a quella delle variabili da proteggere: $\boldsymbol{\varepsilon}$ tale che $\Sigma_\varepsilon = \alpha \ \text{diag}\left(\sigma_1^2, \cdots \sigma_k^2\right)$. In questo caso otteniamo due effetti in particolare: in primo luogo, le covarianze tra le variabili non sono modificate, essendo $\Sigma_\varepsilon$ diagonale; inoltre, se $\alpha$ è piccolo, l'aumento nella variabilità di ciascuna variabile risulta contenuto, essendo per ognuna pari a:

$$\text{Var}(\tilde{X}^j) = \sigma_j^2 (1 + \alpha). \tag{8.3}$$

Di fatto, difficilmente le varianze $\sigma_j^2$ sono note, per cui $\Sigma_\varepsilon$ viene fissata ricorrendo agli usuali stimatori non distorti. Utilizzando la formula per la varianza iterata, si perviene comunque allo stesso risultato (8.3).

Analogamente a quanto visto per le varianze, per le correlazioni risulta:

$$r(\tilde{X}^j, \tilde{X}^h) = \frac{r(X^j, X^h)}{1 + \alpha}, \tag{8.4}$$

la quale è valida approssimativamente anche nel caso in cui le varianze siano stimate.

In base alla relazione (8.3) si può notare che i dati osservati non consentono una stima non distorta delle varianze d'interesse, a meno che non si diffonda, assieme alla tecnica di perturbazione utilizzata, anche il valore del coefficiente $\alpha$. In tal caso, utilizzando la varianza campionaria dei dati perturbati che denotiamo con $\tilde{S}_j^2$ uno stimatore corretto di $\sigma_j^2$ è dato da:

$$\hat{\sigma}_j^2 = \tilde{S}_j^2 / \sqrt{1 + \alpha}.$$

Si veda anche Little (1993) per una modellizzazione del dato originale che, sfruttando le informazioni sul modello di perturbazione, consente di ottenere gli stimatori di massima verosimiglianza per i parametri di interesse e le relazioni che legano i primi ai parametri del modello di perturbazione.

Come già notato, il modello (8.2) con disturbo a componenti incorrelate ($\Sigma_\varepsilon$ diagonale) lascia invariate le covarianze tra le variabili, ma non le varianze e quindi modifica le correlazioni come visto in (8.4). Tuttavia, essendo le correlazioni l'oggetto di interesse primario di molte elaborazioni statistiche, può essere preferibile utilizzare una struttura di covarianza non sferica per $\boldsymbol{\varepsilon}$ che consenta di riprodurre nei dati perturbati le correlazioni originali, pur permettendo, se $\alpha$ è noto, di ricostruire i parametri originali come visto poc'anzi. Si noti tuttavia che la conoscenza di $\alpha$ potrebbe essere utilizzata da un intruso per tentare di invertire la perturbazione così da

ottenere stime dei dati originali.

Se il criterio prescelto è quello di riprodurre nei dati rilasciati i parametri originali, l'aggiunta di disturbo a componenti correlate può offrire una soluzione: infatti se $\Sigma_\varepsilon = \alpha\Sigma$, cioè se anche le covarianze tra i disturbi sono scelte proporzionali alle covarianze tra le variabili da perturbare, si ottiene $\tilde{\Sigma} = (1+\alpha)\Sigma$ e la matrice di correlazione costruita sui dati perturbati coincide con quella stimata sui dati originali. Di conseguenza, in questo caso un modello di regressione costruito su variabili protette fornisce le stesse stime di un modello costruito sui dati originali. Tuttavia per i dati perturbati la somma dei quadrati dei residui fornisce una stima distorta della corrispondente quantità relativa ai dati originali, e ciò crea problemi in sede di inferenza. Ad esempio, alcuni coefficienti di regressione potrebbero risultare più significativi utilizzando i dati protetti o viceversa utilizzando i dati originari. Chiaramente ciò crea problemi in sede di inferenza.

Come visto la tecnica della perturbazione permette di conservare alcune delle caratteristiche originali; tuttavia, in accordo con l'usuale *trade-off* tra preservazione del contenuto informativo e confidenzialità, la protezione raggiunta non è completamente soddisfacente. Soprattutto se il meccanismo di protezione è noto nei suoi dettagli (si ricordi che ciò è necessario per stimare correttamente, a partire dai dati perturbati, alcune quantità di interesse), un intruso può sfruttare l'esistenza di una relazione uno a uno tra ciascun valore perturbato e ciascun valore originale e costruire una stima di quest'ultimo.

Per quello che riguarda l'aggiunta di disturbo, una esperienza per la protezione di database pubblicati sul Web è documentata in Duncan e Mukherjee (2000). Per proteggere da attacchi di intrusi un database disegnato per produrre tabelle (cioè dati aggregati), gli autori utilizzano l'aggiunta di un disturbo casuale, da effettuarsi ad ogni interrogazione o query. L'utilizzo di realizzazioni indipendenti tuttavia consentirebbe, replicando le interrogazioni al database, di stimare con grande precisione i parametri di interesse dell'intruso, mentre d'altra parte l'utilizzo dello stesso disturbo per tutte le interrogazioni (cioè una perturbazione permanente del database) provocherebbe una perdita di informazioni. L'aggiunta di un disturbo a *realizzazioni* positivamente correlate nelle diverse interrogazioni consente di ovviare ai problemi appena discussi. Ad esempio, si propone l'utilizzo di disturbo autocorrelato tra le interrogazioni, allo scopo di ottenere una maggiore protezione e una migliore qualità dei dati. Gli autori definiscono una funzione di perdita che tenga conto delle due esigenze contrapposte di minimizzare la perdita di informazione, misurata tramite una combinazione convessa tra la varianza della componente di disturbo e il numero di interrogazioni del database consentite, e di contenere la possibilità di violazione. Il rischio di violazione è definito come funzione inversa della varianza dello stimatore del parametro di interesse dell'intruso. Una volta stabilita una soglia per il rischio, gli autori derivano la combinazione ottimale di varianza del disturbo e numero di interrogazioni consentite, minimizzando la funzione di perdita. La strategia ottimale è valutata anche in corrispondenza a differenti combinazioni delle due componenti della funzione di perdita. Si dimostra che in generale la strategia di perturbazione a realizzazioni indipendenti raramente è preferibile alla restrizione delle interrogazioni permesse, e

perché ciò avvenga la funzione di perdita deve attribuire un peso molto basso alla perdita di informazione derivante dalla perturbazione; se le interrogazioni non sono ristrette, è sempre preferibile l'utilizzo di un disturbo a realizzazioni correlate ad un disturbo a realizzazioni indipendenti; l'uso combinato di perturbazione correlata e restrizione delle interrogazioni è ottimale quando la funzione di perdita attribuisce peso elevato alla perdita di informazione derivante dal disturbo. In particolare al crescere di tale peso diminuisce il numero di query consentite e la varianza "ottimale" del disturbo.

Come nella proposta di Duncan e Mukherjee (*ibidem*), la perturbazione dei microdati può essere utilizzata come metodo di protezione nella creazione di tabelle; questo approccio è stato proposto anche da Evans *et al.* (1998) come alternativa al metodo della soppressione locale.

La scarsa protettività dei metodi di perturbazione tramite aggiunta di disturbo ha indotto i ricercatori a studiare soluzioni più articolate ancora basate sull'aggiunta di disturbo, ma in combinazione con altre trasformazioni –non lineari- dei dati. Una proposta interessante è quella di Sullivan (1989) (si veda anche Sullivan e Fuller, 1989 e Fuller, 1993); essa è basata su una prima trasformazione dei dati utilizzando una versione lisciata (*smooth)* della funzione di ripartizione empirica. Poiché la funzione di ripartizione è una variabile aleatoria uniforme, applicando a questa la funzione di ripartizione inversa di una variabile aleatoria normale standardizzata, si ottiene una variabile aleatoria normale standardizzata. Ad essa si aggiunge un disturbo casuale normale eventualmente a componenti correlate, dopodiché si procede secondo il percorso inverso (a parte qualche modifica aggiuntiva, in sostanza si tratta di estrarre i ranghi e calcolare la funzione di ripartizione lisciata inversa in corrispondenza di tali ranghi) per riottenere un valore perturbato. Per ciascuno dei valori perturbati, il procedimento prevede anche un controllo della distanza dal valore originale di riferimento, che, per limitare il rischio di re-identificazione, non deve essere la minima delle distanze tra il valore perturbato e un qualunque valore del campione originale.

È interessante sottolineare che, con alcuni accorgimenti e opportune modifiche, il meccanismo si applica anche a variabili qualitative.

Per quanto riguarda la validità inferenziale dei dati, il metodo consente di riprodurre approssimativamente le distribuzioni marginali, con medie pressoché inalterate; Brand (2002) nota che, mentre è possibile stimare correttamente le medie tramite le medie campionarie, nel caso di variabili continue, la variabilità aumenta in funzione di una componente che dipende dai dati osservati e dalla numerosità campionaria, che comunque tende a zero al crescere di $n$. Per quanto riguarda le correlazioni, esse possono variare, specie se la distribuzione che ha generato i dati osservati non è di tipo normale; inoltre, l'effetto del disturbo si distribuisce tra le correlazioni in modo non uniforme. Per ovviare a questo problema, Sullivan (1989) propone di aggiustare iterativamente i valori perturbati in modo tale che la differenza tra le correlazioni calcolate sui dati osservati e quelle calcolate sui dati perturbati sia inferiore a una soglia prestabilita. Data la complessità della procedura, non è possibile derivare analiticamente le proprietà di stimatori di altre quantità. Per modelli di regressione, è opportuno quindi stimare i parametri ricorrendo ai modelli di regressione con errori nelle variabili secondo quanto proposto in Fuller (1993).

In generale ogni metodo di mascheramento, ivi compresa l'aggiunta di disturbo,

porta alla corruzione di alcune particolari caratteristiche della distribuzione, in particolare minimi e massimi, code, numero di zeri, e così via. A causa del *trade-off* protezione-informazione, in generale non ci si può aspettare che tutte le caratteristiche della distribuzione dei dati osservati rimangano invariate, perché altrimenti non si farebbe altro che replicare la distribuzione empirica e non si otterrebbe di conseguenza alcun guadagno in protezione. Per quanto riguarda i parametri che usualmente vengono tenuti fissati, essenzialmente medie e matrici di correlazione, c'è inoltre da dire che, se le distribuzioni sono estremamente asimmetriche -come spesso avviene per variabili di tipo economico- il riferimento ai valori medi non è adeguato e può provocare, in congiunzione con l'aggiunta di disturbo, differenze numeriche che possono essere anche notevoli tra i valori osservati dei parametri di riferimento e quelli ottenuti in base ai dati perturbati.

Quanto alla validità inferenziale, Little (1993) nota che, pur preservando, come già visto, alcuni parametri, questa tecnica non fornisce inferenze *valide*, in quanto non tiene conto dell'aggiunta di incertezza derivante dall'operazione di perturbazione. È opportuno quindi rilasciare dei fattori di correzione del tipo già mostrato nel caso di disturbo incorrelato (fattori che sono specifici del particolare metodo utilizzato), o, in alternativa, rilasciare più realizzazioni dei dati perturbati, per consentire l'applicazione di tecniche di stima basate sul criterio dell'imputazione multipla.

Quanto alla protettività della tecnica, Brand (*ibidem*) sottolinea che data la natura della seconda tecnica di perturbazione, che combina trasformazioni non lineari e aggiunta di disturbo, non tutte le variabili ricevono lo stesso livello di protezione.

In generale, tutte le tecniche basate sull'aggiunta di disturbo hanno un basso livello di protezione, in particolar modo la variante con disturbo a componenti correlate, che addirittura risulta superata dall'aggiunta di *white noise*. La tecnica combinata suggerita da Sullivan e Fuller (1989) e Fuller (1993) sembra raggiungere il più elevato livello di protezione nella classe di tecniche di perturbazione basate sull'aggiunta di disturbo; il meccanismo di perturbazione incorpora in questo caso anche un controllo su una funzione di distanza tra dato perturbato e dato originale, che non deve essere inferiore a una soglia prefissata; tuttavia il livello di protettività è generalmente poco soddisfacente, soprattutto se paragonato alla complessità della tecnica e al numero di parametri da specificare.

L'utilizzo di queste tecniche appare dunque sconsigliabile a meno che non venga effettuato in combinazione con altri metodi di protezione, ad esempio il *data swapping* (Dalenius e Reiss, 1982).

### 8.3.5 Imputazione multipla

Rubin (1993) per primo propone l'uso della simulazione, cioè il rilascio di un insieme di microdati artificiali, per la tutela della riservatezza.

La sua proposta è basata sull'imputazione multipla, cioè sul rilascio di più *data set* simulati da uno stesso modello di riferimento. In questa impostazione la presenza di unità artificiali in quanto simulate elimina *a priori* il problema della tutela della riservatezza, ed è strettamente connessa al criterio che l'informazione statistica risiede

nella verosimiglianza e non nelle singole unità. La presenza di più insiemi di dati permette di quantificare, se il numero di replicazioni è sufficientemente elevato, l'impatto che la protezione ha sulle stime, tramite la variabilità tra le replicazioni (si vedano in proposito i lavori di Raghunathan *et al,* 2003 e di Reiter, 2003). Infine, l'uso di *data set* completi permette di analizzare i dati con le usuali tecniche, senza comportare il ricorso a tecniche sviluppate per dati incompleti, raggruppati o mancanti. Va sottolineato che, trattandosi di simulazione, è possibile utilizzare per i dati da generare una numerosità campionaria e un disegno di campionamento diversi da quelli originari.

In linea con la proposta di Rubin (*op. cit*.), un'applicazione dell'imputazione multipla ai problemi di tutela della riservatezza è stata condotta da Kennickell (1999) per i dati dell'indagine sulle finanze dei consumatori. Diversamente dall'impostazione suggerita da Rubin, il *data set* generato non è completamente sintetico, in quanto l'imputazione viene effettuata soltanto su un sottoinsieme dei dati rilevati, facendo uso di un programma costruito per l'*editing* e l'imputazione dei dati mancanti.

Per identificare il sottogruppo di unità su cui effettuare l'imputazione multipla, le famiglie con un rischio di identificazione eccessivamente elevato vengono individuate e i relativi valori monetari rimossi. Il rischio di identificazione viene considerato elevato laddove i livelli di reddito o di benessere risultano estremi. Insieme alle famiglie a rischio viene selezionato un sottocampione casuale di famiglie i cui valori monetari vengono ugualmente rimossi, allo scopo di ottenere una maggiore protezione. Al sottoinsieme così determinato viene applicata l'imputazione multipla, per la quale possono essere utilizzati diversi modelli operativi. In particolare, l'autore propone tre esperimenti: nel primo, i valori da imputare vengono condizionati a variabili che indicano l'intervallo -di ampiezza predefinita- in cui si colloca il valore osservato; nel secondo, si condiziona ai valori osservati; nel terzo, si esclude qualunque condizionamento ai dati rilevati.

La protezione offerta dalla procedura è stata valutata positivamente, così come il contenuto informativo preservato (Fries *et al.,* 1997) in riferimento al primo degli esperimenti riportati. Si noti che in questo caso, essendo il campione solo in parte sintetico (soltanto alcune unità vengono imputate), il problema della tutela della riservatezza degli individui rilasciati non è completamente eliminato, e viene inoltre trascurato il problema di come identificare gli individui da proteggere.

Una trattazione metodologica completa dell'approccio alla tutela della riservatezza basato sull'imputazione multipla, che copre tanto l'imputazione completa quanto quella parziale (imputazione applicata soltanto a un sottogruppo di unità e/o variabili) si trova in Raghunathan *et al.* (2003) e in Reiter (2003).

### 8.3.6 Proposta di Fienberg, Makov e Steele

In linea con il principio della simulazione, gli autori (Fienberg *et al.*, 1998) propongono il rilascio di dati artificiali sulla base di un modello stimato, di tipo non parametrico o parametrico, che ove possibile tenga conto di quelle fonti di errore non campionario che l'Istituto Statistico possa quantificare, come ad esempio procedure di

*editing*, imputazione, accorpamento di fonti di dati diverse tramite *matching,* problemi di copertura e così via.

Una volta individuato il modello, si procede a campionare da esso, anche in questo caso producendo più replicazioni, allo scopo di permettere una stima corretta della varianza dello stimatore del parametro di interesse.

La procedura viene applicata al rilascio di tabelle di contingenza. Nel caso di dati categoriali il modello loglineare è sicuramente il modello multivariato più noto e semplice per descrivere la frequenza delle singole celle. Essenzialmente, l'idea alla base della proposta di Fienberg *et al*. (*ibidem*) consiste nella definizione di un modello loglineare non saturato e nel rilascio di dati generati da un modello condizionato alla configurazione sufficiente minimale di tabelle indotto dal modello loglineare prescelto. Il fatto che si considerino distribuzioni condizionate alle marginali osservate è una caratteristica da un lato positiva perché intrinsecamente coerente con i dati eventualmente pubblicati, dall'altro negativa perché se alcune delle frequenze marginali sono basse, la protezione assicurata dalla strategia non è soddisfacente. Inoltre il fatto che possano esservi frequenze osservate molto basse (dati sparsi) può inficiare la validità del modello stimato e quindi compromettere la riuscita dell'intera strategia.

L'intera strategia è basata sulla generazione di una particolare tabella dall'insieme delle tabelle con marginali uguali a quelle osservate, secondo la tecnica proposta da Diaconis e Sturmfels (1998). Tale strategia risulta tuttavia computazionalmente molto onerosa, tanto da essere inapplicabile già ad esempio ad una tabella di dimensione 5*4*3 (per una discussione del metodo e dei problemi computazionali connessi si può consultare Duncan *et al.*, 2001).

### 8.3.7 Post-Randomizzazione (Pram)

Accenniamo in questa sede anche al metodo della Post-randomizzazione (Gouweleeuw *et al*., 1998) proposto principalmente per variabili chiave di tipo qualitativo, che frequentemente sono presenti in file di dati elementari provenienti da indagini sociali.

Si tratta di una metodologia basata anch'essa su simulazione di un sottogruppo di variabili per l'intero campione.

Sebbene esso possa essere generalizzato anche al caso di variabili continue, il metodo è proposto con riferimento a variabili qualitative. Trattiamo per semplicità il caso della protezione di una singola variabile $X$. In tal caso i valori da rilasciare si ottengono applicando ad ogni elemento della matrice osservata $X$ la relativa probabilità di transizione dalla modalità $i$ alla modalità $j$ della variabile, $p_{jl} = \Pr(\tilde{X}_i = l \mid X_i = j)$. Le probabilità di transizione sono assegnate sulla base delle caratteristiche delle variabili e della loro influenza sul rischio di identificazione, valutato a partire dalla frequenza delle unità campionarie che presentano una determinata stringa di caratteristiche. Anche in questo caso, unici o quasi unici campionari sono considerati maggiormente a rischio; tuttavia, in linea di principio, è possibile applicare la tecnica selezionando gli individui sulla base di una stima del rischio di violazione individuale del tipo visto nel Capitolo 6.

Nel caso di più variabili, naturalmente vanno specificate le probabilità di

transizione dalla k-pla di modalità osservate $j_1, ..., j_k$ alla k-pla $l_1, ..., l_k$. Rispetto a procedere variabile per variabile in modo indipendente, questo approccio consente di rispettare eventuali vincoli strutturali (è possibile ad esempio imporre probabilità di transizione nulle verso combinazioni del tipo posizione nella professione: inabile al lavoro e posizione nella professione: lavoratore dipendente).

Il Pram può essere considerato come una versione randomizzata del data swapping; anche in questa tecnica, infatti, le modalità di alcune variabili sono scambiate tra record, seguendo tuttavia un preciso criterio probabilistico.

La caratteristica principale della tecnica consiste nel generare una matrice di dati $\tilde{\mathbf{X}}$ che, pur non coincidendo con la matrice originaria $\mathbf{X}$, consente di ottenerne le distribuzioni di frequenza (distribuzioni marginali se il Pram è applicato marginalmente, congiunte se il Pram è applicato a gruppi di variabili). Considerando la matrice delle probabilità di transizione $P=\{p_{jl}\}$, si ottiene infatti per la distribuzione della variabile trattata: $E(\tilde{T}) = P'T$, in cui $T$ indica il vettore delle frequenze marginali della variabile $\mathbf{X}$. Assumendo l'invertibilità di $P$ si può pertanto stimare la distribuzione originaria. Una forma particolare di Pram (Pram invariante), in cui $P'T=P$, assicura che la distribuzione marginale della variabile protetta $\tilde{\mathbf{X}}$ sia in media uguale a quella della variabile originaria $\mathbf{X}$. Ciò dà all'utente il notevole vantaggio di poter lavorare direttamente con i dati rilasciati, senza dover apportare correzioni in sede di analisi.

Quanto alla protettività del metodo, la diminuzione del rischio di re-identificazione si basa in primo luogo sulla considerazione che sono le combinazioni meno frequenti ad essere maggiormente a rischio; con questa premessa, la protezione assicurata può essere valutata tramite le probabilità $p_{jj}$, che per ogni $j$ rappresentano la probabilità di una modalità di rimanere invariata nei dati protetti. Vengono valutate soprattutto le modalità o combinazioni di modalità $j$ che erano considerate maggiormente a rischio nel file originario. Sulla base di queste probabilità, Gouweleeuw *et al.* (1998) propongono particolari misure per valutare il rischio atteso (si tratta infatti di una procedura basata su simulazione) del file rilasciato.

### 8.3.8 Metodi di imputazione da modelli di tipo regressivo

In alcuni casi la simulazione è basata esplicitamente su modelli statistici (sovente di tipo regressivo) che descrivono relazioni condizionate tra le variabili.

Un approccio direttamente legato alla proposta di Rubin è descritto da Abowd e Woodcock (2001), Reiter (2002), Raghunathan *et al.* (*op.cit.*). In parte di questi lavori l'imputazione multipla è intesa, diversamente da quanto descritto nel paragrafo precedente, come simulazione (con $M$ replicazioni) di *tutte* le unità del campione sulla base di un modello multivariato definito in base a relazioni di tipo regressivo tra una variabile e un sottogruppo delle rimanenti. Tutti i lavori citati seguono un approccio bayesiano, pertanto i valori simulati vengono estratti dalle distribuzioni predittive delle variabili sensibili dati i parametri del modello e un insieme di variabili pubbliche o non confidenziali.

Una seconda opzione proposta nel primo dei lavori citati è l'utilizzo della distribuzione predittiva dei dati sensibili date le variabili pubbliche o non confidenziali

e un insieme di informazioni aggregate ritenute rilasciabili (medie, tabelle marginali, coefficienti di modelli di regressione, eccetera).

Tra le procedure sviluppate per la protezione di microdati, Franconi e Stander (2000) e (2003) propongono un metodo di protezione basato su modelli bayesiani. Ambedue le proposte sono sviluppate espressamente in riferimento a microdati di impresa.

Anche in questo caso si utilizza un modello di tipo regressivo definito per le variabili sensibili e che include variabili pubbliche o comunque non confidenziali (numero di addetti, valore delle merci esportate, caratteristiche dell'impresa, area geografica). Il valore aggiunto di questa proposta sta nell'utilizzo di un modello spaziale, con effetti casuali strutturati e non strutturati. A differenza della proposta discussa in precedenza, non si fa direttamente ricorso all'imputazione multipla, anche se si utilizza una impostazione di natura simile, in quanto i valori da rilasciare vengono estratti dalla distribuzione predittiva *a posteriori* delle variabili da proteggere; inoltre si propone di rilasciare, in luogo dei valori osservati, intervalli predittivi, il che permette di tener conto della variabilità del modello. Gli intervalli predittivi vengono ricavati a partire dall'output di simulazioni di tipo MCMC dalla distribuzione predittiva del fatturato.

In Franconi e Stander (2002) e Polettini, Franconi e Stander (2002) viene sviluppato anche un secondo metodo di protezione, non bayesiano, anch'esso basato su modelli di tipo regressivo e anch'esso esplicitamente rivolto alla protezione di dati d'impresa.

La proposta è basata sulla costruzione di un modello di regressione per ciascuna variabile sensibile (che funge da risposta nel modello) dell'indagine; in ogni modello, i regressori sono le altre variabili sensibili, più variabili pubbliche o non confidenziali.

Per il rilascio, si utilizzano i valori stimati in ciascun modello per ciascuna variabile sensibile; per ottenere una maggiore protezione, i valori previsti sono ulteriormente modificati con una sorta di *shrinkage* selettivo dei valori di ciascuna variabile da proteggere.

Diversamente dalle proposte di imputazione viste in precedenza, in questo caso il *data set* protetto è unico (si tratta di *imputazione singola* e non multipla), e pertanto non è possibile fornire stime non distorte della varianza degli stimatori, a meno di non introdurre un modello inferenziale che formalizzi esplicitamente la relazione tra dati protetti e dati originali tramite la struttura del meccanismo di protezione. Questo impone di rilasciare per lo meno i parametri dei modelli utilizzati per generare i dati protetti, e naturalmente implica maggiori rischi di re-identificazione tramite inferenza.

Infine, in Burridge (2003) e Polettini (2003) vengono proposte due diverse metodologie per il rilascio di insiemi di dati sintetici, entrambe basate sulla modellazione della distribuzione congiunta delle variabili d'indagine. Obiettivo comune ad ambedue le proposte è quello di preservare un insieme di statistiche sufficienti per un particolare modello statistico; i dati rilasciati preservano tali statistiche sufficienti *esattamente* con la prima tecnica, *in media* con la seconda, con differenze che si attenuano al crescere della dimensione del campione rilasciato.

### 8.3.9 Metodo di Dandekar basato sul campionamento da ipercubo latino (*latin hypercube sampling*)

Si tratta di una procedura che può essere definita di tipo *bootstrap* modificato per la protezione di microdati di tipo quantitativo. La protezione è ottenuta rilasciando un campione di dati simulati; in particolare, il modello di protezione utilizza un funzionale della funzione di ripartizione empirica per estrarre unità "sintetiche". Di fatto alla protezione concorrono sia l'operazione di ricampionamento (si noti che non si tratta di un *bootstrap tout court*, perché in tal caso non si conseguirebbe una sufficiente protezione), sia una trasformazione dei dati che tende a ridurre l'apporto informativo degli stessi. La proposta presenta il vantaggio di mantenere inalterati caratteristiche univariate ed indici di correlazione tra ranghi, e quindi conserva le associazioni tra variabili, che sono generalmente di interesse nelle applicazioni.

Più nel dettaglio, la procedura utilizza le distribuzioni marginali empiriche e una matrice di correlazione tra i ranghi $\mathbf{T}$ che rappresenta il parametro da mantenere inalterato nella fase di protezione. Oltre alla matrice di correlazione tra i ranghi, dei dati viene utilizzata soltanto l'informazione relativa ai ranghi marginali riferiti alle singole variabili. La procedura è disegnata per individuare in primo luogo una matrice dei dati che approssimi la matrice di correlazione tra i ranghi prefissata.

Utilizzando la matrice di correlazione tra i ranghi che si vuole mantenere fissata, si modificano i valori assunti dalle distribuzioni marginali empiriche e si determina una matrice di dimensioni ($n, k$) le cui colonne rappresentano i valori assunti dalle funzioni di ripartizione empirica marginali la cui distribuzione congiunta realizza una correlazione tra i ranghi approssimativamente uguale a quella di partenza.

A tali valori vengono infine applicate le rispettive funzioni di ripartizione empirica inversa per ottenere dei valori numerici.

Per un migliore accostamento alla matrice-obiettivo, si ripete l'operazione di creazione di tali dati fintantoché la matrice di correlazione tra i ranghi dei dati fittizi non approssima con sufficiente protezione quella dei dati originali.

Inoltre per rendere possibile l'applicazione di analisi su sottogruppi senza compromettere la validità dei risultati basati sui valori simulati, gli autori propongono di suddividere la popolazione o il campione originale in sottogruppi sui quali applicare la tecnica appena vista. Ciò è chiaramente conseguibile soltanto se i sottogruppi così definiti (eventualmente dopo accorpamenti) hanno numerosità sufficientemente elevata. In alternativa, gli autori suggeriscono di introdurre tra le variabili da considerare anche delle variabili categoriche indicanti gli strati. Questo naturalmente implica che la protezione verrà applicata anche agli strati. Come elemento aggiuntivo di protezione gli autori suggeriscono di generare campioni sintetici di numerosità superiore a quella dei dati osservati. Questo consente di evitare o quantomeno ridurre il problema dei casi unici campionari, ed inoltre dovrebbe consentire, se le distribuzioni marginali sono trattate come visto poc'anzi, di eliminare quasi completamente la presenza di *outlier*. Se l'aumento della numerosità campionaria è notevole, la procedura risente tuttavia del fatto che, per costruzione, il supporto della distribuzione è fissato entro il supporto dei dati osservati, quando invece è ragionevole richiedere che minimo e massimo campionari, al crescere di n, tendano agli estremi del supporto della distribuzione da cui

provengono. Questo crea anche problemi per la stima delle code della distribuzione.

A proposito della scelta della numerosità del campione sintetico, gli autori (Dandekar *et al.,* 2002a) ipotizzano anche di utilizzare una numerosità inferiore a quella del campione osservato in modo da ottenere una microaggregazione.

Una ulteriore possibilità suggerita in congiunzione con il metodo basato sul campionamento da ipercubo latino è il cosiddetto "mascheramento ibrido" (si vedano i lavori Dandekar *et al.*, 2002a e Dandekar, *et al.* 2002b sull'argomento), in cui il *data set* sintetico generato come sopra descritto viene combinato con il *data set* originale, secondo uno schema che gli autori suggeriscono di tipo additivo o moltiplicativo. La procedura può anche essere generalizzata in modo da ottenere un numero di unità in $\tilde{\mathbf{X}}$ diverso da quello di $\mathbf{X}$, semplicemente partendo nella procedura di accoppiamento da un *data set* ricampionato con ripetizione da $\mathbf{X}$.

Le applicazioni pubblicate in letteratura mostrano che generalmente caratteristiche delle distribuzioni marginali quali media e varianza nonché naturalmente la struttura di associazione prefissata rimangono simili ai valori originali. Per quanto riguarda invece la protettività di tale tecnica, in Dandekar *et al.*, (2002a) e Dandekar, *et al.* (2002b) sono documentati esperimenti di *linkage* basato sul criterio della minima distanza euclidea multivariata. Tuttavia, la differente numerosità dei *data set* originale e rilasciato, e l'uso di medie delle percentuali di reidentificazione ottenute in base all'impiego di un numero crescente di variabili chiave fanno sì che i risultati di tali esperimenti risultino di difficile interpretazione. La valutazione della protezione offerta da queste, come altre, tecniche di tutela della riservatezza rimane pertanto un problema tuttora aperto.

### 8.3.10 Modello mistura di Grim, Boček e Pudil

Sebbene il lavoro (Grim *et al.*, 2001) sia stato formalizzato con riferimento specifico alla diffusione di microdati censuari, l'idea di fondo della metodologia può essere applicata anche a microdati d'impresa senza modifiche sostanziali.

Il metodo è basato sulla definizione e la stima di un modello multivariato che rappresenti in modo sufficientemente accurato la distribuzione che ha generato i dati. In particolare, per tale distribuzione viene assunta la struttura di una mistura discreta a componenti *indipendenti.* Si ragiona quindi in termini di strati e all'interno degli strati le distribuzioni sono ipotizzate fattorizzabili nel prodotto delle marginali. Quest'ultima assunzione è certamente la più delicata e, in effetti, la meno verosimile, dell'intero lavoro, in quanto di fatto presuppone l'esistenza di variabili di stratificazione che spieghino la struttura associativa tra le variabili. La struttura generale del modello come mistura discreta di componenti tuttavia è valida, in quanto si può realisticamente rappresentare la distribuzione congiunta come una somma ponderata di sottodistribuzioni condizionate:

$$P(x_1,\cdots,x_k) = \sum_{j=1}^{M} w_j F(x_1,\cdots,x_k \mid j), \qquad \sum_j w_j = 1 \tag{8.5}$$

Come si diceva, l'ulteriore assunzione imposta dagli autori è la seguente:

$$F(x_1,\cdots,x_k\,|\,j)=\prod_{h=1}^{k}p_h(x_h\,|\,j)\,.$$

Il modello (8.5) viene stimato tramite procedure basate sulla verosimiglianza. Trattandosi di un modello mistura, la stima è basata sull'algoritmo EM (Dempster *et al*., 1977; Little e Rubin, 1987); tra gli aspetti delicati dell'algoritmo sono da ricordare la possibilità di raggiungere un massimo locale e la scelta del numero di componenti.

L'opportunità di fornire agli utenti o agli Istituti Statistici un modello formale da cui sia possibile estrarre campioni di individui sintetici è un punto di forza della strategia.

La procedura viene valutata considerando l'accostamento, in termini di massimo errore assoluto, delle distribuzioni *marginali* stimate alle distribuzioni marginali empiriche e delle distribuzioni univariate condizionate a ciascuna variabile del modello; una verifica globale sulla rappresentazione delle relazioni multivariate, tuttavia, manca.

Un modello probabilistico che, sia pure in modo parsimonioso o approssimato, tenga conto delle relazioni di dipendenza tra le variabili può sicuramente fornire il superamento delle limitazioni di cui il modello discusso indubbiamente soffre.

### 8.3.11 Mascheramento Matriciale

Il mascheramento matriciale non è in sé una tecnica di protezione, in quanto rappresenta una formalizzazione unificante che racchiude, come casi particolari, alcune delle tecniche di protezione già trattate.

Come già introdotto in precedenza, denotiamo con $\mathbf{X}$ la matrice osservata e con

$$\tilde{\mathbf{X}}=\mathbf{AXB}+\mathbf{C} \tag{8.6}$$

un *mascheramento matriciale*, ossia una trasformazione dei dati osservati.

La notazione così come riportata sopra è stata introdotta da Cox (1994). Per ottenere una maggiore generalità, la matrice $\mathbf{X}$ nella formula (8.6) può anche essere sostituita da una preventiva trasformazione $\hat{\mathbf{X}}$, risultante dall'applicazione, eventualmente in sequenza, di un cosiddetto *mascheramento elementare* del tipo $\hat{\mathbf{X}}=\mathbf{AX}$, $\hat{\mathbf{X}}=\mathbf{XB}$, o $\hat{\mathbf{X}}=\mathbf{X}+\mathbf{C}$.

Come già notato, la matrice $\mathbf{A}$ opera trasformazioni sugli individui, ragione per cui è detta matrice di trasformazione dei record. Ad esempio, può essere utilizzata per estrarre sottocampioni: in questo caso $\mathbf{A}$ è una matrice stocastica con elementi $a_{ij}\in\{0,1\}$; ogni vettore riga di $\mathbf{A}$ contiene al più un 1 e, se $k$ rappresenta la dimensione del sottocampione da estrarre,

$$\sum_{ij}a_{ij}=k\,.$$

Una opportuna scelta della matrice $\mathbf{A}$ consente anche di ottenere un file di dati *microaggregati* (Defays e Anwar, 1998). L'operazione definita in letteratura come microaggregazione consiste nella selezione, tramite opportuni criteri, di un sottogruppo di $n_i$ unità del campione e nella sostituzione di ciascuno dei valori osservati con somme

o medie degli $n_i$ valori rilevati su tali unità. Tale operazione può essere effettuata su più sottogruppi - naturalmente disgiunti- del campione e chiaramente consente di preservare i totali delle variabili microaggregate, mantenendo, entro certi limiti, incogniti i valori individuali.

Trattandosi della tecnica sicuramente più utilizzata per il rilascio di microdati di impresa, essa viene esaminata in maggior dettaglio nel Paragrafo 8.4, dove, in particolare, viene discusso il criterio di raggruppamento delle unità ai fini dell'aggregazione, aspetto che rappresenta l'elemento cruciale di tale tecnica.

Una volta scelto il criterio di selezione delle unità da aggregare ed ordinata la matrice dei dati secondo tale criterio, la microaggregazione in $k$ gruppi di numerosità $n_1, n_{2,\ldots,} n_k$ delle prime $m$ unità (dove $m = \sum_i n_i$) consiste nella sostituzione delle medie ai valori individuali. Tale operazione può essere conseguita tramite un particolare mascheramento matriciale $\tilde{\mathbf{X}} = \mathbf{AX}$ in cui $\mathbf{A}$ è definita come una matrice a blocchi di dimensione $n \times n$ i cui elementi sono le matrici quadrate $\mathbf{J}_{n_i}$, ciascuna di elemento generico $\frac{1}{n_i}$, la matrice identità $\mathbf{I}_{n-m}$ e le matrici di nullità, di dimensione $n_i \times n - n_i$ o $n - m \times m$ (a seconda che esse affianchino le matrici $\mathbf{J}_{n_i}$ identicamente uguali a $\frac{1}{n_i}$ o la matrice identità).

$$\mathbf{A} = \begin{pmatrix} \mathbf{J}_{n_1} & \mathbf{0}_{n_1 \times n - n_1} \\ \mathbf{J}_{n_2} & \mathbf{0}_{n_2 \times n - n_2} \\ \vdots & \vdots \\ \mathbf{J}_{n_k} & \mathbf{0}_{n_k \times n - n_k} \\ \mathbf{0}_{n-m \times m} & \mathbf{I}_{n-m} \end{pmatrix}, \quad \mathbf{J}_{n_i} = \begin{pmatrix} n_i^{-1} & \cdots & n_i^{-1} \\ \vdots & n_i^{-1} & \vdots \\ n_i^{-1} & \cdots & n_i^{-1} \end{pmatrix}$$

$\mathbf{B}$, la matrice di trasformazione delle variabili, è invece utilizzata per trasformare i valori delle modalità.

La soppressione locale, ossia la rimozione di prespecificate modalità di unità considerate ad elevato rischio di violazione, è ottenibile tramite l'applicazione successiva di mascheramenti elementari, definendo $\mathbf{A}$ in modo che identifichi i sottogruppi di unità in cui va rimossa la stessa variabile e $\mathbf{B}$ come una matrice a blocchi di elementi matrici identità e matrici di nullità.

La matrice $\mathbf{C}$, infine, detta matrice di spostamento, consente, nella forma base $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{C}$, di aggiungere disturbo casuale ad alcune variabili, di effettuare il cosiddetto *rounding* o raggruppamento di variabili a valori interi come l'età.

## 8.4 La microaggregazione

Dal momento che le imprese sono considerate ad alto rischio di identificazione, per il rilascio di dati elementari di impresa non si effettua una stima del rischio di re-identificazione come visto in precedenza, piuttosto, le tecniche proposte in letteratura

per la protezione dei microdati di impresa puntano alla produzione di un file di unità statistiche fittizie i cui record non siano più riconducibili ai dati originali, mantenendo al contempo le caratteristiche dei dati originali dal punto di vista dell'analisi statistica.

Un approccio adottato in alcuni casi è quello suggerito da Defays e Nanopoulos (1992) e Anwar (1993); esso consiste nel produrre dati *microaggregati* per le variabili quantitative. In pratica si producono *pseudo-record* sintetizzando (normalmente utilizzando la media aritmetica) i valori di un gruppo di unità statistiche che presentano caratteristiche simili tra di loro. Il principio si basa sulla regola della soglia utilizzata per rilasciare dati aggregati (il numero minimo di unità statistiche in ciascun gruppo è almeno pari a tre). La sintesi può essere effettuata su tutte le variabili contemporaneamente, di fatto creando unità statistiche identiche tra loro, oppure una variabile per volta. Nel primo caso l'efficacia di queste tecniche dal punto di vista della tutela della riservatezza è notevole anche perché il controllo del rischio di violazione è preliminare alla produzione del file da rilasciare. L'utilità dei microdati prodotti dal punto di vista dell'analisi statistica sconta, invece, il fatto che non propriamente di microdati si tratta ma, appunto, di microaggregati (Corsini *et al*., 1999; Contini *et al*., 1998). Nel secondo caso, analisi empiriche mostrano che il livello di protezione non è, almeno nel caso trattato, sufficiente (si veda Pagliuca e Seri, 1999b).

Nonostante le riserve metodologiche sui dati microaggregati, le tecniche di microaggregazione continuano a destare l'interesse degli Istituti nazionali di statistica, compreso l'Istat che ha prodotto in via sperimentale dei dati microaggregati per l'indagine sui conti economici delle imprese relativa agli anni 1994 e 1995 (vedi Paragrafo 9.2).

Con il termine microaggregazione ci si riferisce ad un insieme di tecniche (Defays e Anwar, 1998) sviluppate con l'intendimento di rendere possibile il rilascio di dati d'impresa sotto il vincolo del segreto statistico e della tutela della riservatezza. L'idea su cui si basano tali tecniche è quella di creare delle unità fittizie che assumono dei valori ottenuti come sintesi di quelli osservati su un insieme di numerosità minima di imprese (di solito costante e indicata con $k$). Il criterio di sintesi utilizzato è generalmente la media per le variabili quantitative. Questo rende evidente il fatto che tanto più le imprese nel microgruppo presentano valori simili fra loro tanto più piccola è la distorsione introdotta nei dati (perdita di informazione).

La questione può quindi essere vista come un problema di analisi dei gruppi con il vincolo che il numero di imprese in un gruppo sia non inferiore a una soglia fissata. Con tale vincolo si intende garantire *a priori* il rispetto delle norme sul segreto statistico e la tutela della riservatezza dei dati individuali attualmente in uso per il rilascio e la diffusione dei dati poiché ogni singola unità rilasciata si riferisce a un numero minimo di unità (regola della soglia). Inoltre, può essere considerata anche la regola della dominanza per cui un'unità statistica del file originale assegnata a un microgruppo non deve contribuire con i propri valori in misura superiore a una percentuale $p$ fissata sulla somma dei valori nel gruppo.

Un modo semplice di applicare la microaggregazione è quello di ordinare le unità statistiche (in senso crescente o decrescente) secondo i valori di una variabile (metodi su asse singolo o microaggregazione univariata). Questa rappresenta il criterio di similarità con cui viene misurata la distanza fra le unità statistiche e può essere scelta tra quelle

presenti nel file o determinata secondo altri criteri (ad esempio viene suggerito di scegliere la prima componente principale per un insieme di variabili). Successivamente, a partire dalle prime *k* nell'ordinamento, si creano gruppi di *k* unità consecutive in modo da formare una partizione dell'insieme di partenza. I valori medi ponderati con i coefficienti di riporto all'universo di ciascuna variabile sono i valori attribuiti alle unità (imprese fittizie) presenti nel file. L'esempio rappresentato in Tabella 8.1 e in Tabella 8.2 utilizza valori fittizi e mostra come agisce il processo di microaggregazione su asse singolo.

**Tabella 8.1 Valori dell'asse di ordinamento calcolato come combinazione dei valori standardizzati delle variabili Fatturato e Addetti, posizione delle unità statistiche rispetto all'asse di ordinamento (Ordine) e valori osservati**

| | | | Valori osservati | | |
|---|---|---|---|---|---|
| Asse | Valore | Ordine | Fatturato | Addetti | Export |
| 1 | 0.07 | 5 | 100000 | 70 | 17200 |
| 2 | -0.09 | 4 | 64000 | 90 | 10300 |
| 3 | 0.90 | 6 | 166000 | 50 | 2500 |
| 4 | 1.45 | 10 | 190000 | 50 | 18700 |
| 5 | 1.10 | 9 | 160000 | 60 | 11300 |
| 6 | -1.25 | 2 | 130000 | 10 | 22400 |
| 7 | -3.28 | 1 | 41000 | 10 | 29000 |
| 8 | 1.07 | 8 | 100000 | 100 | 22000 |
| 9 | 0.96 | 7 | 110000 | 90 | 20000 |
| 10 | -0.94 | 3 | 99000 | 40 | 14600 |
| Totale | | | 1160000 | 570 | 168000 |

**Tabella 8.2 Valori (crescenti) dell'asse di ordinamento, posizione delle unità statistiche rispetto all'asse di ordinamento (Ord), identificazione del gruppo cui viene attribuita l'unità statistica (Gr), valori osservati e valori microaggregati**

| | | | | Valori osservati | | | Dati microaggregati | | |
|---|---|---|---|---|---|---|---|---|---|
| Asse | Valore | Ord | Gr | Fatturato | Addetti | Export | Fatturato | Addetti | Export |
| 7 | -3.28 | 1 | 1 | 41000 | 10 | 29000 | 90000 | 20 | 22000 |
| 6 | -1.25 | 2 | 1 | 130000 | 10 | 22400 | 90000 | 20 | 22000 |
| 10 | -0.94 | 3 | 1 | 99000 | 40 | 14600 | 90000 | 20 | 22000 |
| 2 | -0.09 | 4 | 2 | 64000 | 90 | 10300 | 110000 | 70 | 10000 |
| 1 | 0.07 | 5 | 2 | 100000 | 70 | 17200 | 110000 | 70 | 10000 |
| 3 | 0.90 | 6 | 2 | 166000 | 50 | 2500 | 110000 | 70 | 10000 |
| 9 | 0.96 | 7 | 3 | 110000 | 90 | 20000 | 140000 | 75 | 18000 |
| 8 | 1.07 | 8 | 3 | 100000 | 100 | 22000 | 140000 | 75 | 18000 |
| 5 | 1.10 | 9 | 3 | 160000 | 60 | 11300 | 140000 | 75 | 18000 |
| 4 | 1.45 | 10 | 3 | 190000 | 50 | 18700 | 140000 | 75 | 18000 |
| Totale | | | | 1160000 | 570 | 168000 | 1160000 | 570 | 168000 |

Dall'esempio si può notare anche che, se il numero di unità statistiche sottoposto a microaggregazione non è un multiplo di *k*, le imprese rimanenti vengono attribuite all'ultimo gruppo (o al primo secondo convenienza).

La perdita di informazione dovuta alla "perturbazione" delle informazioni originali rispetto ai dati microaggregati viene, in genere, misurata come perdita di variabilità.

Data la matrice dei dati originali $\mathbf{X}(n,p)$ relativa ad n unità statistiche su cui sono rilevate p variabili quantitative, il problema della microaggregazione consiste nel determinare una partizione dell'insieme delle n unità statistiche in classi di numerosità non inferiore a una soglia fissata k (k-partizione) e che minimizzi la perdita di

informazione.

Misuriamo l'informazione contenuta nel file con una misura di variabilità come la somma degli scarti al quadrato dalla media. Assumendo per semplicità che $p$=1 e applicando il principio di scomposizione della devianza, abbiamo:

$$\begin{aligned}
\text{Dev(T)} &= \sum_{i(1,n)}(x_i - x_M)^2 \\
&= \sum_{r(1,m)}\sum_{i(1,n_r)}(x_{ir} - x_M)^2 \\
&= \sum_{r(1,m)}\sum_{i(1,n_r)}(x_{ir} - x_{M_r})^2 + \sum_{r(1,m)} n_r(x_{M_r} - x_M)^2 \\
&= \text{Dev(W)} + \text{Dev(B)} \\
&= \text{Dev(W)} + \text{Dev}(Y)
\end{aligned}$$

dove:

Dev(T)= devianza totale della variabile $X$
i= indice contatore delle unità statistiche
$x_M$= media di $X$
r= indice contatore dei gruppi
m= numero dei gruppi che formano la partizione
$n_r$ = numero delle unità statistiche nell'r-mo gruppo
$x_{M_r}$ = media di $X$ nell'r-mo gruppo
Dev(W)= devianza nei gruppi
Dev(B)= devianza tra i gruppi
Dev($Y$)= devianza della variabile $Y$ che sostituisce $X$ nel file rilasciato

Quindi, Dev(W) esprime la perdita di informazione dovuta alla microaggregazione in quanto differenza fra la variabilità dei dati originali e quella dei dati microaggregati. Pertanto, la soluzione ottima per il problema della microaggregazione è data dalla k-partizione che minimizza Dev(W). Normalmente si considera come misura relativa da minimizzare il rapporto Dev(W)/Dev(T) o, equivalentemente, il rapporto ad esso complementare a 1, Dev($Y$)/Dev(T) da massimizzare. Quest'ultimo esprime la qualità di rappresentazione dei dati originali da parte dei dati microaggregati.

Ai fini della rappresentazione della qualità dei dati questa misura può essere espressa analiticamente per ciascuna delle p variabili oppure sintetizzata in un indice unico. In tal caso si possono introdurre nel calcolo dei pesi che esprimono l'importanza che si intende attribuire a ciascuna variabile. Tanto più sono omogenee le unità statistiche nei gruppi e tanto più piccola è Dev(W).

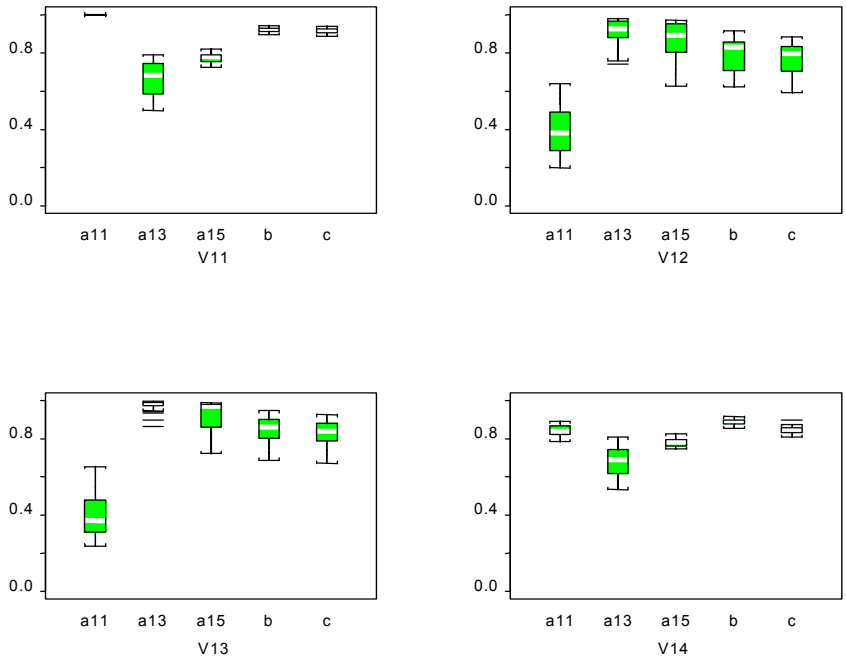
La soluzione ottima può essere ottenuta esaminando in maniera esaustiva tutte le possibili $k$-partizioni sulle n unità statistiche (Defays e Nanopoulos, 1992), ma il costo computazionale di questa operazione cresce in modo esponenziale rispetto ad $n$ (per $n$=100 e $k$=3 Domingo-Ferrer e Mateo-Sanz, 2001 calcolano circa 23 giorni di tempo di elaborazione) e l'algoritmo di ricerca è di difficile implementazione.

Non presentano problemi di costo computazionale gli algoritmi proposti inizialmente nella pratica del tipo accennato nel precedente paragrafo. Il meccanismo di partizione in questi algoritmi consiste nell'ordinare le unità statistiche in base a qualche criterio "ragionevole" e costruire i gruppi prendendo $k$ (costante) unità consecutive nell'ordinamento fino ad esaurimento dell'insieme. Nel caso in cui n non è multiplo di $k$

le unità rimanenti vengono attribuite al gruppo più vicino nell'ordinamento. I criteri proposti sono di tipo unidimensionale, ossia viene utilizzata una variabile come criterio per ordinare le unità (*single axis methods*). Tale variabile rappresenta in sostanza il criterio con cui si misura la dissimilarità tra le unità statistiche. Questi metodi sono tanto più efficaci quanto più le *p* variabili sono correlate tra loro (e con l'asse di ordinamento). Come già anticipato, la variabile-asse di ordinamento può essere scelta tra quelle presenti nel file, ma una scelta ragionevole sulla base della precedente considerazione è quella di utilizzare come asse di ordinamento la prima componente principale (*first principal component method*) per la sua qualità di massimizzare la somma delle correlazioni con le variabili originali. Un criterio alternativo è quello di ordinare le unità secondo la somma degli scarti standardizzati (*sum of Zed score method*). Secondo questo criterio le variabili vengono standardizzate e per ogni unità statistica viene calcolata la somma dei valori così ottenuti su tutte le variabili.

Nella sperimentazione condotta su dati reali (Corsini *et al*., 1999; Contini *et al*., 1998) descritta nel Paragrafo 9.2, questi due ultimi criteri hanno mostrato di essere sostanzialmente equivalenti, mentre rispetto alla scelta di una singola variabile i risultati sono stati mediamente migliori, nel senso che la scelta di una variabile fornisce buoni risultati su se stessa e sulle variabili strettamente correlate con essa ma peggiori rispetto alle variabili meno correlate. Questa considerazione è ben evidenziata in Figura 8.1. dove i criteri sono indicati con le lettere b (prima componente principale), c (somma degli scarti standardizzati), a11, a13 e a15 rispettivamente i criteri basati sulla variabile v11, v13 e v15.

**Figura 8.1 Boxplot del rapporto di varianze relativamente alle variabili da V11 a V14 per metodo di microaggregazione (indagine Sistema dei conti delle imprese con 1-19 addetti, anni di rilevazione '93 e '94, parametro di aggregazione k=3, 4 e 5, sei divisioni di attività economica)**



Nel primo grafico si nota che la qualità di rappresentazione (espressa con il rapporto fra le varianze) rispetto al criterio a11 è elevata rispetto alla variabile v11 (se stessa) al variare di *k*, divisione di attività economica e su entrambi gli anni di rilevazione presi in considerazione, mentre nel terzo grafico si nota che lo stesso criterio fornisce risultati inferiori e più variabili rispetto alla variabile v13. I criteri b e c si comportano, invece, in maniera più stabile rispetto alle variabili e ai parametri presi in considerazione.

Considerati i buoni risultati su una variabile, in termini di perdita di variabilità, quando l'ordinamento è definito dalla variabile stessa, è stato suggerito un criterio di microaggregazione per cui le variabili sono trattate indipendentemente l'una dall'altra (*individual ranking method*). Tuttavia, nonostante il metodo mantenga normalmente più informazione rispetto ai metodi su asse singolo, non garantisce allo stesso modo la riservatezza dei dati e per questo non viene qui trattato (Pagliuca e Seri, 1999b).

Uno sviluppo notevole delle tecniche di microaggregazione univariata è stato introdotto da Domingo-Ferrer e Mateo-Sanz (2001). Nel lavoro sono dimostrate alcune proposizioni importanti nella ricerca di una soluzione ottimale (e subottimale) per il problema della microaggregazione. Riportiamo di seguito i più significativi.

*Definizione 1* (*Connected group*) dato un insieme S (ordinabile) di dati, un gruppo $G \subseteq S$ si dice connesso se per ogni $x,y,z \in S$ tali che $x \leq z \leq y$, con $x,y \in G$ si ha che $z \in G$.

*Lemma 1* Dato un insieme di n dati, con $n \geq 2k$, allora la partizione ottima (secondo la definizione data sopra) in due gruppi entrambi di numerosità non inferiore a *k* è costituita da due gruppi entrambi connessi.

*Teorema 1* La soluzione ottima del problema detto della *k*-partizione di un insieme di dati è costituita da gruppi connessi.

*Corollario 1* Una soluzione ottima per il problema della *k*-partizione di un insieme di dati esiste ed è tale che i suoi gruppi sono connessi ed hanno numerosità compresa tra *k* e 2*k*.

Grazie a questi risultati l'insieme delle possibili partizioni tra le quali ricercare la soluzione ottima si riduce notevolmente perché è costituito dalle sole partizioni costituite da gruppi connessi di numerosità compresa tra *k* e 2*k*-1. Su queste basi sono stati definiti i metodi euristici presentati nello stesso lavoro e che descriviamo brevemente di seguito.

Il primo metodo è derivato dal metodo di Ward per il *clustering* gerarchico (Ward, 1963) e per questo denominato *k*-Ward. Facendo riferimento al caso unidimensionale, si tratta di ordinare i dati e formare un gruppo con le prime k unità nell'ordinamento e un gruppo con le ultime *k*, mentre le rimanenti unità statistiche formano ciascuna dei gruppi unidimensionali. Successivamente si applica il metodo di Ward fino a che tutti gli elementi del file apparterranno a un gruppo con almeno *k* elementi evitando di unire gruppi che abbiano entrambi numerosità superiore o uguale a *k*. Se al termine del processo la partizione risultante è costituita da gruppi di numerosità superiore o uguale a 2*k*, l'algoritmo si ripete su ogni singolo gruppo ricorsivamente fino ad ottenere una soluzione ammissibile secondo il Corollario 1.

La complessità computazionale del metodo *k*-Ward è $O(n^2)$ e questo può essere un

problema per file di grandi dimensioni.

La seconda proposta è basata sull'uso degli algoritmi genetici (si veda ad esempio Ribeiro *et al*, 1994). L'adattamento degli algoritmi genetici al problema della *k*-partizione ottima consiste in:

- rappresentare una k-partizione con una stringa binaria di n elementi in cui l'elemento i-mo rappresenta la i-ma unità statistica nel file ordinato. Il primo elemento di un gruppo viene rappresentato da un 1 mentre gli altri vengono rappresentati con uno 0. In tal modo le soluzioni ammissibili sono le stringhe in cui gli 1 sono separati da almeno k-1 e non più di 2k-2 0;
- gli operatori "*crossover*" e "*mutation*"[15] sono modificati in modo che ogni stringa generata rappresenti a sua volta una k-partizione.

Un limite di questo approccio risiede nell'arbitrarietà nella scelta dei parametri dell'algoritmo, in particolare, le probabilità con cui intervengono gli operatori e il criterio di convergenza. Inoltre, essendo un metodo basato su un criterio di ricerca casuale, fornisce soluzioni che non sono riproducibili nel senso che applicandolo più di una volta sullo stesso file si potrebbero ottenere risultati ogni volta diversi. La complessità computazionale è lineare, O(*n*).

Per quanto riguarda la perdita di informazione, nel caso in cui non sia possibile condurre una ricerca esaustiva della soluzione ottima, il metodo *k*-Ward è quello che sembra preferibile anche se al crescere di *n* e di *k* i risultati tendono a essere simili per tutti i metodi. Quando la necessità di un tempo di elaborazione ridotto prevale (ad esempio, se si vuole consentire il calcolo dei microaggregati via internet), i metodi basati sugli algoritmi genetici o su singolo asse sono da preferire. I primi per le migliori performance in termini di perdita di informazione, i secondi per la semplicità di implementazione.

Il metodo *k*-Ward presenta una ulteriore peculiarità per essere più facilmente generalizzabile al caso multidimensionale in cui un'unità statistica è rappresentata con un vettore di valori (microaggregazione multivariata) e i confronti fra unità avvengono sulla base di una distanza che fornisce un ordinamento solo parziale all'insieme dei dati. Per tale motivo il concetto di "gruppo connesso" perde di significato e i risultati teorici precedenti non sono più applicabili.

Un criterio di ordinamento per la microaggregazione multidimensionale suggerito da Domingo-Ferrer e Mateo-Sanz (2001) basato sulla definizione di una distanza fra imprese (unità statistiche) consiste nell'individuare, nella matrice delle distanze, le unità più distanti fra di loro (unità *estremi*). Successivamente si associano le *k*-1 unità più vicine a ciascun estremo formando due gruppi, delle *prime k* unità e delle *ultime k* unità. Escludendo le unità nei due gruppi si replica il procedimento sulle rimanenti unità fino ad averle assegnate tutte ad un gruppo.

Algoritmi per la microaggregazione di variabili quantitative sono implementati all'interno del software μ-Argus,[16] sviluppato all'interno del progetto europeo Casc.

[15] Date due stringhe l'operatore crossover le combina con probabilità Pcross o le lascia inalterate con probabilità 1-Pcross.
Una mutazione si applica alterando ciascun elemento di una stringa da 1 a 0 o viceversa con probabilità Pmut. Questa probabilità è scelta generalmente bassa (ad esempio, 0,05).

[16] http://neon.vb.cbs.nl/casc/MU.html

# PARTE QUARTA

## Alcune esperienze in Istat

# Capitolo 9. Comunicazione di dati a soggetti non Sistan(*)

## 9.1 Il rilascio dei *file standard* in Istat

Il Decreto legislativo 322/1989 costitutivo del Sistan prevede la possibilità per l'Istat di rilasciare "… 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." (art.10) che nell'accezione Istat hanno preso il nome di *file standard*. In questo paragrafo si descrivono brevemente le procedure applicate dall'Istat per la produzione dei *file standard*, ossia collezioni di dati elementari che vengono rilasciate, per motivi di ricerca e dietro la sottoscrizione di un accordo contratto, ad utenti non appartenenti ad uffici del Sistema statistico nazionale.

In ottemperanza alla norma sopra richiamata, l'Istat rilascia *file standard* da circa dieci anni per le seguenti indagini:

1) Censimento popolazione e abitazioni (da un campione all'1 per cento per il 1991);
2) Indagine trimestrale sulle forze di lavoro (file trimestrali e recentemente file longitudinali);
3) Indagine sui Consumi delle famiglie;
4) Indagine multiscopo sulle famiglie;
5) Panel europeo delle famiglie (Echp, *European Community Household Panel*);
6) Inserimento professionale dei laureati;
7) Percorsi di studio e lavoro dei diplomati.

A seguito dell'entrata in vigore del *Commission Regulation* 831/2002, sono in corso di predisposizione da parte di Eurostat *anonymised microdata file* (equivalenti ai nostri *file standard*) relativi alle indagini *Continuing Vocational Training Survey* (Cvts) sulla formazione professionale in ambito lavorativo e *Community Innovation Survey* (Cis) sull'innovazione tecnologica delle imprese.

Inizialmente, nella fase di approntamento della procedura di rilascio dei *file standard*, sono stati condotti studi su alcune indagini sulle famiglie volti ad analizzare dal punto di vista statistico la tutela della riservatezza nel rilascio di *file standard*. Per misurare il rischio di violazione di un file è stato in tale fase adottato un criterio di tipo globale (vedi Paragrafo 6.2), relativo cioè all'intero file e quantificato come numero atteso di re-identificazioni nel file da rilasciare; i risultati degli studi appena richiamati hanno permesso di stabilire quali provvedimenti (metodi di protezione) adottare per contenere tale rischio al di sotto di una soglia accettabile. Poiché la misura del rischio utilizzata in questa fase era di tipo *globale*, anche le misure di protezione non potevano essere limitate a record specifici e di conseguenza gli interventi di protezione venivano applicati alla totalità degli individui; questi consistevano sostanzialmente nella

(*) Capitolo redatto da Giovanni Seri eccetto il paragrafo 9.1 redatto da Luisa Franconi

ricodifica delle variabili con conseguente riduzione del loro dettaglio. La procedura appena descritta è stata in un primo tempo adottata per la creazione di tutti i *file standard* prodotti dall'Istituto (ad eccezione del file per l'indagine Echp costruito con criteri armonizzati a livello europeo).

Dato che il meccanismo di protezione descritto agisce su tutti i record indistintamente, questa procedura di rilascio ha chiaramente un notevole impatto sul contenuto informativo del file rilasciato, contenuto che viene considerevolmente ridotto. Per superare tale limitazione, è stata recentemente introdotta una misura di rischio individuale (Benedetti e Franconi, 1998) definita per ogni singolo record (vedi Paragrafo 6.3). Tale misura consente di applicare la protezione ai soli record ad elevato rischio di re-identificazione, e allo stesso tempo contenendo entro valori tollerabili anche misure di rischio globale quali il numero atteso di re-identificazioni nel file. E' quindi da qualche anno che, in corrispondenza di un rinnovamento delle indagini, o per indagini per le quali non sono stati precedentemente prodotti *file standard*, viene applicata la nuova procedura di rilascio. Ciò è avvenuto finora per l'Indagine sui consumi delle famiglie e per l'Indagine sull'inserimento professionale dei laureati.

La nuova procedura è basata sull'adozione della misura di rischio individuale già descritta nel Paragrafo 6.3; una volta definito un valore di soglia, i record il cui rischio di re-identificazione supera tale soglia vengono definiti *a rischio* e successivamente protetti tramite il metodo della soppressione locale descritto nel Paragrafo 8.3.1. La protezione è pertanto selettiva, ossia applicata a singole variabili e singoli record, e ciò ha permesso di rilasciare *file standard* con una più alta valenza informativa rispetto alla procedura precedentemente adottata. Alcune variabili, soprattutto quelle geografiche, vengono tuttora sottoposte a ricodifica globale nel caso in cui il numero di record a rischio sia troppo elevato e conseguentemente il numero di soppressioni eccessivo.

Per dare una misura dell'azione di questa nuova procedura rispetto alla precedente presentiamo alcuni dati. Così, ad esempio, per l'indagine sui Consumi il file gerarchico (dal quale è possibile ricostruire la struttura familiare) prevedeva un dettaglio geografico per il "luogo di residenza" a livello di ripartizione geografica (anni precedenti al 1997). Con la procedura attuale (dati del file relativo all'anno di indagine 1997) è stato possibile migliorare il dettaglio a livello regionale lasciando inalterati circa l'87 per cento dei record (su circa 64 mila in totale). Inoltre, circa l'80 per cento dei record protetti hanno subito una sola soppressione (i rimanenti non più di due). Analoga percentuale per le informazioni a livello familiare (circa 22 mila famiglie), quasi l'88 per cento non subiscono alcuna soppressione. Tuttavia, in questo caso le famiglie che subiscono più di una soppressione è più alto. Ciò è dovuto al fatto che alcune variabili, come la regione di residenza, se vengono soppresse per un record devono essere soppresse anche per tutti i componenti della stessa famiglia. Il confronto fra i due file (qui presentato in maniera semplificata) rende comunque evidente il pregio di questo nuovo approccio. L'intervento di protezione, e quindi di riduzione dell'informazione, è limitato a una porzione ridotta del file (circa il dieci per cento) invece che all'intero file. Di contro, il prezzo che viene pagato per un maggior dettaglio nell'informazione geografica è nel non avere alcune informazioni (come se si avesse una mancata risposta in almeno una variabile di ciascun record protetto). Non è immediatamente confrontabile il rischio di identificazione nelle due procedure, comunque, si tenga conto

che nel caso del rischio individuale la soglia di rischio è stata fissata in $\alpha=2.5\times10^{-5}$, il che vuol dire che il rischio di identificazione per ciascun record non è superiore a 1/40000.

E' importante sottolineare ancora una volta il fatto che l'applicazione di qualunque tecnica di protezione riduce la quantità e il dettaglio delle informazioni rilasciate determinando una diminuzione del contenuto informativo del file dei microdati. Introducendo, oltre a misure di rischio di violazione, anche misure di perdita di informazione, è possibile formalizzare il problema del rilascio come un problema di ottimizzazione; l'uso di opportuni algoritmi di ricerca operativa consente, quindi, di tener conto di ambedue i criteri appena menzionati. A questo proposito, de Waal e Willenborg (1998) hanno proposto un algoritmo per la soppressione locale ottimale (che minimizza, cioè, la perdita di informazione associata all'applicazione del metodo). Hurkens e Tiourine (1998), hanno recentemente suggerito di modificare i microdati tramite la ricodifica globale e/o la soppressione locale in modo da minimizzare una misura di perdita di informazione. La procedura adottata dall'Istat accoglie, da una parte, la proposta di Hurkens e Tiourine, limitandola ai record a rischio, e dall'altra la integra con l'algoritmo di ottimizzazione di de Waal e Willenborg; in particolare, la ricodifica globale viene applicata per limitare il numero di record a rischio a un livello ragionevole mentre l'algoritmo di ottimizzazione viene utilizzato per stabilire su quali variabili applicare le soppressioni locali. Tale tecnica di protezione e la misura di rischio individuale adottate dall'Istat sono implementate all'interno del software μ-Argus, sviluppato nell'ambito del progetto europeo Casc (*Computational Aspects of Statistical Confidentiality*). Il pacchetto, attualmente in fase di verifica, è interamente dedicato alla protezione dei dati elementari e contiene le principali tecniche proposte in letteratura. Il software è disponibile a titolo gratuito presso il sito del progetto Casc.[17]

## 9.2 La microaggregazione dei dati microaggregati del sistema dei conti economici delle imprese italiane. Anni 1995 e 1996

La crescente richiesta di microdati di impresa da parte di ricercatori e studiosi ha spinto l'Istituto nazionale di statistica italiano a sperimentare, per la prima volta, il rilascio di dati d'impresa microaggregati (si veda a proposito della microaggregazione il Paragrafo 8.4), compiendo in tal modo un passo in avanti rispetto alla tradizionale diffusione dei dati in forma tabellare. Si tratta, in particolare, della diffusione dei dati relativi ai costi, ai ricavi, all'occupazione, al costo del lavoro e agli investimenti delle imprese rispondenti alle rilevazioni sui conti economici delle imprese industriali e dei servizi nel 1995 e nel 1996 (Istat, 2001), aggregati secondo le più piccole numerosità consentite dall'obbligo di tutela del segreto statistico e della riservatezza dei dati riferiti alle unità rispondenti alle rilevazioni.

Per una corretta interpretazione dei dati occorre precisare che per le imprese plurisettoriali l'attribuzione dell'attività economica è avvenuta secondo il criterio della prevalenza. Nel caso di imprese plurilocalizzate la ripartizione geografica è stata stabilita in relazione alla ripartizione di appartenenza della sede legale. I dati sono prodotti da due

[17] http://neon.vb.cbs.nl/casc/MU.html

indagini distinte: per le imprese con addetti da 1 a 19 un'indagine campionaria, per le imprese con 20 addetti e oltre un'indagine totale sottoposta a integrazione da altre fonti per le mancate risposte totali. Le variabili considerate rappresentano gli aggregati comuni alle due indagini (Istat, 1998).

Il file di dati è stato preventivamente trattato prima di essere sottoposto a microaggregazione. Le imprese sono state analizzate per tipo di attività economica (codici Ateco 1991 a tre cifre) e dettaglio territoriale (ripartizioni geografiche) al fine di ottenere dei domini omogenei con numerosità congrue. Alcuni gruppi di attività economica, in cui le numerosità erano insufficienti e l'aggregazione con altri gruppi giudicata distorsiva o poco significativa, sono stati eliminati. Ogni dominio di imprese definito dalla combinazione di attività economica e dettaglio territoriale è stato trattato indipendentemente dagli altri.

In ciascun dominio sono state individuate le imprese più influenti sulla variabilità di alcune variabili principali. Per migliorare il risultato del successivo processo di microaggregazione, tali imprese sono state accorpate in un gruppo, insieme ad altre imprese scelte in maniera casuale. La qualità dell'informazione fornita dall'impresa fittizia rappresentativa di tale gruppo non è apprezzabile in termini di approssimazione dei dati originali in quanto viene ottenuta non tenendo conto del criterio di similarità implicito nel metodo.[18]

Il resto delle unità è stato sottoposto al processo di microaggregazione utilizzando una tecnica su asse singolo. L'asse di ordinamento è stato calcolato come combinazione lineare delle seguenti variabili centrate e ridotte: numero di addetti, fatturato totale, ricavi accessori, costi per materie prime, costi per servizi vari, costo del lavoro, investimenti e valore aggiunto. Per queste variabili, che contribuiscono direttamente alla formazione dell'asse di ordinamento, la rappresentazione attraverso i dati microaggregati è risultata chiaramente migliore rispetto alle restanti. La sintesi all'interno dei gruppi è stata operata mediante media aritmetica ponderata con i coefficienti di riporto all'universo (vedi Tabella 9.1).

Per valutare i risultati solitamente si misura la perdita di informazione con la perdita di variabilità che si registra dopo la microaggregazione. In maniera equivalente, misurando la qualità dell'informazione invece della perdita, è stato calcolato il rapporto tra la varianza di una variabile modificata dalla microaggregazione e la varianza originale della stessa. Tanto più questo rapporto è vicino all'unità, tanto migliore è il risultato ottenuto in termini di perdita di informazione. L'indice è stato calcolato analiticamente per ogni variabile e per ogni dominio (vedi Tabella 9.2).

Mediamente il valore di questo indice di "variabilità mantenuta" si attesta intorno a 0.75, tuttavia i risultati sono molto variabili da dominio a dominio e da variabile a variabile. Ad esempio per la variabile "numero di addetti" l'indice registra nei vari domini valori che vanno da un minimo di 0.36 a un massimo di 0.97; per il fatturato si va da 0.34 a 0.99. Valori analoghi si hanno per le variabili che contribuiscono al calcolo dell'asse di ordinamento. Per le altre, come era lecito aspettarsi anche per la loro natura, la perdita di informazione è maggiore con una percentuale più elevata di valori bassi dell'indice.

[18] Tale unità, complementare rispetto al resto dei dati, garantisce la coerenza dei totali delle variabili nei domini. Nel file rilasciato queste unità sono contrassegnate da un flag.

**Tabella 9.1 Esempio di dati microaggregati estratti dal file rilasciato**

| Ateco | Anno | Rip. Geo. | Addetti | Dipend | Fatt. Tot. | Costi-1 | Costi-2 | Cos. Lav. | .... | Tot. Inv. | Val. Agg. | Coeff. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 141 | 95 | 1 | 10.6 | 8.2 | 950.9 | 202.3 | 334.4 | 202.2 | .... | 147.0 | 422.6 | 20.1 |
| 141 | 95 | 1 | 14.5 | 12.5 | 1060.8 | 303.2 | 546.9 | 386.5 | .... | 60.9 | 198.6 | 26.8 |
| 141 | 95 | 1 | 25.0 | 23.3 | 5122.0 | 1931.0 | 1571.7 | 1163.7 | .... | 534.0 | 2255.7 | 3.0 |
| 141 | 95 | 1 | 28.0 | 26.7 | 5275.3 | 1281.0 | 1467.7 | 1522.3 | .... | 153.7 | 2688.3 | 3.0 |
| 141 | 95 | 1 | 28.7 | 27.7 | 4820.7 | 1709.3 | 1153.3 | 1461.0 | .... | 94.7 | 2168.7 | 3.0 |
| 141 | 95 | 1 | 43.3 | 41.3 | 9007.7 | 1938.7 | 2124.3 | 2497.0 | .... | 266.0 | 4801.7 | 3.0 |
| 141 | 95 | 1 | 46.3 | 44.3 | 10190.7 | 2977.0 | 2372.3 | 2411.3 | .... | 2374.0 | 4621.0 | 3.0 |
| 141 | 95 | 1 | 48.0 | 47.0 | 9209.3 | 2784.7 | 1987.7 | 2587.3 | .... | 666.3 | 4131.7 | 3.0 |
| 141 | 95 | 2 | 2.8 | 1.2 | 606.4 | 180.4 | 173.4 | 52.7 | .... | 19.3 | 262.8 | 93.7 |
| 141 | 95 | 2 | 2.9 | 1.1 | 163.6 | 33.7 | 53.9 | 43.7 | .... | 27.4 | 91.9 | 31.1 |
| 141 | 95 | 2 | 4.8 | 3.8 | 728.1 | 170.6 | 245.7 | 150.6 | .... | 248.8 | 370.8 | 16.8 |
| 141 | 95 | 2 | 8.5 | 6.9 | 971.4 | 282.7 | 448.0 | 275.0 | .... | 25.2 | 259.9 | 14.6 |
| 141 | 95 | 2 | 10.1 | 7.2 | 1314.6 | 190.1 | 610.1 | 309.5 | .... | 378.4 | 530.8 | 7.9 |
| 141 | 95 | 2 | 14.7 | 13.7 | 1649.7 | 273.9 | 262.0 | 739.8 | .... | 47.8 | 1129.1 | 31.5 |
| 141 | 95 | 2 | 15.1 | 14.1 | 2440.9 | 519.2 | 608.5 | 718.3 | .... | 208.3 | 1260.6 | 4.1 |
| 141 | 95 | 2 | 21.3 | 20.3 | 1905.3 | 317.0 | 584.0 | 1039.0 | .... | 304.0 | 1357.7 | 3.0 |

**Tabella 9.2 Esempio di rapporto tra le varianze dei dati microaggregati e dei dati originali. Variabile Addetti. Anno 1995**

| | Italia | | | |
|---|---|---|---|---|
| | Nord | | Centro-sud | |
| Ateco | Nord-ovest | Nord-est | Centro | Sud e Isole |
| 141 | 0.92 | 0.92 | 0.90 | 0.92 |
| 142-143-144-145 | 0.94 | 0.90 | 0.88 | 0.86 |
| 151 | 0.87 | 0.94 | 0.87 | 0.84 |
| 152-153-154 | 0.92 | 0.88 | 0.84 | 0.90 |
| 155 | 0.48 | 0.88 | 0.85 | 0.94 |
| 156 | 0.63 | 0.77 | 0.85 | 0.92 |
| 157-158 | 0.96 | 0.80 | 0.87 | 0.76 |
| 159 | 0.90 | 0.91 | 0.81 | 0.93 |
| … | … | … | … | … |
| 177 | 0.91 | 0.83 | 0.89 | 0.77 |
| 181 | 0.89 | 0.92 | 0.84 | 0.85 |
| 182-183 | 0.77 | 0.74 | 0.81 | 0.81 |
| 191 | 0.92 | 0.96 | 0.90 | 0.93 |
| 192 | 0.96 | 0.79 | 0.80 | 0.90 |
| 193 | 0.85 | 0.91 | 0.90 | 0.82 |
| 201 | 0.89 | 0.88 | 0.88 | 0.90 |
| 202-204 | 0.86 | 0.70 | 0.92 | 0.86 |
| 203 | 0.94 | 0.96 | 0.96 | 0.88 |
| 205 | 0.95 | 0.93 | 0.88 | 0.89 |
| 211-212 | 0.96 | 0.93 | 0.90 | 0.77 |
| 221-222-223 | 0.72 | 0.81 | 0.95 | 0.93 |
| 231-232-233 | | | 0.88 | |
| 241 | 0.92 | 0.87 | 0.85 | 0.79 |
| … | … | … | … | |

Come ulteriore termine di confronto vengono riportate insieme ai dati microaggregati le principali caratteristiche statistiche dei dati originali: numero di osservazioni; media e varianza di ogni variabile; matrice delle somme dei quadrati e dei prodotti incrociati, matrice di varianze e covarianze e matrice di correlazione per ogni sottopopolazione.

Le elaborazioni sono state effettuate con il pacchetto Sas *System* e, per quanto riguarda la microaggregazione, utilizzando l'applicazione software Masq (Microaggregazione su Asse Singolo per variabili Quantitative) sviluppata con il modulo Sas/AF (Pagliuca e Seri, 1999a).

Per quanto riguarda la valutazione della qualità dell'informazione che viene fornita con i metodi di microaggregazione, una sperimentazione condotta sui dati delle rilevazioni sui conti economici delle imprese riferite agli anni 1993 e 1994 (Corsini *et al*., 1999; Contini *et al*., 1998) ha confermato che i criteri di ordinamento del tipo "singolo asse", applicati per produrre le basi dati allegate, tendono a fornire risultati tanto più soddisfacenti quanto più l'asse prescelto per la microaggregazione è correlato (positivamente o negativamente) con le variabili oggetto di microaggregazione. Questo spiega perché, quando si intende mettere a disposizione degli utenti le singole variabili microeconomiche derivanti dai bilanci di impresa, l'utilizzo di dati microaggregati comporta perdite di informazione relativamente ridotte; quasi tutte le poste di un conto economico sono infatti correlate alla dimensione dell'impresa.

Le cose non stanno più in questi termini quando si considerano indicatori nella forma di rapporti caratteristici, cioè funzioni (non lineari) delle variabili rilevate, come ad esempio il "valore aggiunto per addetto" o il costo del lavoro per unità di prodotto.

La sperimentazione ha mostrato che le quote di variabilità e di correlazione preservate rispetto alle variabili originali possono divenire particolarmente basse nel caso dei rapporti caratteristici. E' necessario quindi tenere conto di queste problematiche nell'uso dei dati microaggregati per il calcolo di questi rapporti.

L'utilizzo di dati microaggregati per costruire indicatori dinamici richiede cautele ancora maggiori rispetto a quelle evidenziate a proposito dei rapporti caratteristici. Gli indicatori dinamici vengono normalmente utilizzati per effettuare analisi microeconometriche su variabili che colgono la dinamica delle imprese nel tempo (ad esempio, crescita dell'occupazione, saggio di variazione del costo del lavoro per unità di prodotto, eccetera). Anche in questo caso, fra l'asse prescelto per la microaggregazione e gli indicatori dinamici vi è generalmente scarsa correlazione; di conseguenza le quote di variabilità e di correlazione rispetto alle variabili originali possono divenire particolarmente basse.

Risultati più accettabili, sempre dal punto di vista empirico, si sono verificati nella stima dei parametri di alcuni modelli econometrici (Corsini *et al*., 1999; Contini *et al*., 1998) che comunque vanno interpretati con prudenza anche alla luce della considerazione intuitiva che un processo di aggregazione, per quanto ridotto, comporta una perdita di variabilità e una crescita della correlazione tra le variabili.

## 9.3 Il Laboratorio Adele per l'analisi dei dati elementari[19]

Abbiamo visto nei capitoli precedenti che, tradizionalmente, la statistica ufficiale, sia in Italia che all'estero, viene esternata in forma aggregata o, comunque, in modo che non si possano trarre riferimenti individuali tali da consentire il collegamento con soggetti identificabili.

Il soddisfacimento del vincolo di tutela della riservatezza impone, infatti, un limite nel livello di dettaglio che può essere raggiunto nella descrizione dei fenomeni oggetto della rilevazione. Anche nel caso di rilascio di collezioni campionarie di dati elementari (*file standard*) che, sia dal punto dei vista lecito dell'analisi statistica che da quello illecito dell'identificabilità di un interessato forniscono il massimo potenziale informativo, si ricorre a metodologie statistiche che limitano il contenuto di informazioni del file originale. Inoltre, solo alcune indagini in ambito sociale producono un *file standard*. Per far fronte in maniera soddisfacente alle esigenze conoscitive degli studiosi gli Istituti nazionali di statistica sono stati sollecitati ad intraprendere iniziative alternative anche di carattere amministrativo. Sono stati così costituiti i cosiddetti *Data Analysis Center* (Dac) che sono dei luoghi fisici cui possono accedere ricercatori e studiosi per effettuare le proprie analisi statistiche sotto il controllo diretto dell'Istituto di statistica titolare dei dati. I Dac esistenti sono iniziative relativamente recenti e ancora poco diffuse (in Europa la Cbs olandese ne è dotata, Eurostat lo ha appena istituito). Il Dac italiano nasce nel 1998, e si chiama Laboratorio Adele per l'analisi dei dati elementari.

Principale obiettivo del Laboratorio Adele è offrire a un'utenza esterna "esperta" la possibilità di analizzare dati elementari delle principali indagini dell'Istat, spostando la fase di verifica della tutela della riservatezza sull'*output* dell'analisi statistica piuttosto che sull'input, come avviene nel caso dei *file standard*. Questo approccio nasce dalla considerazione che la Statistica è, per sua natura, pratica di sintesi nel descrivere fenomeni che, nella maggior parte dei casi, si esprimono come indici o modelli (parametri di modelli) sicuri per quanto riguarda la tutela della riservatezza degli interessati. Contestualmente la tutela della riservatezza viene garantita sotto diversi aspetti:

- legalmente, attraverso la sottoscrizione di un modulo contratto che impegna l'utente al rispetto di norme di comportamento specifiche;
- fisicamente, attraverso il controllo dell'ambiente di lavoro. Il Laboratorio è collocato presso la sede centrale dell'Istat, è prevista la presenza di addetti che attendono al controllo della sala, le postazioni di lavoro sono isolate dalla rete telematica, le operazioni di input e *output* sono inibite agli utenti e vengono eseguite su richiesta dagli addetti;
- statisticamente, tramite il controllo cui sono sottoposti i risultati dell'analisi dell'utente preventivamente al rilascio.

La normativa sulla tutela delle persone e di altri soggetti rispetto al trattamento dei

[19] informazioni possono essere richieste a adele@istat.it oppure recandosi sul sito http://www.istat.it seguendo il percorso PRODOTTI E SERVIZI/LABORATORIO ANALISI DATI ELEMENTARI

dati personali distingue due diverse forme di rilascio dei dati: "comunicazione" o "diffusione" a seconda che venga data conoscenza di dati personali a "uno o più soggetti determinati diversi dall'interessato" o a "soggetti indeterminati". Il Laboratorio Adele si configura come una forma di comunicazione di dati a soggetti non facenti parte del Sistema statistico nazionale (il Sistan per il quale la circolazione dei dati è privilegiata e sottoposta a specifico regolamento) e si rivolge ad un'utenza "specializzata" che, per esclusivi motivi di ricerca, abbia necessità di elaborare dati elementari di impresa oppure dati provenienti da indagini in campo sociale, per i quali non risulta sufficiente il livello di dettaglio del *file standard*, se disponibile.

Il Laboratorio Adele si rivolge principalmente ad utenti "specializzati" in grado di essere autonomi per quanto riguarda l'elaborazione ed interpretazione dei dati, nonché l'utilizzo degli strumenti hardware e software messi a disposizione.

Il Laboratorio Adele si configura, quindi, come un ulteriore strumento a disposizione del mondo della ricerca scientifica per accedere all'informazione statistica ufficiale. Cioè, Adele non sostituisce i tradizionali canali di diffusione dell'Istat (*file standard*, elaborazioni personalizzate) ma serve per rispondere efficacemente a quelle esigenze conoscitive che questi strumenti non riescono a soddisfare. Pensiamo, ad esempio, alla stima di parametri di modelli econometrici o alla costruzione di indicatori socio-demografici che richiedono l'elaborazione diretta dei dati elementari.

I dati che si trovano presso il Laboratorio Adele sono i dati convalidati dagli uffici preposti alle singole indagini e, in linea di principio, mantengono il massimo contenuto informativo. L'impegno dell'Istat, in tal senso, è quello di mettere a disposizione della ricerca scientifica quanti più dati sia possibile, compatibilmente con le norme sulla tutela della riservatezza e le caratteristiche dei dati.

L'accesso al Laboratorio per l'analisi di dati elementari, privi di elementi identificativi diretti, è consentito solo per motivi di ricerca ad utenti appartenenti ad organizzazioni, pubbliche o private, (che abbiano sottoscritto il *Codice di deontologia e buona condotta per i trattamenti di dati personali per scopi statistici e di ricerca scientifica*) rientranti in una delle seguenti categorie:

(a) istituti di istruzione universitaria o post-universitaria;

(b) enti, istituti o società scientifiche per i quali l'attività di ricerca scientifica risulti dagli scopi istituzionali o sia altrimenti documentabile.

Ogni singola richiesta di accesso al Laboratorio Adele per essere accolta deve ricevere l'autorizzazione da parte del Presidente dell'Istat. Il richiedente è vincolato dalla firma di un modulo-contratto in cui deve:

(a) indicare le proprie generalità;

(b) fornire un'adeguata descrizione del progetto di ricerca facendo risultare in maniera chiara le finalità di studio dello stesso;

(c) indicare con sufficiente dettaglio l'insieme dei dati che intende analizzare: il file dati tra quelli disponibili o l'indagine di interesse, l'anno o il periodo cui si riferiscono i dati o l'indagine, le variabili o i quesiti della rilevazione che si intende utilizzare nella rilevazione;

(d) indicare le modalità di trattamento degli stessi;

(e) indicare il software tra quelli disponibili presso il Laboratorio, che intende utilizzare; in alternativa specificare il software commerciale con licenza valida che egli stesso si impegna a fornire su supporto originale per l'installazione presso il Laboratorio;

(f) specificare in maniera dettagliata la tipologia di *output* che intende prelevare al termine dell'analisi;

(g) fornire una stima preventiva del tempo necessario per l'elaborazione;

L'Istat provvede ad un'istruttoria preliminare della richiesta. Se necessario, richiede la documentazione o eventuali informazioni mancanti. L'istruttoria ha lo scopo, in particolare, di:

- verificare che non sia possibile il raggiungimento degli obiettivi dichiarati tramite un *file standard* esistente o attraverso altre forme di diffusione più consone;
- verificare la disponibilità dei dati richiesti;
- verificare la compatibilità della richiesta con le norme sulla comunicazione dei dati.

La richiesta viene, quindi, sottomessa al Presidente dell'Istat per l'autorizzazione. Ogni singola richiesta di accesso al Laboratorio Adele per essere accolta deve ricevere l'autorizzazione da parte del Presidente dell'Istat.

E' importante sottolineare che i risultati delle elaborazioni condotte presso il Laboratorio Adele da ricercatori esterni al Sistan non costituiscono in alcun modo fonte di statistica ufficiale e sono piena responsabilità degli autori stessi.

I risultati delle elaborazioni vengono rilasciati agli utenti solo dopo essere stati controllati per assicurarsi che non contengano violazioni dal punto di vista della tutela della riservatezza.

# APPENDICE A.1

Decreto legislativo 6 settembre 1989, n. 322

Pubblicato nella Gazz. Uff. 22 settembre 1989, n.222

"Norme sul Sistema statistico nazionale e sulla riorganizzazione dell'Istituto nazionale di statistica, ai sensi dell'art.24 della legge 23 agosto 1988, n. 400"[20]

[20] Versione aggiornata al mese di ottobre 2001.

**Decreto legislativo 6 settembre 1989, n. 322** - Pubblicato nella Gazz. Uff. 22 settembre 1989, n.222

**Norme sul Sistema statistico nazionale e sulla riorganizzazione dell'Istituto nazionale di statistica, ai sensi dell'art.24 della legge 23 agosto 1988, n. 400.**

IL PRESIDENTE DELLA REPUBBLICA

Visti gli articoli 76 e 87 della Costituzione;
Visto l'art.24 della legge 23 agosto 1988n.400, recante delega al Governo per l'emanazione di norme di riforma degli enti e degli organismi pubblici di informazione statistica;
Acquisto il parere delle competenti commissioni parlamentari previsto dal citato articolo 24;
Sulla proposta del Presidente del Consiglio dei ministri e del Ministro per gli Affari regionali ed i problemi istituzionali, di concerto con i Ministri dell'Interno, dell'Agricoltura e delle foreste, della Sanità, del Bilancio e della programmazione economica e del Tesoro;

EMANA

il seguente decreto legislativo:

## CAPO I
## SISTEMA STATISTICO NAZIONALE

### Art.1

*1.1. Oggetto della disciplina*

1. Il presente decreto disciplina, in base ai principi ed ai criteri direttivi di cui all'art.24 della legge 23 agosto 1988, n.400, le attività di rilevazione, elaborazione, analisi e diffusione e archiviazione dei dati statistici svolte dagli enti ed organismi pubblici di informazione statistica, al fine di realizzare l'unità di indirizzo, l'omogeneità organizzativa e la razionalizzazione dei flussi informativi a livello centrale e locale, nonché l'organizzazione e il funzionamento dell'Istituto nazionale di statistica.

2. L'informazione statistica ufficiale è fornita al Paese e agli organismi internazionali attraverso il Sistema statistico nazionale.

### Art.2

*1.2. Ordinamento del Sistema*

*1.3. statistico nazionale*

1. Fanno parte del Sistema statistico nazionale[1] :

*a)* l'Istituto nazionale di statistica (ISTAT);

*b)* gli uffici di statistica centrali e periferici delle amministrazioni dello Stato e delle amministrazioni ed aziende autonome, istituiti ai sensi dell'art.3;

*c)* gli uffici di statistica delle regioni e delle province autonome;

*d)* gli uffici di statistica delle province;

*e)* gli uffici di statistica dei comuni singoli o associati e delle unità sanitarie locali;

*f)* gli uffici di statistica delle camere di commercio, industria, artigianato e agricoltura;

*g)* gli uffici di statistica, comunque denominati, di amministrazioni e enti pubblici individuati ai sensi dell'art.4 ;

*h)* gli altri enti ed organismi pubblici di informazione statistica individuati con decreto del Presidente del Consiglio dei ministri .

### Art.3

*1.4. Uffici di statistica*

1. Presso le amministrazioni centrali dello Stato e presso le aziende autonome sono istituiti uffici di statistica, posti alle dipendenze funzionali dell'ISTAT.

2. Gli uffici di statistica sono ordinati anche secondo le esigenze di carattere tecnico indicate dall'Istat. Ad ogni ufficio è preposto un dirigente o funzionario designato dal Ministro competente, sentito il presidente dell'ISTAT.

3. Le attività e le funzioni degli uffici statistici delle province, dei comuni e delle camere di commercio, industria, artigianato e agricoltura sono regolate dalla legge 16 novembre 1939, n. 1823, e dalle relative norme di attuazione, nonché dal presente decreto nella parte applicabile. Entro sei mesi dalla data di entrata in vigore del presente decreto, gli enti locali, ivi comprese le unità sanitarie locali che non vi abbiano ancora provveduto, istituiscono l'ufficio di statistica anche in forma associata o consortile. I comuni con più di 100.000 abitanti

[1] Ai sensi della legge n.125/98,art.2, comma 1, "Al Sistema statistico (Sistan) di cui al decreto legislativo 6 settembre 1989, n.322, partecipano i soggetti privati che svolgono funzioni o servizi d'interesse pubblico o si configurino come essenziali per il raggiungimento degli obiettivi del Sistema stesso. Tali soggetti sono individuati con decreto del Presidente del Consiglio dei ministri, secondo criteri che garantiscano il rispetto dei principi di imparzialità e completezza dell'informazione statistica. Ad essi si applicano le disposizioni di cui al citato decreto legislativo n.322 del 1989".

istituiscono con effetto immediato un ufficio di statistica che fa parte del Sistema statistico nazionale.

4. Gli uffici di statistica costituiti presso le prefetture assicurano, fatte salve le competenze a livello regionale del commissario del Governo previste dall'art.13, comma 1, lettera c), della legge 23 agosto 1988, n.400, anche il coordinamento, il collegamento e l'interconnessione a livello provinciale di tutte le fonti pubbliche preposte alla raccolta ed alla elaborazione dei dati statistici, come individuate dall'Istat.

5. Gli uffici di statistica di cui ai commi 2, 3 e 4 esercitano le proprie attività secondo le direttive e gli atti di indirizzo emanati dal comitato di cui all'art.17.

**Art.4**

*1.5. Uffici di statistica di enti*
*1.6. e di amministrazioni pubbliche*

1. Presso enti ed organismi pubblici può essere costituito, sulla base di direttive del Presidente del Consiglio dei ministri, sentiti il Ministro vigilante ed il presidente dell'Istat, un ufficio di statistica, cui attribuire i compiti di cui all'art.6.

2. Gli uffici di statistica di cui al comma 1 sono costituiti tenendo conto dell'importanza delle attività svolte dall'ente o dall'amministrazione ai fini della informazione statistica nazionale e delle esigenze di completamento del sistema informativo nazionale. Nella individuazione degli uffici, si terrà conto del grado di specializzazione e della capacità di elaborazione del sistema informativo degli enti e degli organismi medesimi.

3. Gli uffici costituiti ai sensi del comma 1 sono inseriti nell'ambito del Sistema statistico nazionale di cui all'art.2 e sono sottoposti alla disciplina del presente decreto, in quanto applicabile.

4. Gli enti che svolgono la loro attività nelle materie contemplate nell'art.1 del decreto legislativo del Capo provvisorio dello Stato 17 luglio 1947, n. 691, ancorché non rientranti nel Sistema statistico nazionale, forniranno allo stesso i dati aggregati elaborati nell'ambito delle rilevazioni statistiche di competenza. Essi uniformano la propria attività statistica ai principi del presente decreto ed a quelli definiti in sede comunitaria per l'armonizzazione delle legislazioni nazionali in materia di prevenzione e repressione dell'utilizzo dei proventi derivanti da attività illegali.

5. Le sanzioni di cui all'art.11 si applicano anche alle violazioni delle disposizioni statistiche emanate in materia valutaria, fermo restando il procedimento sanzionatorio disciplinato dal testo unico delle norme di legge in materia valutaria, approvato con decreto del Presidente della Repubblica 31 marzo 1988, n. 148.

**Art.5**

*1.7. Uffici di statistica delle regioni*
*1.8. e delle province autonome*

1. Spetta a ciascuna regione ed alle province autonome di Trento e Bolzano istituire con propria legge uffici di statistica.

2. Il Consiglio dei ministri adotta atti di indirizzo e di coordinamento ai sensi dell'art.2, comma 3, lettera d), della legge 23 agosto 1988, n. 400, per assicurare unicità di indirizzo dell'attività statistica di competenza delle regioni e delle province autonome.

3. L'ISTAT esercita nei confronti degli uffici di cui al comma 1 poteri di indirizzo e coordinamento tecnici, allo scopo di renderne omogenee le metodologie.

**Art.6**

*1.9. Compiti degli uffici di statistica*

1. Gli uffici di statistica del Sistema statistico nazionale, oltre agli altri compiti attribuiti dalla normativa che li riguarda:
*a)* promuovono e realizzano la rilevazione, l'elaborazione, la diffusione e l'archiviazione dei dati statistici che interessano l'amministrazione di appartenenza, nell'ambito del programma statistico nazionale;
*b)* forniscono al Sistema statistico nazionale i dati informativi previsti del programma statistico nazionale relativi all'amministrazione di appartenenza, anche in forma individuale ma non nominativa ai fini della successiva elaborazione statistica;
*c)* collaborano con le altre amministrazioni per l'esecuzione delle rilevazioni previste dal programma statistico nazionale;
*d)* contribuiscono alla promozione e allo sviluppo informatico a fini statistici degli archivi gestionali e delle raccolte di dati amministrativi.

2. Gli uffici attuano l'interconnessione ed il collegamento dei sistemi informativi dell'amministrazione di appartenenza con il Sistema statistico nazionale. Per attuare il collegamento tra il Sistema informativo dell'anagrafe tributaria ed il Sistema statistico nazionale, la Presidenza del

Consiglio dei ministri promuove, entro sei mesi dalla data di entrata in vigore del presente decreto, specifiche intese tra il Ministero delle finanze e l'Istituto nazionale di statistica anche al fine di assicurare il pieno rispetto dell'anonimato dei singoli contribuenti e del segreto fiscale.

3. Per i compiti di cui al comma 1, gli uffici di statistica hanno accesso a tutti i dati statistici in possesso dell'amministrazione di appartenenza, salvo eccezioni relative a categorie di dati di particolare riservatezza espressamente previste dalla legge. Essi possono richiedere all'amministrazione di appartenenza elaborazioni di dati necessarie alle esigenze statistiche previste dal programma statistico nazionale.

4. Per esigenze particolari, connesse a determinate rilevazioni statistiche previste dal programma statistico nazionale, il presidente dell'Istat, sentito il comitato di cui all'art.17, può richiedere la comunicazione al Sistema, da parte degli uffici, di categorie di dati in forma nominativa. Sono fatte salve le riserve previste dalla legge.
5. In casi particolari, l'amministrazione o gli enti di appartenenza possono individuare ulteriori categorie di dati assoggettabili anche per tempi determinati a vincolo di riservatezza, dandone comunicazione al comitato di cui all'art.17.

6. Gli uffici di statistica inoltrano entro il 31 marzo di ciascun anno al presidente dell'Istat e all'amministrazione di appartenenza un rapporto annuale sull'attività svolta.

***Art.6 bis***

1.10. Trattamenti di dati personali

*1. I soggetti che fanno parte o partecipano al Sistema statistico nazionale possono raccogliere ed ulteriormente trattare i dati personali necessari per perseguire gli scopi statistici previsti dal presente decreto, dalla legge o dalla normativa comunitaria, qualora il trattamento di dati anonimi non permetta di raggiungere i medesimi scopi.*

*2. Nel programma statistico nazionale sono illustrate le finalità perseguite e le garanzie previste dal presente decreto e dalla legge 31 dicembre 1996, n. 675. Il programma indica anche i dati di cui agli articoli 22 e 24 della medesima legge, le rilevazioni per le quali i dati sono trattati e le modalità di trattamento. Il programma è adottato sentito il Garante per la protezione dei dati personali.*

*3. Quando sono raccolti per altri scopi, i dati personali possono essere ulteriormente trattati per scopi statistici, se ciò è previsto dal presente decreto, dalla legge, dalla normativa comunitaria o da un regolamento.*

*4. I dati personali raccolti specificamente per uno scopo possono essere trattati dai soggetti di cui al comma 1 per altri scopi statistici di interesse pubblico previsti ai sensi del comma 3, quando questi ultimi sono chiaramente determinati e di limitata durata. Tale eventualità, al pari di quella prevista del medesimo comma 3, è chiaramente rappresentata agli interessati al momento della raccolta o quando ciò non è possibile, è resa preventivamente nota al pubblico e al Garante nei modi e nei termini previsti dal codice di deontologia e di buona condotta.*

*5. I dati personali sono resi anonimi dopo la raccolta o quando la loro disponibilità non sia più necessaria per i propri trattamenti statistici.*

*6. I dati identificativi, qualora possano essere conservati, sono custoditi separatamente da ogni altro dato personale salvo che ciò, in base ad un atto motivato per iscritto, risulti impossibile in ragione delle particolari caratteristiche del trattamento o comporti un impiego di mezzi manifestamente sproporzionato. I dati personali trattati per scopi statistici sono conservati separatamente da ogni altro dato personale trattato per finalità che non richiedano il loro utilizzo.*

*7. I dati identificativi, qualora possano essere conservati, sono abbinabili ad altri dati, sempre che l'abbinamento sia temporaneo ed essenziale per i propri trattamenti statistici.*

*8. In caso di esercizio dei diritti dell'interessato ai sensi dell'articolo 13 della legge 31 dicembre 1996, n. 675, l'aggiornamento, la rettificazione o l'integrazione dei dati sono annotate senza modificare questi ultimi qualora il risultato di tali operazioni non produca effetti significativi sull'analisi statistica o sui risultati statistici.*

**Art.7**
**Obbligo di fornire dati statistici**

1. Salvo diversa indicazione del comitato di cui all'art.17, è fatto obbligo a tutte le amministrazioni, enti ed organismi pubblici di fornire tutti i dati e le notizie che vengono loro richiesti per rilevazioni previste dal programma statistico nazionale. Sono sottoposti al medesimo

obbligo i soggetti privati per le rilevazioni statistiche, rientranti nel programma stesso, espressamente indicate con delibera del Consiglio dei ministri.

*2. Non rientrano nell'obbligo di cui al comma 1 i dati personali di cui agli articoli 22 e 24 della legge 31 dicembre 1996, n.675.*

3. Coloro che, richiesti di dati e notizie ai sensi del comma 1, non li forniscano, ovvero li forniscono scientemente errati o incompleti, sono soggetti ad una sanzione amministrativa pecuniaria, nella misura di cui all'art.11, che è applicata secondo il procedimento ivi previsto.

**Art.8**

*1.11. Segreto di ufficio degli addetti*
*1.12. agli uffici di statistica*

1. Le norme in materia di segreto d'ufficio previste dal vigente ordinamento dell'impiego civile dello Stato si applicano a tutti gli addetti agli uffici di statistica previsti dagli articoli 3, 4 e 5.
2. Resta fermo il disposto dell'art.15 del decreto del Presidente della Repubblica 2 novembre 1976, n. 784.

**Art.9**

*1.13. Disposizioni per la tutela*
*1.14. del segreto statistico*

1. I dati raccolti nell'ambito di rilevazioni statistiche comprese nel programma statistico nazionale da parte degli uffici di statistica non possono essere esternati se non in forma aggregata, *in modo che non se ne possa trarre alcun riferimento relativamente a persone identificabili e possono essere utilizzati solo per scopi statistici.*

*2. I dati di cui al comma 1 non possono essere comunicati o diffusi, se non in forma aggregata e secondo modalità che rendano non identificabili gli interessati ad alcun soggetto esterno, pubblico o privato, né ad alcun ufficio della pubblica amministrazione. In ogni caso, i dati non possono essere utilizzati al fine di identificare nuovamente gli interessati.*

3. In casi eccezionali, l'organo responsabile dell'amministrazione nella quale è inserito l'ufficio di statistica può, sentito il comitato di cui all'art.17, chiedere al Presidente del Consiglio dei ministri l'autorizzazione ad estendere il segreto statistico anche a dati aggregati.

4. Fatto salvo quanto previsto dall'art.8, non rientrano tra i dati tutelati dal segreto statistico gli estremi identificativi di persone o di beni, o gli atti certificativi di rapporti, *provenienti da pubblici registri, elenchi, atti o documenti conoscibili da chiunque.*

**Art.10**
**Accesso ai dati statistici**

1. I dati elaborati nell'ambito delle rilevazioni statistiche comprese nel programma statistico nazionale sono patrimonio della collettività e vengono distribuiti per fini di studio e di ricerca a coloro che li richiedono secondo la disciplina del presente decreto, fermi restando i divieti di cui all'art.9.

2. 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.

3. Presso la sede centrale dell'Istat in Roma, presso le sedi regionali dell'Istat, nonché presso gli uffici di statistica delle prefetture, sono costituiti uffici di collegamento del Sistema statistico nazionale con il pubblico. Gli altri uffici di statistica di cui all'art.2 possono costituire uffici di collegamento del Sistema statistico nazionale col pubblico, dandone comunicazione all'Istat.

4. Enti od organismi pubblici, persone giuridiche, società, associazioni e singoli cittadini hanno il diritto di accedere ai dati di cui al comma 1 facendone richiesta agli uffici di cui al comma 3. I dati, se non immediatamente disponibili, vengono consegnati ai richiedenti nel tempo strettamente necessario per la riproduzione, con rimborso delle spese, il cui importo è stabilito dall'Istat.

5. Il comitato di cui all'art.17 stabilisce le modalità di funzionamento degli uffici costituiti ai sensi del comma 3.

6. Alle amministrazioni e agli enti pubblici che fanno parte del Sistema statistico nazionale vengono periodicamente trasmessi, a cura dell'Istat, i dati elaborati dal Sistema statistico nazionale.

7. Le procedure per l'accesso, da parte della Camera dei deputati e del Senato della Repubblica e dei loro organi, nonché dei singoli loro componenti ai dati elaborati dal Sistema statistico

nazionale sono disciplinate dai regolamenti parlamentari.

**Art.11**

*1.15. Sanzioni amministrative*

1. Le sanzioni amministrative pecuniarie, di cui all'art.7, sono stabilite:

*a)* nella misura minima di lire quattrocentomila e massima di lire quattro milioni per le violazioni da parte di persone fisiche;

*b)* nella misura minima di lire un milione e massima di lire dieci milioni per le violazioni da parte di enti e società.

2. L'accertamento delle violazioni, ai fini dell'applicazione delle sanzioni amministrative pecuniarie, è effettuato dagli uffici di statistica, facenti parte del Sistema statistico nazionale di cui all'art.2, che siano venuti a conoscenza della violazione.

3. Il competente ufficio di statistica redige motivato rapporto in ordine alla violazione e, previa contestazione degli addebiti agli interessati secondo il procedimento di cui agli articoli 13 e seguenti della legge 24 novembre 1981, n. 689, lo trasmette al prefetto della provincia, il quale procede ai sensi dell'art.18 e seguenti della medesima legge. Dell'apertura del procedimento è data comunicazione all'Istat.

**Art.12**

*1.16. Commissione per la garanzia dell'informazione statistica*

1. Al fine di garantire il principio della imparzialità e della completezza dell'informazione statistica è istituita, presso la Presidenza del Consiglio dei ministri, la commissione per la garanzia dell'informazione statistica. In particolare, la commissione vigila:

*a)* sulla imparzialità e completezza dell'informazione statistica *e contribuisce alla corretta applicazione delle norme* che disciplinano la tutela della riservatezza delle informazioni fornite all'Istat e ad altri enti del Sistema statistico nazionale, *segnalando anche al Garante per la protezione dei dati personali i casi di inosservanza delle medesime norme o assicurando altra collaborazione nei casi in cui la natura tecnica dei problemi lo richieda;*

*b)* sulla qualità delle metodologie statistiche e delle tecniche informatiche impiegate nella raccolta, nella conservazione e nella diffusione dei dati;

c) sulla conformità delle rilevazioni alle direttive degli organismi internazionali e comunitari.

2. La commissione, nell'esercizio delle attività di cui al comma 1, può formulare osservazioni e rilievi al presidente dell'Istat, il quale provvede a fornire i necessari chiarimenti entro trenta giorni dalla comunicazione, sentito il Comitato di cui all'art.17; qualora i chiarimenti non siano ritenuti esaustivi, la commissione ne riferisce al Presidente del Consiglio dei ministri. Esprime inoltre parere sul programma statistico nazionale ai sensi dell'art.13, *ed è sentita ai fini della sottoscrizione dei codici di deontologia e di buona condotta relativi al trattamento dei dati personali nell'ambito del Sistema statistico nazionale.*

3. La commissione è composta di nove membri, nominati, entro sei mesi dalla data di entrata in vigore del presente decreto, con decreto del Presidente della Repubblica, su proposta del Presidente del Consiglio dei ministri, dei quali sei scelti tra professori ordinari in materie statistiche, economiche ed affini o direttori di istituti di statistica o di ricerca statistica non facenti parte del Sistema statistico nazionale, e tre tra alti dirigenti di enti e amministrazioni pubbliche, che godano di grande prestigio e competenza nelle discipline e nei campi collegati alla produzione, diffusione e analisi delle informazioni statistiche e che non siano preposti ad uffici facenti parte del Sistema statistico nazionale. Possono essere nominati anche cittadini di Paesi comunitari che abbiano i medesimi requisiti.

4. Il presidente della commissione è eletto dagli stessi membri.

5. I membri della commissione restano in carica sei anni e non possono essere confermati.

6. La commissione si riunisce almeno due volte all'anno e redige un rapporto annuale, che si allega alla relazione al Parlamento sull'attività dell'Istat.

7. Partecipa alle riunioni il presidente dell'Istat.

8. Alle funzioni di segreteria della commissione provvede il Segretario generale della Presidenza del Consiglio dei Ministri che istituisce, a questo fine, un apposito ufficio, che può avvalersi anche di esperti esterni ai sensi della legge 23 agosto 1988, n. 400.

9. I compensi di cui all'art.20 per i membri della commissione sono posti a carico del bilancio dell'Istat.

**Art.13**

*1.17. Programma statistico nazionale*

1. Le rilevazioni statistiche di interesse pubblico affidate al Sistema statistico nazionale ed i relativi obiettivi sono stabiliti nel programma statistico nazionale.

2. Il programma statistico nazionale ha durata triennale e viene tenuto aggiornato.

3. Il programma statistico nazionale è predisposto dall'Istat, sottoposto al parere della commissione per la garanzia dell'informazione statistica di cui all'art.12 ed approvato con decreto del Presidente della Repubblica[2], su proposta del Presidente del Consiglio dei ministri, previa deliberazione del CIPE[3].

4. Gli aggiornamenti del programma statistico nazionale sono predisposti e approvati con la stessa procedura di cui al comma 3.

## CAPO II
## ORGANIZZAZIONE E FUNZIONI DELL'ISTAT

### Art.14

*1.18. Istituto nazionale di statistica*

1. L'Istituto centrale di statistica, istituito con legge 9 luglio 1926, n. 1162, assume la denominazione di Istituto nazionale di statistica (ISTAT).

2. L'Istituto nazionale di statistica è persona giuridica di diritto pubblico ed ha ordinamento autonomo secondo le disposizioni del presente decreto.

3. Sono organi dell'Istituto:
*a)* il presidente;
*b)* il comitato per l'indirizzo e il coordinamento dell'informazione statistica;
*c)* il consiglio;
*d)* il collegio dei revisori dei conti.

4. L'ISTAT è sottoposto alla vigilanza del Presidente del Consiglio dei ministri.

### Art.15

*1.19. Compiti dell'Istat*

1. L'ISTAT provvede:
*a)* alla predisposizione del programma statistico nazionale;
*b)* alla esecuzione dei censimenti e delle altre rilevazioni statistiche previste dal programma statistico nazionale ed affidate alla esecuzione dell'istituto;
*c)* all'indirizzo e al coordinamento delle attività statistiche degli enti ed uffici facenti parte del Sistema statistico nazionale di cui all'art.2;
*d)* all'assistenza tecnica agli enti ed uffici facenti parte del Sistema statistico nazionale di cui all'art.2, nonché alla valutazione, sulla base dei criteri stabiliti dal comitato di cui all'art.17, dell'adeguatezza dell'attività di detti enti agli obiettivi del programma statistico nazionale;
*e)* alla predisposizione delle nomenclature e metodologie di base per la classificazione e la rilevazione dei fenomeni di carattere demografico, economico e sociale. Le nomenclature e le metodologie sono vincolanti per gli enti ed organismi facenti parte del Sistema statistico nazionale;
*f)* alla ricerca e allo studio sui risultati dei censimenti e delle rilevazioni effettuate, nonché sulle statistiche riguardanti fenomeni d'interesse nazionale e inserite nel programma triennale;
*g)* alla pubblicazione e diffusione dei dati, delle analisi e degli studi effettuati dall'istituto ovvero da altri uffici del Sistema statistico nazionale che non possano provvedervi direttamente; in particolare alla pubblicazione dell'annuario statistico italiano e del Bollettino mensile di statistica;
*h)* alla promozione e allo sviluppo informatico a fini statistici degli archivi gestionali e delle raccolte di dati amministrativi;
*i)* allo svolgimento di attività di formazione e di qualificazione professionale per gli addetti al Sistema statistico nazionale;
*l)* ai rapporti con enti ed uffici internazionali operanti nel settore dell'informazione statistica;
*m)* alla promozione di studi e ricerche in materia statistica;
*n)* alla esecuzione di particolari elaborazioni statistiche per conto di enti e privati, remunerate a condizioni di mercato.

2. Per lo svolgimento dei propri compiti l'Istat si può avvalere di enti pubblici e privati e di società mediante rapporti contrattuali e convenzionali, nonché mediante partecipazione al capitale degli enti e società stessi.

3. L'ISTAT, nell'attuazione del programma statistico nazionale, si avvale degli uffici di statistica di cui all'art.2, come precisato dagli articoli 3 e 4.

---

[2] Il Psn è ora approvato con DPCM per effetto dell'art.2 della L.13/1991

[3] Ai sensi dell'art.6-bis, comma 2, "Il programma statistico è adottato sentito il Garante per la protezione dei dati personali"

4. L'ISTAT, per l'esercizio delle sue funzioni, procede con periodicità, almeno biennale, alla convocazione di una Conferenza nazionale di statistica.

5. L'ISTAT si avvale del patrocinio e della consulenza dell'avvocatura dello Stato.

**Art.16**

*1.20. Presidente*

1. Il presidente dell'Istituto nazionale di statistica, scelto tra i professori ordinari in materie statistiche, economiche ed affini, è nominato, ai sensi dell'art.3 della legge 23 agosto 1988, n. 400, con decreto del Presidente della Repubblica, su proposta del Presidente del Consiglio, previa deliberazione del Consiglio dei ministri. Egli ha la legale rappresentanza e provvede all'amministrazione dell'Istituto, assicurandone il funzionamento.

2. Il presidente può adottare provvedimenti di competenza del comitato di cui all'art.17 nei casi di urgente necessità, salvo ratifica dello stesso organo, da convocare immediatamente e comunque entro trenta giorni dalla data del provvedimento.

3. Il presidente, in caso di assenza o di impedimento, può delegare la legale rappresentanza e le altre funzioni inerenti al suo ufficio ad un membro del consiglio.

4. Il presidente può delegare, per l'esercizio di particolari attribuzioni, la legale rappresentanza dell'Istituto al direttore generale, ai direttori centrali, nonché ai dirigenti dei servizi ed uffici dell'Istituto stesso, nei limiti e con le modalità che saranno previsti nel regolamento di organizzazione di cui all'art.22.

5. Il presidente dura in carica quattro anni e può essere confermato una sola volta. Ad esso spetta una indennità di carica da determinarsi con decreto del Presidente del Consiglio dei ministri, di concerto con il Ministro del tesoro.

**Art.17**

*1.21. Comitato di indirizzo e coordinamento dell'informazione statistica*

1. E' costituito il comitato di indirizzo e coordinamento dell'informazione statistica per l'esercizio delle funzioni direttive dell'Istat nei confronti degli uffici di informazione statistica costituiti ai sensi dell'art.3.

2. Il comitato è composto:

*a)* dal presidente dell'Istituto che lo presiede;
*b)* da dieci membri in rappresentanza delle amministrazioni statali, di cui tre delle amministrazioni finanziarie, dotate dei più complessi sistemi di informazione statistica, indicate dal Presidente del Consiglio dei Ministri, sentito il presidente dell'Istat;
*c)* da un rappresentante delle regioni designato tra i propri membri dalla Conferenza permanente per i rapporti tra lo Stato, le regioni e le province autonome, di cui all'art.12 della legge 23 agosto 1988, n.400;
*d)* da un rappresentante dell'UPI;
*e)* da un rappresentante dell'Unioncamere;
*f)* da tre rappresentanti dell'ANCI;
*g)* da due rappresentanti di enti pubblici tra quelli dotati dei più complessi sistemi d'informazione;
*h)* dal direttore generale dell'Istat;
*i)* da due esperti scelti tra i professori ordinari di ruolo di prima fascia in materie statistiche, economiche ed affini.

3. Il comitato può essere integrato, su proposta del presidente, da rappresentanti di altre amministrazioni statali competenti per specifici oggetti di deliberazione.

4. I membri di cui alle lettere b), c), d), e), f) e g) del comma 2 sono nominati con decreto del Presidente del Consiglio dei Ministri, su proposta del Ministro o del rappresentante degli organismi interessati; i membri di cui alla lettera i) sono nominati con decreto del Presidente del Consiglio dei Ministri, su proposta del Ministro dell'università e della ricerca scientifica e tecnologica.

5. Il comitato dura in carica quattro anni. I suoi membri possono essere confermati per non più di due volte.

6. Il comitato emana direttive vincolanti nei confronti degli uffici di statistica costituiti ai sensi dell'art.3, nonché atti di indirizzo nei confronti degli altri uffici facenti parte del Sistema statistico nazionale di cui all'art.2. Le direttive sono sottoposte all'assenso dell'amministrazione vigilante, che si intende comunque dato qualora, entro trenta giorni dalla comunicazione, la stessa non formula rilievi. Delibera, su proposta del presidente, il programma statistico nazionale.

7. Il comitato si riunisce su convocazione del presidente ogni volta che questi o le amministrazioni e gli enti rappresentati ne ravvisino la necessità.

8. Il comitato è costituito con la nomina della maggioranza assoluta dei propri membri.

**Art.18**

*1.22. Consiglio dell'Istat*

1. Il consiglio dell'Istat programma, indirizza e controlla l'attività dell'Istituto.

2. Il consiglio è composto:
*a)* dal presidente dell'istituto, che lo presiede;
*b)* da tre membri designati, tra i propri componenti, dal comitato di cui all'art.17;
*c)* da cinque membri nominati dal Presidente del Consiglio dei Ministri, dei quali due professori ordinari oppure direttori di istituti di statistica o di ricerca statistica;
*d)* dal presidente della Commissione per la garanzia dell'informazione statistica di cui all'art.12.
3. Il direttore generale dell'Istituto partecipa alle riunioni del consiglio e ne è il segretario.

4. I membri del consiglio sono nominati con decreto del Presidente del Consiglio dei Ministri. I membri di cui alle lettere b) e c) del comma 2 durano in carica quattro anni; allo scadere del termine i singoli membri cessano dalle funzioni anche se siano stati nominati nel corso del quadriennio.

5. Il Consiglio è costituito con la nomina della maggioranza assoluta dei propri membri

**Art.19**

*1.23. Collegio dei revisori dei conti*

1. Il collegio dei revisori dei conti è nominato, per la durata di tre anni, con decreto del Presidente del Consiglio dei ministri ed è composto da:
*a)* un magistrato del Consiglio di Stato, con funzioni di presidente;
*b)* un dirigente della Presidenza del Consiglio dei ministri;
*c)* un dirigente del Ministero del tesoro.

2. Con il medesimo decreto sono nominati due membri supplenti.

3. Il collegio dei revisori dei conti accerta la regolare tenuta della contabilità e la corrispondenza del bilancio consuntivo alle risultanze dei libri e delle scritture contabili; verifica i risultati conseguiti rispetto agli obiettivi; esamina le giustificazioni fornite dall'Istituto in merito ad eventuali scostamenti. I componenti del collegio sono invitati alle sedute del Consiglio.

4. Ai fini della relazione annuale al Parlamento sulla gestione finanziaria, l'Istat trasmette alla Corte dei conti il conto consuntivo e gli allegati, nel termine di cui all'art.23, comma 3.

**Art.20**

*1.24. Compensi ai componenti degli organi collegiali dell'Istat*

1. I compensi per i componenti degli organi collegiali di cui agli articoli 12, 17, 18 e 19 sono determinati con decreto del Presidente del Consiglio dei Ministri, di concerto con il Ministro del tesoro.

**Art.21**

*1.25. Direttive e atti di indirizzo*

1. Le direttive e gli atti di indirizzo del comitato previsti dal comma 6 dell'art.17 hanno ad oggetto:
*a)* gli atti di esecuzione del programma statistico nazionale;
*b)* le iniziative per l'attuazione del predetto programma;
*c)* i criteri organizzativi e la funzionalità degli uffici di statistica delle amministrazioni dello Stato, anche ad ordinamento autonomo, nonché degli enti e degli uffici facenti parte del Sistema statistico nazionale;
*d)* i criteri e le modalità per l'interscambio dei dati indicati dall'art.6 fra gli uffici di statistica delle amministrazioni e degli enti facenti parte del Sistema statistico nazionale, assicurando, in ogni caso, il rispetto delle disposizioni di cui all'art.8.

**Art.22**

*1.26. Compiti del Consiglio*

1. Il presidente convoca il consiglio e fissa le materie da portare alla sua discussione.

2. Spetta al Consiglio:
*a)* di deliberare, entro il 30 aprile di ciascun anno, un piano annuale che evidenzi gli obiettivi, le spese previste per il successivo triennio e le previsioni annuali di entrata, con indicazioni separate di quelle proprie e di quelle a carico del bilancio statale, seguendone periodicamente lo stato di attuazione. In tale documento è altresì inserito, con atto separato, il piano annuale di attuazione del programma statistico nazionale di cui all'art.13;
*b)* di deliberare il bilancio preventivo, le relative variazioni e il conto consuntivo;
c) di deliberare il disegno organizzativo dell'Istituto, determinando gli uffici centrali e periferici e la loro organizzazione, fissandone i compiti e la dotazione di personale e di mezzi,

nonché il regolamento organico e la pianta organica del personale;
*d)* di deliberare i regolamenti sulla gestione finanziaria, economica e patrimoniale, tenendo conto della natura specifica e della autonomia dell'Istat;
*e)* di deliberare la partecipazione dell'Istat al capitale di enti e società, ai sensi dell'art.15, comma 2;
*f)* di nominare su proposta del presidente il direttore generale e i direttori centrali dell'Istituto.

3. Per la validità delle sedute del consiglio occorre la presenza di almeno sei componenti. Per la validità delle deliberazioni occorre il voto favorevole della maggioranza dei presenti. In caso di parità di voti prevale quello del presidente.

4. Le deliberazioni sugli oggetti di cui alle lettere a), b), c), d) ed e) del comma 2 sono approvate con decreto del Presidente del Consiglio dei Ministri, di concerto, quanto alla lettera c), con i Ministri del tesoro e per la funzione pubblica e, quanto alle lettere d) ed e), con il Ministro del tesoro.

**Art.23**

*1.27. Gestione finanziaria*

1. La gestione finanziaria dell'Istat si svolge sulla base di un bilancio pluriennale, redatto in relazione ai piani di attività e alle previsioni pluriennali di spesa di cui all'art.22, comma 2, lettera a).

2. Per ciascun esercizio la gestione finanziaria si svolge in base ad un bilancio preventivo annuale, coincidente con l'anno solare, deliberato dal consiglio entro il 31 ottobre dell'anno precedente e trasmesso alla Presidenza del Consiglio dei Ministri entro quindici giorni dalla deliberazione.

3. Entro il mese di aprile il consiglio delibera il conto consuntivo dell'esercizio precedente, che viene trasmesso alla Presidenza del Consiglio dei Ministri entro quindici giorni dalla deliberazione. Oltre alle relazioni del presidente e del collegio dei revisori dei conti, ad esso è allegato un documento sulla situazione patrimoniale, sulla dimostrazione dei risultati economici conseguiti e sulla situazione amministrativa.
4. Il sistema di classificazione, gli schemi del bilancio e dei conti e i documenti consuntivi saranno disciplinati dai regolamenti di cui all'art.22, comma 2, lettera d).

5. La relazione al bilancio deve illustrare anche gli aspetti economici della gestione, ponendo in evidenza lo stato di attuazione della programmazione, i costi ed i risultati conseguiti, nonché gli eventuali scostamenti.

**Art.24**

**1.28.** *Relazione al Parlamento*

1. Il Presidente del Consiglio dei Ministri trasmette al Parlamento, entro il 31 maggio di ciascun anno, una relazione sull'attività dell'Istat, sulla raccolta, trattamento e diffusione dei dati statistici della pubblica amministrazione, nonché sullo stato di attuazione del programma statistico nazionale in vigore.

2. Alla relazione è allegato il rapporto annuale di cui al comma 6 dell'art.12.

**Art.25**
**Abrogazioni di precedenti norme**

1. Sono abrogati nella parte incompatibile il regio decreto-legge 27 maggio 1929, n. 1285, convertito dalla legge 21 dicembre 1929, n. 2238, la legge 16 novembre 1939, n. 1823 (26), la legge 6 agosto 1966, n. 628, la legge 19 dicembre 1969, n. 1025, e tutte le altre norme incompatibili con il presente decreto.

**Art.26**

*1.29. Norme transitorie*

1. Entro tre mesi dalla data di entrata in vigore del presente decreto, le amministrazioni e gli enti di cui agli articoli 3 e 4 inviano alla Presidenza del Consiglio dei Ministri una relazione sulla situazione degli uffici di statistica esistenti e sui provvedimenti necessari per il loro adeguamento alle norme del presente decreto. Entro i successivi tre mesi, le amministrazioni e gli enti provvedono, anche sulla base delle eventuali direttive della Presidenza del Consiglio dei Ministri, alla riorganizzazione o istituzione degli uffici di statistica, secondo le norme del presente decreto.
2. L'ordinamento previsto dal presente decreto acquista efficacia sei mesi dopo la sua entrata in vigore.
3. Le disposizioni recate dal presente decreto non comportano oneri a carico del bilancio dello Stato.
Il presente decreto, munito del sigillo dello Stato, sarà inserito nella Raccolta ufficiale degli atti normativi della Repubblica italiana. E' fatto obbligo a chiunque spetti di osservarlo e di farlo osservare.

# APPENDICE A.2

Pubblicato nella Gazzetta Ufficiale del 1 ottobre 2002

"Codice di deontologia e di buona condotta per i trattamenti di dati personali a scopi statistici e di ricerca scientifica effettuati nell'ambito del sistema statistico nazionale"[21]

[21] Testo contenuto nella deliberazione del Garante per la protezione dei dati personali n. 13 del 31/7/2002

## CODICE DI DEONTOLOGIA E DI BUONA CONDOTTA PER I TRATTAMENTI DI DATI PERSONALI A SCOPI STATISTICI E DI RICERCA SCIENTIFICA EFFETTUATI NELL'AMBITO DEL SISTEMA STATISTICO NAZIONALE

*PREAMBOLO*

***Il presente codice è volto a garantire che l'utilizzazione di dati di carattere personale per scopi di statistica, considerati dalla legge di rilevante interesse pubblico e fonte dell'informazione statistica ufficiale intesa quale patrimonio della collettività, si svolga nel rispetto dei diritti, delle libertà fondamentali e della dignità delle persone interessate, in particolare del diritto alla riservatezza e del diritto all'identità personale.***

***Il codice è sottoscritto in attuazione degli articoli 6 e 10, comma 6, del decreto legislativo 30 luglio 1999, n. 281 e si applica ai trattamenti per scopi statistici effettuati nell'ambito del sistema statistico nazionale, per il perseguimento delle finalità di cui al decreto legislativo 6 settembre 1989, n. 322.***

La sua sottoscrizione è effettuata ispirandosi alle pertinenti fonti e documenti internazionali in materia di attività statistica e, in particolare:

a) alla Convenzione europea per la salvaguardia dei diritti dell'uomo e delle libertà fondamentali del 4 novembre 1950, ratificata dall'Italia con legge 4 agosto 1955, n. 848;

b) alla Carta dei diritti fondamentali dell'Unione Europea del 18 dicembre 2000, con specifico riferimento agli artt. 7 e 8;

c) alla Convenzione n. 108 adottata a Strasburgo il 28 gennaio 1981, ratificata in Italia con legge 21 febbraio 1989, n. 98;

d) alla direttiva n. 95/46/CE del Parlamento europeo e del Consiglio dell'Unione Europea del 24 ottobre 1995;

e) alla Raccomandazione del Consiglio d'Europa n. R(97)18, adottata il 30 settembre 1997;

f) all'articolo 10 del Regolamento (CE) n. 322/97 del Consiglio dell'Unione Europea del 17 febbraio 1997.

***Gli enti, gli uffici e i soggetti che applicano il seguente codice sono chiamati ad osservare anche il principio di imparzialità e di non discriminazione nei confronti di altri utilizzatori, in particolare, nell'ambito della comunicazione per scopi statistici di dati depositati in archivi pubblici e trattati da enti pubblici o sulla base di finanziamenti pubblici.***

### Capo I - AMBITO DI APPLICAZIONE E PRINCIPI GENERALI

#### Art.1
#### Ambito di applicazione

**1.** Il codice si applica ai trattamenti di dati personali per scopi statistici effettuati da:
a) enti ed uffici di statistica che fanno parte o partecipano al sistema statistico nazionale, per l'attuazione del programma statistico nazionale o per la produzione di informazione statistica, in conformità ai rispettivi ambiti istituzionali;

b) strutture diverse dagli uffici di cui alla lettera *a)*, ma appartenenti alla medesima amministrazione o ente, qualora i relativi trattamenti siano previsti dal programma statistico nazionale e gli uffici di statistica attestino le metodologie adottate, osservando le disposizioni contenute nei decreti legislativi 6 settembre 1989, n. 322 e 30 luglio 1999, n. 281, e loro successive modificazioni e integrazioni, nonché nel presente codice.

**Art.2**
**Definizioni**

**1.** Ai fini del presente codice si applicano le definizioni elencate nell'art.1 della legge 31 dicembre 1996, n. 675 (di seguito denominata "Legge"), nel decreto legislativo 30 luglio 1999, n. 281, e loro successive modificazioni e integrazioni. Ai fini medesimi, si intende inoltre per:

a) "*trattamento per scopi statistici*", qualsiasi trattamento effettuato per finalità di indagine statistica o di produzione**,** conservazione e diffusione di risultati statistici in attuazione del programma statistico nazionale o per effettuare informazione statistica in conformità agli ambiti istituzionali dei soggetti di cui all'articolo 1;
b) "*risultato statistico*", l'informazione ottenuta con il trattamento di dati personali per quantificare aspetti di un fenomeno collettivo;
c) "*variabile pubblica*", il carattere o la combinazione di caratteri, di tipo qualitativo o quantitativo, oggetto di una rilevazione statistica che faccia riferimento ad informazioni presenti in pubblici registri, elenchi, atti, documenti o fonti conoscibili da chiunque;
d) "*unità statistica*", l'entità alla quale sono riferiti o riferibili i dati trattati.

**Art.3**
**Identificabilità dell'interessato**

**1.** Agli effetti dell'applicazione del presente codice:

a) un interessato si ritiene identificabile quando, con l'impiego di mezzi ragionevoli, è possibile stabilire un'associazione significativamente probabile tra la combinazione delle modalità delle variabili relative ad una unità statistica e i dati identificativi della medesima;
b) i mezzi ragionevolmente utilizzabili per identificare un interessato afferiscono, in particolare, alle seguenti categorie:
   - risorse economiche;
   - risorse di tempo;
   - archivi nominativi o altre fonti di informazione contenenti dati identificativi congiuntamente ad un sottoinsieme delle variabili oggetto di comunicazione o diffusione;
   - archivi, anche non nominativi, che forniscano ulteriori informazioni oltre a quelle oggetto di comunicazione o diffusione;
   - risorse *hardware* e *software* per effettuare le elaborazioni necessarie per collegare informazioni non nominative ad un soggetto identificato, tenendo anche conto delle effettive possibilità di pervenire in modo illecito alla sua identificazione in rapporto ai sistemi di sicurezza ed al *software* di controllo adottati;
   - conoscenza delle procedure di estrazione campionaria, imputazione, correzione e protezione statistica adottate per la produzione dei dati;
c) in caso di comunicazione e di diffusione, l'interessato può ritenersi non identificabile se il rischio di identificazione, in termini di probabilità di identificare l'interessato stesso tenendo conto dei dati comunicati o diffusi, è tale da far ritenere sproporzionati i mezzi eventualmente necessari per procedere all'identificazione rispetto alla lesione o al pericolo di lesione dei diritti degli interessati che può derivarne, avuto altresì riguardo al vantaggio che se ne può trarre.

**Art.4**
**Criteri per la valutazione del rischio di identificazione**

**1.** Ai fini della comunicazione e diffusione di risultati statistici, la valutazione del rischio di identificazione tiene conto dei seguenti criteri:

a) si considerano dati aggregati le combinazioni di modalità alle quali è associata una frequenza non inferiore a una soglia prestabilita, ovvero un'intensità data dalla sintesi dei valori assunti da un numero di unità statistiche pari alla suddetta soglia. Il valore minimo attribuibile alla soglia è pari a tre;
b) nel valutare il valore della soglia si deve tenere conto del livello di riservatezza delle informazioni;
c) i risultati statistici relativi a sole variabili pubbliche non sono soggetti alla regola della soglia;
d) la regola della soglia può non essere osservata qualora il risultato statistico non consenta ragionevolmente l'identificazione di unità statistiche, avuto riguardo al tipo di rilevazione e alla natura delle variabili associate;
e) i risultati statistici relativi a una stessa popolazione possono essere diffusi in modo che non siano possibili collegamenti tra loro o con altre fonti note di informazione, che rendano possibili eventuali identificazioni;
f) si presume che sia adeguatamente tutelata la riservatezza nel caso in cui tutte le unità statistiche di una popolazione presentino la medesima modalità di una variabile.

**2.** Nel programma statistico nazionale sono individuate le variabili che possono essere diffuse in forma disaggregata, ove ciò risulti necessario per soddisfare particolari esigenze conoscitive anche di carattere internazionale o comunitario.

**3.** Nella comunicazione di collezioni campionarie di dati, il rischio di identificazione deve essere per quanto possibile contenuto. Tale limite e la metodologia per la stima del rischio di identificazione sono individuati dall'Istat che, attenendosi ai criteri di cui all'art.3, comma 1, lett. *d*), definisce anche le modalità di rilascio dei dati dandone comunicazione alla Commissione per la garanzia dell'informazione statistica.

## Art.5
## Trattamento di dati sensibili da parte di soggetti privati

**1.** I soggetti privati che partecipano al sistema statistico nazionale ai sensi della legge 28 aprile 1998, n. 125, raccolgono o trattano ulteriormente dati sensibili per scopi statistici di regola in forma anonima, fermo restando quanto previsto dall'art.6-*bis*, comma 1, del decreto legislativo 6 settembre 1989, n. 322, come introdotto dal decreto legislativo 30 luglio 1999, n. 281, e successive modificazioni e integrazioni.

**2.** In casi particolari in cui scopi statistici, legittimi e specifici, del trattamento di dati sensibili non possono essere raggiunti senza l'identificazione anche temporanea degli interessati, per garantire la legittimità del trattamento medesimo è necessario che concorrano i seguenti presupposti:

a) l'interessato abbia espresso liberamente il proprio consenso sulla base degli elementi previsti per l'informativa;
b) il titolare adotti specifiche misure per mantenere separati i dati identificativi già al momento della raccolta, salvo che ciò risulti irragionevole o richieda uno sforzo manifestamente sproporzionato;
c) il trattamento risulti preventivamente autorizzato dal Garante, anche sulla base di un'autorizzazione relativa a categorie di dati o tipologie di trattamenti, o sia compreso nel programma statistico nazionale.

**3.** Il consenso è manifestato per iscritto. Qualora la raccolta dei dati sensibili sia effettuata con particolari modalità quali interviste telefoniche o assistite da elaboratore che rendano particolarmente gravoso per l'indagine acquisirlo per iscritto, il consenso, purché espresso, può essere documentato per iscritto. In tal caso, la documentazione dell'informativa resa all'interessato e dell'acquisizione del relativo consenso è conservata dal titolare del trattamento per tre anni.

## Capo II - INFORMATIVA, COMUNICAZIONE E DIFFUSIONE

### Art.6
### Informativa

**1.** Oltre alle informazioni di cui all'art.10 della Legge, all'interessato o alle persone presso le quali i dati personali dell'interessato sono raccolti per uno scopo statistico è rappresentata l'eventualità che essi possono essere trattati per altri scopi statistici, in conformità a quanto previsto dai decreti legislativi 6 settembre 1989, n. 322 e 30 luglio 1999, n. 281, e loro successive modificazioni e integrazioni.

**2.** Quando il trattamento riguarda dati personali non raccolti presso l'interessato e il conferimento dell'informativa a quest'ultimo richieda uno sforzo sproporzionato rispetto al diritto tutelato, in base a quanto previsto dall'art.10, comma 4 della Legge, l'informativa stessa si considera resa se il trattamento è incluso nel programma statistico nazionale o è oggetto di pubblicità con idonee modalità da comunicare preventivamente al Garante il quale può prescrivere eventuali misure ed accorgimenti.

**3.** Nella raccolta di dati per uno scopo statistico, l'informativa alla persona presso la quale i dati sono raccolti può essere differita per la parte riguardante le specifiche finalità, le modalità del trattamento cui sono destinati i dati, qualora ciò risulti necessario per il raggiungimento dell'obiettivo dell'indagine –in relazione all'argomento o alla natura della stessa– e purché il trattamento non riguardi dati sensibili. In tali casi, il completamento dell'informativa deve essere fornito all'interessato non appena vengano a cessare i motivi che ne avevano ritardato la comunicazione, a meno che ciò comporti un impiego di mezzi palesemente sproporzionato. Il soggetto responsabile della ricerca deve redigere un documento -successivamente conservato per almeno due anni dalla conclusione della ricerca e reso disponibile a tutti i soggetti che esercitano i diritti di cui all'art.13 della Legge- in cui siano indicate le specifiche motivazioni per le quali si è ritenuto di differire l'informativa, la parte di informativa differita, nonché le modalità seguite per informare gli interessati quando sono venute meno le ragioni che avevano giustificato il differimento.

**4.** Quando le circostanze della raccolta e gli obiettivi dell'indagine sono tali da consentire ad un soggetto di rispondere in nome e per conto di un altro, in quanto familiare o convivente, l'informativa all'interessato può essere data anche per il tramite del soggetto rispondente.

### Art.7
### Comunicazione a soggetti non facenti parte del sistema statistico nazionale

**1.** Ai soggetti che non fanno parte del sistema statistico nazionale possono essere comunicati, sotto forma di collezioni campionarie, dati individuali privi di ogni riferimento che ne permetta il collegamento con gli interessati e comunque secondo modalità che rendano questi ultimi non identificabili.

**2.** La comunicazione di dati personali a ricercatori di università o ad istituti o enti di ricerca o a soci di società scientifiche a cui si applica il codice di deontologia e di buona condotta per i trattamenti di dati personali per scopi statistici e di ricerca scientifica effettuati fuori dal sistema statistico nazionale, di cui all'articolo 10, comma 6, del decreto legislativo 30 luglio 1999, n. 281 e successive modificazioni e integrazioni, è consentita nell'ambito di specifici laboratori costituiti da soggetti del sistema statistico nazionale, a condizione che:

a) i dati siano il risultato di trattamenti di cui i medesimi soggetti del sistema statistico nazionale siano titolari;
b) i dati comunicati siano privi di dati identificativi;
c) le norme in materia di segreto statistico e di protezione dei dati personali, contenute anche nel presente codice, siano rispettate dai ricercatori che accedono al laboratorio anche sulla base di una preventiva dichiarazione di impegno;
d) l'accesso al laboratorio sia controllato e vigilato;
e) non sia consentito l'accesso ad archivi di dati diversi da quello oggetto della comunicazione;
f) siano adottate misure idonee affinché le operazioni di immissione e prelievo di dati siano inibite ai ricercatori che utilizzano il laboratorio;

g) il rilascio dei risultati delle elaborazioni effettuate dai ricercatori che utilizzano il laboratorio sia autorizzato solo dopo una preventiva verifica, da parte degli addetti al laboratorio stesso, del rispetto delle norme di cui alla lettera *c*).

**3.** Nell'ambito di progetti congiunti, finalizzati anche al perseguimento di compiti istituzionali del titolare del trattamento che ha originato i dati, i soggetti del sistema statistico nazionale possono comunicare dati personali a ricercatori operanti per conto di università, altre istituzioni pubbliche e organismi aventi finalità di ricerca, purché sia garantito il rispetto delle condizioni seguenti:

a) i dati siano il risultato di trattamenti di cui i medesimi soggetti del sistema statistico nazionale sono titolari;
b) i dati comunicati siano privi di dati identificativi;
c) la comunicazione avvenga sulla base di appositi protocolli di ricerca sottoscritti da tutti i ricercatori che partecipano al progetto;
d) nei medesimi protocolli siano esplicitamente previste, come vincolanti per tutti i ricercatori che partecipano al progetto, le norme in materia di segreto statistico e di protezione dei dati personali contenute anche nel presente codice.

**4.** È vietato ai ricercatori ammessi alla comunicazione dei dati di effettuare trattamenti per fini diversi da quelli esplicitamente previsti dal protocollo di ricerca, di conservare i dati comunicati oltre i termini di durata del progetto, di comunicare ulteriormente i dati a terzi.

### Art.8
### Comunicazione dei dati tra soggetti del sistema statistico nazionale

**1.** La comunicazione di dati personali, privi di dati identificativi, tra i soggetti del sistema statistico nazionale è consentita per i trattamenti statistici, strumentali al perseguimento delle finalità istituzionali del soggetto richiedente, espressamente determinati all'atto della richiesta, fermo restando il rispetto dei principi di pertinenza e di non eccedenza.

**2.** La comunicazione anche dei dati identificativi di unità statistiche tra i soggetti del sistema statistico nazionale è consentita, previa motivata richiesta in cui siano esplicitate le finalità perseguite ai sensi del decreto legislativo 6 settembre 1989, n. 322, ivi comprese le finalità di ricerca scientifica per gli enti di cui all'art.2 del decreto legislativo medesimo, qualora il richiedente dichiari che non sia possibile conseguire altrimenti il medesimo risultato statistico e, comunque, nel rispetto dei principi di pertinenza e di stretta necessità.

**3.** I dati comunicati ai sensi dei commi 1 e 2 possono essere trattati dal soggetto richiedente, anche successivamente, per le sole finalità perseguite ai sensi del decreto legislativo 6 settembre 1989, n. 322, ivi comprese le finalità di ricerca scientifica per gli enti di cui all'art.2 del decreto legislativo medesimo, nei limiti previsti dal decreto legislativo 30 luglio 1999, n. 281, e nel rispetto delle misure di sicurezza previste dall'art.15 della Legge e successive modificazioni e integrazioni.

### Art.9
### Autorità di controllo

**1.** La Commissione per la garanzia dell'informazione statistica di cui all'articolo 12 del decreto legislativo 6 settembre 1989, n. 322 contribuisce alla corretta applicazione delle disposizioni del presente codice e, in particolare, di quanto previsto al precedente art.8, segnalando al Garante i casi di inosservanza.

## Capo III - SICUREZZA E REGOLE DI CONDOTTA

### Art.10
### Raccolta dei dati

**1.** I soggetti di cui all'art.1 pongono specifica attenzione nella selezione del personale incaricato della raccolta dei dati e nella definizione dell'organizzazione e delle modalità di rilevazione, in modo da garantire il rispetto del presente codice e la tutela dei diritti degli interessati, procedendo altresì alla designazione degli incaricati del trattamento, secondo le modalità di legge.

**2.** In ogni caso, il personale incaricato della raccolta si attiene alle disposizioni contenute nel presente codice e alle istruzioni ricevute. In particolare:

a) rende nota la propria identità, la propria funzione e le finalità della raccolta, anche attraverso adeguata documentazione;
b) fornisce le informazioni di cui all'art.10 della Legge e di cui all'art.6 del presente codice, nonché ogni altro chiarimento che consenta all'interessato di rispondere in modo adeguato e consapevole, evitando comportamenti che possano configurarsi come artifici o indebite pressioni;
c) non svolge contestualmente presso gli stessi interessati attività di rilevazione di dati per conto di più titolari, salvo espressa autorizzazione;
d) provvede tempestivamente alla correzione degli errori e delle inesattezze delle informazioni acquisite nel corso della raccolta;
e) assicura una particolare diligenza nella raccolta di dati personali di cui agli articoli 22, 24 e 24 bis della Legge.

**Art.11**
**Conservazione dei dati**

**1.** I dati personali possono essere conservati anche oltre il periodo necessario per il raggiungimento degli scopi per i quali sono stati raccolti o successivamente trattati, in conformità all'art.9 della Legge e all'art.6-*bis* del decreto legislativo 6 settembre 1989, n. 322 e successive modificazioni e integrazioni. In tali casi, i dati identificativi possono essere conservati fino a quando risultino necessari per:

- indagini continue e longitudinali;
- indagini di controllo, di qualità e di copertura;
- definizione di disegni campionari e selezione di unità di rilevazione;
- costituzione di archivi delle unità statistiche e di sistemi informativi;
- altri casi in cui ciò risulti essenziale e adeguatamente documentato per le finalità perseguite.

**2.** Nei casi di cui al comma 1, i dati identificativi sono conservati separatamente da ogni altro dato, in modo da consentirne differenti livelli di accesso, salvo che ciò risulti impossibile in ragione delle particolari caratteristiche del trattamento o comporti un impiego di mezzi manifestamente sproporzionati rispetto al diritto tutelato.

**Art.12**
**Misure di sicurezza**

**1.** Nell'adottare le misure di sicurezza di cui all'art.15, comma 1, della Legge e di cui al regolamento previsto dal comma 2 del medesimo articolo, il titolare del trattamento determina anche i differenti livelli di accesso ai dati personali con riferimento alla natura dei dati stessi e alle funzioni dei soggetti coinvolti nei trattamenti.

**2.** I soggetti di cui all'art.1 adottano le cautele previste dagli articoli 3 e 4 del decreto legislativo 11 maggio 1999, n. 135 in riferimento ai dati di cui agli articoli 22 e 24 della Legge.

**Art.13**
**Esercizio dei diritti dell'interessato**

**1.** In caso di esercizio dei diritti di cui all'art.13 della Legge, l'interessato può accedere agli archivi statistici contenenti i dati che lo riguardano per chiederne l'aggiornamento, la rettifica o l'integrazione, sempre che tale operazione non risulti impossibile per la natura o lo stato del trattamento, o comporti un impiego di mezzi manifestamente sproporzionati.

**2.** In attuazione dell'art.6-*bis*, comma 8, del decreto legislativo 6 settembre 1989, n. 322, il responsabile del trattamento annota in appositi spazi o registri le modifiche richieste dall'interessato, senza variare i dati originariamente immessi nell'archivio, qualora tali operazioni non producano effetti significativi sull'analisi statistica o sui risultati statistici connessi al trattamento. In particolare, non si procede alla variazione se le modifiche richieste contrastano con le classificazioni e con le metodologie statistiche adottate in conformità alle norme internazionali comunitarie e nazionali.

**Art.14**
**Regole di condotta**

**1.** I responsabili e gli incaricati del trattamento che, anche per motivi di lavoro, studio e ricerca abbiano legittimo accesso ai dati personali trattati per scopi statistici, conformano il proprio comportamento anche alle seguenti disposizioni:

a) i dati personali possono essere utilizzati soltanto per gli scopi definiti all'atto della progettazione del trattamento;
b) i dati personali devono essere conservati in modo da evitarne la dispersione, la sottrazione e ogni altro uso non conforme alla legge e alle istruzioni ricevute;
c) i dati personali e le notizie non disponibili al pubblico di cui si venga a conoscenza in occasione dello svolgimento dell'attività statistica o di attività ad essa strumentali non possono essere diffusi, né altrimenti utilizzati per interessi privati, propri o altrui;
d) il lavoro svolto deve essere oggetto di adeguata documentazione;
e) le conoscenze professionali in materia di protezione dei dati personali devono essere adeguate costantemente all'evoluzione delle metodologie e delle tecniche;
f) la comunicazione e la diffusione dei risultati statistici devono essere favorite, in relazione alle esigenze conoscitive degli utenti, purché nel rispetto delle norme sulla protezione dei dati personali.

**2.** I responsabili e gli incaricati del trattamento di cui al comma 1 sono tenuti a conformarsi alle disposizioni del presente codice, anche quando non siano vincolati al rispetto del segreto d'ufficio o del segreto professionale. I titolari del trattamento adottano le misure opportune per garantire la conoscenza di tali disposizioni da parte dei responsabili e degli incaricati medesimi.

**3.** I comportamenti non conformi alle regole di condotta dettate dal presente codice devono essere immediatamente segnalati al responsabile o al titolare del trattamento.

# Riferimenti bibliografici


Abowd, J.M. e Woodcock, S.D. (2001). Microdata protection through noise addition. Doyle, P., Lane, J.I., Theeuwes, J.J.M. e Zayatz, L. (Eds), *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies*. Amsterdam: Elsevier Science, 215–277.

Adam, N.R. e Wortmann, J.C. (1989). Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys*, 21, 515-556.

Anwar, M.N. (1993). Microaggregation: the small aggregates method. *Internal Report*, Eurostat.

Benedetti, R. e Franconi, L. (1998). Statistical and technological solutions for controlled data dissemination. *Pre-proceedings del convegno NTTS '98, New Techniques and Technologies for Statistics*, Sorrento, 1, 225-232.

Benedetti, R., Capobianchi, A. e Franconi, L. (2003). An estimation method for individual risk of disclosure based on sampling design. *Contributi Istat*, 11/2003.

Bethlehem, J. C., Keller, W. J. e Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association,* 85, 409, 38-45.

Biggeri, L. e Zannella, F. (1991). Release of microdata and statistical disclosure control in the new national system of Italy: main problems, some technical solutions, experiments. *Bullettin of the International Statistical Institute, Proceedings*, Tome LIV, Book 1, 1-25.

Bishop, Y.M.M, Fienberg, S.E. e Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: MIT Press.

Blakemore, M. (2001). The potential and perils of remote access. Doyle, P., Lane, J.I., Theeuwes, J.J.M. e Zayatz, L. (Eds), *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies*. Amsterdam: Elsevier Science, 315-337.

Brand, R. (2002). Microdata protection through noise addition. In Domingo-Ferrer, J. (Ed.), *Inference Control in Statistical Databases*: *From Theory to Practice*. Lecture Notes in Computer Science, 2316, Berlin: Springer-Verlag, 97-116.

Burridge, J. (2003). Information preserving statistical obfuscation. *Statistics and Computing,* 13. 321-327.

Buzzigoli, L. e Giusti, A. (1999). An algorithm to calculate the lower and upper bounds of the elements of an array given its marginals. In *Statistical Data Protection, Proceedings of the Conference*, Luxembourg, 131-147.

Carlson, M. e Salabasis, M. (1998). A data-swapping technique for generating synthetic samples: a method for disclosure control *Pre-proceedings del convegno NTTS '98, New Techniques and Technologies for Statistics*, Sorrento, 2, 109-114.

Carvalho, F.D., Dellaert, N.P. e Osòrio, M.S. (1994). Statistical disclosure in two-dimensional tables: general tables. *Journal of the American Statistical Association*, 89, 1547-1557.

Cicalini, L. (2001). Principi giuridici e procedure per la diffusione delle statistiche del commercio con l'estero. *Istat, documento interno*.

Coccia, G. (1993). Disclosure risk in Italian current population surveys. In *International Seminar on Statistical Confidentiality*, *Proceeedings*, Dublin, 8-10 September 1992. Luxembourg: Office for the Official Publications of the European Communities, 415-423.

Contini, B. e Monducci, R. (1996). Analisi microeconomica e indagini sulle imprese: bisogni informativi e proposte. *Società Italiana di Statistica. Atti della XXXVIII Riunione scientifica,* 1, 351-362.

Contini, B., Corsini, V., Franconi, L., Pagliuca, D., Papa, P., Piersimoni, F., Seri, G., Siesto, G. e Taccini, P. (1998). Metodi di microaggregazione per il rilascio di dati di impresa. *Documenti Istat*, n. 17/1998.

Corsini, V., Franconi, L., Pagliuca, D. e Seri, G. (1999). An application of microaggregation methods to Italian business survey. In *Statistical Data Protection, Proceedings of the Conference.* Lisbon, Luxembourg: Eurostat, 109-113.

Cox, L.H. (1981). Linear sensitivity measures in statistical disclosure control. *Journal of Statistical Planning and Inference*, 5, 153-164.

Cox, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 398, 520-524.

Cox, L.H. (1994). Matrix masking methods for disclosure limitation in microdata. *Survey Methodology*, 20, 165-169.

Cox, L.H. (1995a). Network models for complementary cell suppression. *Journal of the American Statistical Association*, 75, 377-385.

Cox, L.H. (1995b). Protecting confidentiality in business surveys. In *Business Survey Methods*, Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. e Kott, P.S. (Eds.), New-York: Wiley, 443-476.

Cox, L.H. (2000). Towards a bayesian perspective on statistical disclosure limitation. presentato a *ISBA 2000* - George, E.I. (Ed.), *The Sixth World Meeting of the International Society for Bayesian Analysis,* 91-98.

Cox, L.H. (2001). Bounding entries in 3-dimensional contigency tables. In *Proceedings of the conference Eustat 2001,* Lisbon.

Crescenzi, F. (1993). Estimating population uniques: methodological proposals and applications on Italian census data. In *International Seminar on Statistical Confidentiality*, *Proceeedings*, Dublin, 8-10 September 1992. Luxembourg: Office for the Official Publications of the European Communities, 247-260.

Cuppen, M. e Willenborg, L. (2003). Source data perturbation and consistent sets of safe tables. *Statistics and Computing,* 13, 355-362.

Dalenius, T. e Reiss, S.P. (1982). Data-swapping: a technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73-85.

Dandekar, R., Cohen, M. e Kirkendall, N. (2001). Applicability of Latin Hypercube Sampling to create multi variate synthetic micro data. In *ETK-NTTS 2001 Pre-proceedings of the Conference*. European Communities, Luxembourg, 839-847.

Dandekar, R., Cohen, M. e Kirkendall, N. (2002a). Sensitive micro data protection using Latin Hypercube Sampling technique. In Domingo-Ferrer, J. (Ed.), *Inference Control in Statistical Databases*: *From Theory to Practice*. Lecture Notes in Computer Science, 2316, Berlin: Springer-Verlag, 117-225.

Dandekar, R., Domingo-Ferrer, J. e Sebé, F. (2002b). LHS based hybrid microdata vs rank swapping and micoraggregation for numeric microdata protection. In Domingo-Ferrer, J. (Ed.), *Inference Control in Statistical Databases*: *From Theory to Practice*. Lecture Notes in Computer Science, 2316, Berlin: Springer-Verlag, 153-162.

David, M.H. (1998). Killing with kindness: the attack on public use data. *Proceedings of the Section on Government Statistics*, *American Statistical Association,* 3-7.

de Waal, A.G. e Willenborg, L.C.R.J. (1996). A view on statistical disclosure control for microdata. *Survey Methodology*, 22, 95-103.

de Waal, A.G. e Willenborg, L.C.R.J. (1998). Optimal local suppression in microdata. *Journal of Official Statistics*, 14, 421-435.

de Wolf, P. P. (2001). Notes for the TES course on SDC held in Sept. 2001 at CBS, Voorburg (NL).

Defays, D. e Anwar, M.N. (1998). Masking microdata using micro-aggregation. *Journal of Official Statistics*, 14, 449-461.

Defays, D. e Nanopoulos, P. (1992). Panels of enterprises and confidentiality: the small aggregates method. *Proceedings of Statistics Canada Symposium 92, Design and Analysis of Longitudinal Surveys, November 1992,* 195-204.

Dempster, A.P., Laird, N.M. e Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 40, 1-38.

Diaconis, P. e Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics,* 26, 363-397.

Dobra, A. (2000). Measuring the disclosure risk in multiway tables with fixed marginals corresponding to decomposable loglinear models. *Technical Report,* Department of Statistics, Carnegy Mellon University.

Dobra, A. (2001). Computing sharp integer bounds for entries in contigency tables given a set of fixed marginals. *Technical Report,* Department of Statistics, Carnegy Mellon University.

Dobra, A. e Fienberg, S.E. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Inaugural Article, PNAS 97:* 11885-11892. Reperibile anche al sito http://www.pnas.org/.

Dobra, A. e Fienberg, S.E. (2001). Bounds for cell entries in contingency tables induced by fixed marginal totals with applications to disclosure limitation. *Statistical Journal of the United Nations ECE,* 18, 363-371. Disponibile su http://www.niss.org/dg/technicalreports.html.

Dobra, A., Karr, A.F., Sanil, A.P. e Fienberg, S.E. (2002). Software systems for tabular data releases. *International Journal on Uncertainty Fuzziness and Knowledge-Based Systems*, 10, 5, 529-544. Disponibile su http://www.niss.org/dg/technicalreports.html.

Domingo-Ferrer, J. (1998). Pros and cons of new information technologies for statistical data protection. *Pre-proceedings* del convegno *New Techniques and Technologies for Statistics*, Napoli, 1, 233-240.

Domingo-Ferrer, J., e Mateo-Sanz, J.M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering,* 14, 189-201.

Doyle, P., Lane, J.I., Theeuwes, J.J.M. e Zayatz, L.V. (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies.* Amsterdam: North-Holland.

Duncan, G.T., Keller-McNulty, S.A. e Stokes, S.L. (2001). Disclosure risk vs. data utility: the R-U confidentiality map. *Technical Report LA-UR-01-6428*, Los Alamos National Laboratory.

Duncan, G.T. e Lambert, D. (1986). Disclosure limited data dissemination, (with discussion), *Journal of the American Statistical Association*, 81, 10-28.

Duncan, G. e Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics.*

Duncan, G.T. e Mukherjee, S. (2000). Optimal disclosure limitation strategy in statistical databases: deterring tracker attacks through additive noise. *Journal of the American Statistical Association,* 95, 720-729.

Duncan, G.T. e Pearson, R.W. (1991). Enhancing access to microdata while protecting confidentiality: prospects for the future. *Statistical Science*, 6, 219-239.

Evans, B.T., Zayatz, L. e Slanta, J. (1998). Using noise for disclosure limitation of establishment tabular data. *Journal of Official Statistics*, 14, 537-551.

Fazio, N. e Menghinello, S. (2001). Aspetti statistici e informatici connessi alla realizzazione della banca dati *on line* sulle statistiche del commercio con l'estero. *Istat, documento interno*.

Fellegi, I.P. (1975). Controlled random rounding. *Survey Methodology*, 1, 123-133.

Fienberg, S.E. (1999). Fréchet and Bonferroni bounds for multi-way tables of counts with applications to disclosure limitations. In *Statistical Data Protection, Proceedings of the Conference*. Lisbon, Luxembourg: Eurostat, 115-129.

Fienberg, S.E. (2000). Confidentiality and data protection through disclosure limitation: evolving principles and technical advances. *The Philippine Statistician* 49, 1-12. Reperibile al sito http://www.niss.org/dg/technicalreports.html.

Fienberg, S.E. e Makov, U.E. (1998). Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics*, 14, 385-397.

Fienberg, S.E., Makov, U.E. e Steele, R.J., (1998). Disclosure limitation using perturbation and related methods for categorical data (with discussion). *Journal of Official Statistics*, 14, 485-502.

Fischetti, M. e Salazar-Gonzales, J.J. (1999). Models and algorithms for the 2-dimensional cell suppression problem in statistical disclosure control. *Mathematical Programming*, 84, 283-312.

Fischetti, M. e Salazar-Gonzales, J.J. (2000). Models and algorithms for optimizing cell suppression in tabular data with linear constraints. *Journal of the American Statistical Association,* 95, 916-928.

Fischetti, M. e Salazar-Gonzales, J.J. (2003). Partial cell suppression: a new methodology for statistical disclosure control. *Statistics and Computing,* 13, 13-21.

Franconi, L. (1999). La diffusione dei censimenti del 2000: nuove metodologie e tecniche per la tutela della riservatezza. *Relazione invitata alla XLI Riunione Scientifica della Società Italiana di Statistica*, Udine, 295-310.

Franconi, L. e Merola, G. (2003). Strategies for the application of statistical disclosure control methods in Web-based systems for data dissemination. *Accettato per la pubblicazione su La rivista di Statistica Ufficiale*.

Franconi, L. e Stander, J. (2000). Model based disclosure limitation for business microdata. In *Proceedings of the International Conference on Establishment Surveys-II, June 17-21, 2000. Buffalo, New York*, 887-896.

Franconi, L. e Stander, J. (2002). A model based method for disclosure limitation of business microdata. *Journal of the Royal Statistical Society, D*, 51, 1-11.

Franconi, L. e Stander, J. (2003). Spatial and non-spatial model-based protection procedures for the release of business microdata. *Statistics and Computing,* 13, 295-305.

Fries, G., Johnson, B. and Woodburn, R. L. (1997). Analyzing the disclosure review procedures for the 1995 Survey of Consumer Finances. *ASA Proceedings of the Social Statistics Section*, 311-316, American Statistical Association (Alexandria, VA).

Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9, 383-406.

Galambos, J. e Simonelli, I. (1996). *Bonferroni-type Inequalities with Applications.* New York: Springer-Verlag.

Gomatam, S., Karr, A.F. e Sanil, A. (2003). A risk-utility framework for categorical data swapping. *NISS Technical Report* n.132, February, 2003.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. e de Wolf, P.-P. (1998). Post randomisation for statistical disclosure control: theory and implementation, Journal of Official Statistics, 14, 463-478.

Grim, J., Boček, P. e Pudil, P. (2001). Safe dissemination of census results by means of interactive probabilistic models. In *ETK-NTTS 2001 Pre-proceedings of the Conference*. Luxembourg: European Communities, 849-856.

Hurkens, C. A. J. e Tiourine, S. R. (1998). Models and Methods for the Microdata Protection Problem. *Journal of Official Statistics*, 14, 437--447

Istat (1998). Conti economici delle imprese 1995. *Informazioni n. 102.*

Istat (2001). Microaggregazione dei dati economici strutturali delle imprese industriali e dei servizi - Anni 1995-1996. *Informazioni n. 34.*

Karr, A.F., Lee, J., Sanil, A.P., Hernandez, J., Karimi, S. e Litwin, K. (2002). Web-based systems that disseminate information from data but protect confidentiality. Apparirà in *Advances in Digital Government*, McIver W.J. e Elmagarmid A.W. (eds). Kluwer Academic Publishers. Disponibile su http://www.niss.org/dg/technicalreports.html.

Keller-McNulty, S. e Unger, E.A. (1998). A database system prototype for remote access to information based on confidential data. *Journal of Official Statistics*, 14, 347-360.

Kelly, J.P. (1990). Confidentiality protection in two and three-dimensional tables. Tesi di Dottorato, University of Maryland, College Park, Maryland.

Kelly, J.P., Golden, B.L. e Assad, A.A. (1992). Cell suppression: disclosure protection for sensitive tabular data. *Networks*, 22, 397-417.

Kennickell, A.B. (1999). Multiple imputation and disclosure protection. . In *Statistical Data Protection, Proceedings of the Conference*. Lisbon, Luxembourg: Eurostat, 381-400.

Kim, J. (1986). A method for limiting disclosure of microdata based on random noise and transformation. In *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 370-374.

Kooiman, P. (1998). Discussion of "Disclosure limitation using perturbation and related methods for categorical data". *Journal of Official Statistics*, 14, 503-508.

Little, R.J.A. e Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Malvestuto, F.M. e Moscarini, M. (1999). An audit expert for large statistical databases. *Statistical Data Protection Proceedings of the Conference,* European Commission, Theme 9: Research and development, Collection: Studies and research, 29-43.

Merola, G. (2003a). Safety rules in statistical disclosure control for tabular data. *Contributi Istat*, 1/2003.

Merola, G. (2003b). Safety rules in statistical disclosure control for tabular data. *Presentato per la pubblicazione*.

Mokken, R.J., Kooiman, P., Pannekoek, J. e Willenborg, L.C.R.J. (1992). Disclosure risks for microdata. *Statistica Neerlandica*, 46, 49-67.

Paass, G. (1988). Disclosure Risk and Disclosure Avoidance for Microdata, *Journal of Business and Economic Statistics*, Vol. 6, 487-500.

Pagliuca, D. e Seri, G. (1999a). M.A.S.Q. Manuale utente. *Documento interno Istat*.

Pagliuca, D. e Seri, G. (1999b). Some results of individual ranking method on the System of Enterprise Accounts Annual Survey. *Esprit SDC Project, Deliverable MI-3/D2*.

Polettini, S. Franconi, L. e Stander, J. (2002). Model Based Disclosure Protection. In Domingo-Ferrer, J. (Ed.), *Inference Control in Statistical Databases*: *From Theory to Practice*. Lecture Notes in Computer Science, 2316, Berlin: Springer-Verlag, 83-96

Polettini, S. (2003). Maximum entropy simulation for microdata protection. *Statistics and Computing,* 13. 307-320.

Reiter, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18, 531-543.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata. *Survey Methodology*, in corso di pubblicazione.

Raghunathan, T.E., Reiter, J.P. e Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation, *Journal of Official Statistics*, 19, 1-16.

Raghunathan, T.E. e Rubin, D.B. (2000). Bayesian multiple imputation to preserve confidentiality in public-use data sets. Presentato a *ISBA 2000* - George, E.I. (Ed.), *The Sixth World Meeting of the International Society for Bayesian Analysis*.

Ribeiro Filho, J.L. e Treleaven, P.C. (1994). Genetic algorithm programming environments. *IEEE Computer*, 27, 28-43.

Rubin, D.B. (1993). Discussion of "Statistical disclosure limitation". *Journal of Official Statistics*, 9, 461-468.

Sande, G. (1984). Automated cell suppression to preserve confidentiality of business statistics. *Statistical Journal of the United Nations,* ECE 2, 33-41.

Schouten, B. e Cigran, M. (2003). Remote access systems for statistical analysis of microdata. *Statistics and Computing,* 13, 381-389.

Shackis, D. (1993). Manual on disclosure control methods. *Report,* Luxembourg: Eurostat.

Skinner, C.J. e Holmes, D.J. (1993). Modelling population uniqueness. In *International Seminar on Statistical Confidentiality*, *Proceeedings*, Dublin, 8-10 September 1992. Luxembourg: Office for the Official Publications of the European Communities, 175-199.

Skinner, C.J. e Holmes, D.J. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14, 361-372.

Skinner, C.J., Marsh, C., Openshaw, S. e Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics*, 10, 31-51.

Sullivan, G.R. (1989). The use of added error to avoid disclosure in microdata releases. Tesi di dottorato non pubblicata, Iowa State University.

Sullivan, G.R. e Fuller, W.A. (1989). The use of added error to avoid disclosure in microdata releases. In *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 602-610.

Trottini, M. (2001). A decision-theoretic approach to data disclosure problems. *Research in Official Statistics* 4**,** 7-22. Disponibile su http://www.niss.org/dg/technicalreports.html.

Trottini, M. e Fienberg, S.E. (2002). Modelling user uncertainty for disclosure risk and data utility. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, 10, 511-527. Disponibile su http://www.stat.cmu.edu/~fienberg/DLindex.html.

Ward, J.K. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58**,** 236-244.

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. (2001). Elements of statistical disclosure control. Lecture Notes in Statistics, 115, New York: Springer-Verlag.

Winkler, W.E., Yancey, W.E. e Creecy, R.H. (2002). Disclosure risk assessment in perturbative microdata protection via record linkage. In Domingo-Ferrer, J. (Ed.), *Inference Control in Statistical Databases*: *From Theory to Practice*. Lecture Notes in Computer Science, 2316, Berlin: Springer-Verlag, 135-152.

Zayatz, L. (2002). SDC in the 2000 U.S. Decennial Census. In Domingo-Ferrer, J. (Ed.), *Inference Control in Statistical Databases*: *From Theory to Practice*. Lecture Notes in Computer Science, 2316, Berlin: Springer-Verlag, 193-202.

Zucchetti, A. e altri (2004). Codice della privacy (Commento al Decreto Legislativo 30 giugno 2003, n.196). Ed. Giuffrè, Collana: Le nuove leggi amministrative, ISBN 8814107688.