Documenti Istat

Review dei sistemi di accesso remoto: schematizzazione e analisi comparativa.

A. Capobianchi (*)

(*) ISTAT – Servizio Progettazione e supporto metodologico nei processi di produzione statistica

Sommario

La crescente richiesta da parte di ricercatori di utilizzare dati individuali per analisi sempre più specifiche, ha creato, per i vari Istituti di statistica, l'esigenza di sviluppare metodologie di accesso e tecniche di protezione che, da una parte soddisfino al meglio tali richieste e dall'altra garantiscano il rispetto del principio della tutela della riservatezza dei rispondenti.
Negli ultimi anni sempre più Istituti hanno implementato o stanno progettando l'implementazione del canale di "*accesso remoto"* come canale di accesso a microdati.
La diffusione di tale canale è strettamente legata alla sua duttilità di implementazione, ovvero alla facilità di adattarsi alle diverse esigenze che un Istituto o un'Agenzia di Statistica presenta in termini di prodotti da offrire ai propri utenti. Altri elementi che caratterizzano il crescente interesse a tale canale è lo sviluppo di tecnologie informatiche, che rendono sempre più sicure le comunicazioni via web, e la diffusa utilizzazione di internet.
Nel presente documento si descrivono le soluzioni adottate da alcuni paesi e si effettua un'analisi dettagliata delle fasi più significative e caratterizzanti del sistema stesso.

Indice

1. Introduzione

Per rispondere alle sempre più pressanti richieste della comunità scientifica di informazione statistica sotto forma di microdati, i vari Istituti ed Agenzie Nazionali di Statistica hanno sviluppato tre canali di accesso ai dati ognuno con delle peculiarità specifiche. Tali canali sono: rilascio di file di dati su CD, Accesso remoto e Laboratorio di dati elementari (Capobianchi, 2005). Ognuno di essi risponde a delle caratteristiche particolari della richiesta dati e nel loro insieme definiscono la maggior parte dell'offerta di microdati dell'Istituto a cui si riferiscono. In particolare l'accesso remoto può essere definito come:

"quel canale di accesso ai dati che permette ad utenti esterni di eseguire in remoto, ovvero attraverso una rete informatica, analisi statistiche su file di microdati senza aver un accesso diretto ad essi" (Trewin, 2006).

Si può notare come la definizione sia particolarmente generale e di conseguenza gli Istituti e le Agenzie Nazionali, adeguandoli a delle specifiche esigenze, hanno potuto diversificare i propri sistemi di accesso remoto agendo su diversi aspetti che coinvolgono la determinazione del sistema stesso (Rowland, 2003). Gli aspetti strutturali che possono permettere una diversificazione nell'implementazione dei vari sistemi possono riguardare:

- *I dati messi a disposizione*: i dati su cui fare le elaborazioni possono essere protetti con le usuali tecniche di protezione (data swapping, ricodifiche ecc) oppure possono essere resi disponibili file di dati cosiddetti "master" ovvero file risultanti dall'indagine in cui vengono eliminati esclusivamente le variabili identificative dirette (nome, cognome, indirizzo ecc);

- *Chi può accedere al servizio e con quale sistema di comunicazione con l'utente*: in alcuni casi l'accesso al servizio può essere ristretto ad una particolare tipologia di utenza; la comunicazione tra l'utente e il sistema di accesso remoto può avvenire sia attraverso il canale di posta elettronica che attraverso le rete internet;

- *I programmi statistici*: i programmi messi a disposizione dell'utente possono essere software statistici commerciali tra quelli maggiormente utilizzati nel campo della ricerca statistica (SAS, Stata Spss) o programmi costruiti appositamente per eseguire le richieste degli utenti;

- *I tempi di presa rilascio dell'output*: gli output risultanti dalle analisi condotte sui dati, ed in particolare quelli intermedi, possono essere visionati dall'utente in maniera quasi immediata sul proprio computer oppure possono essere inviati all'utente via mail in un momento successivo. Proprio questo aspetto caratterizza maggiormente i due tipi base di accesso remoto. Il primo prende il nome di "*Remote Facilities"* mentre il secondo, e più diffuso, prende il nome di "*Remote execution";*

- *Tipo di output ammesso*: alcuni sistemi prevedono il rilascio di output esclusivamente sotto forma di tabelle (*table server*), mentre altri prevedono anche il rilascio di risultati legati all'applicazione *di modelli* (*model server*);

- *Il tipo di controllo degli output*: i controlli per verificare se l'output di una determinata analisi può essere rilasciato, in quanto non ci sono violazioni della riservatezza dei rispondenti, possono essere fatti manualmente, da personale dell'Istituto, o in maniera automatica. Il controllo automatico avviene attraverso un sistema di interrogazioni che analizza sia le richieste degli utenti (ovvero i file contenenti i codici di programma- filtering software for set-up) che i risultati stessi delle elaborazioni.

Dall'analisi della definizione di accesso remoto si può notare come l'elemento caratterizzante tale canale sia legato al fatto che i dati sui quali si vuole condurre l'analisi statistica restino fisicamente all'interno dell'Istituto e non sia possibile né accedere direttamente ad essi né estrarre dei sottocampioni ma sia possibile ottenere esclusivamente i risultati (preferibilmente conclusivi) di un progetto statistico di ricerca applicato ai dati stessi.

Nel seguito analizzeremo le esperienze più indicative nel campo dell'accesso remoto relative sia ad alcuni paesi Europei che a paesi come gli Stati Uniti, Canada e l'Australia.

2. Danimarca (Statistics Denmark)

Nel 2000[1], l'Istituto nazionale di Statistica Danese, ha avviato un progetto per l'accesso remoto a dati elementari caratterizzato da un forte controllo sulle persone che accedono al servizio e sulla modalità di connessione.

Il canale di accesso remoto è rivolto a ricercatori appartenenti a selezionate Università, Enti pubblici, organizzazioni con scopi umanitari, organizzazioni non governative, società di consulenza o imprese che, in maniera stabile, presentano nella loro struttura una unità di ricerca (a queste ultime non è comunque concesso l'accesso a dati di impresa).
Non viene concessa l'autorizzazione a singoli individui e a ricercatori di enti non danesi. I ricercatori stranieri anche se non possono utilizzare il canale dell'accesso remoto, possono in ogni caso effettuare analisi su microdati prodotti dall'Istituto di Statistica Danese utilizzando il canale del Laboratorio di Dati elementari. Caso particolare è quello dei "visiting researcher" i quali possono utilizzare l'accesso remoto durante il loro periodo di permanenza in Danimarca utilizzando le strutture messe a disposizione dall'ente di ricerca che li ospita.

Per ottenere l'autorizzazione all'accesso ai dati si deve presentare un progetto di ricerca che viene valutato da una commissione appositamente preposta. Nel caso in cui il progetto viene accettato l'utente sottoscrive un contratto nel quale si impegna a rispettare le regole di comportamento predisposte dall'Istituto per garantire la riservatezza dei rispondenti. Vengono così assegnati uno o più computer, allestiti appositamente per l'analisi dei dati attraverso accesso remoto, presso l'ambiente di lavoro o di ricerca dell'utente autorizzato. Non vengono abilitati computer che si trovano in ambienti che non possono essere adeguatamente supervisionati (ad esempio quelli che si trovano in appartamenti privati).

L'intera gestione del sistema di Accesso Remoto è amministrata centralmente dall'unità "Division research service" (unità Drs). Il personale di tale unità ha accesso a tutti i dati prodotti dall'Istituto e attraverso una stretta collaborazione con le singole divisioni costruisce le collezioni di dati interdisciplinari richiesti dagli utenti.

I dati messi a disposizione sono forniti così come appaiono nei registri previa eliminazione delle variabili identificative dirette, ovvero non viene applicata nessuna tecnica di protezione (ricodifica, swapping ecc (Istat, 2004)). Le informazioni messe a disposizione soddisfano comunque il principio del "need to know" ovvero vengono messe a disposizione esclusivamente le informazioni necessarie in relazione alla descrizione del progetto di ricerca.
La soluzione tecnica è basata sull'utilizzo di internet. I dati, prodotti dall'Istituto e messi a disposizione del ricercatore che ne fa richiesta, vengono immagazzinati su di un disco collegato con

[1] http://www.dst.dk/HomeUK/ForSale/Research/acces.aspx, 2-2006
"New developments in the Danish system for access to microdata" Lars Borchsenius submitted Joint/Unece Eurostatwork session on statistical data confidentiality (Geneva 2005)

un server Unix appositamente predisposto. Tale server viene utilizzato esclusivamente dagli utenti e non è collegato con la rete di produzione dell'Istituto.
Le comunicazioni via internet sono criptate attraverso la cosiddetta "RSA SecurID card" ovvero una componente hardware che oltre a garantire un sistema di comunicazioni sicure via internet (attraverso un processo di criptazione delle informazioni), assicura che esclusivamente l'utente autorizzato dall'Istituto ottenga l'accesso al sistema. Il software Citrix assicura che il ricercatore dalla propria postazione di lavoro possa vedere l'ambiente Unix collocato all'interno dei locali dell'Istituto. L'intero processo di elaborazione avviene quindi all'interno dell'istituto Danese e i dati non possono essere trasferiti sul computer dell'utente. Il ricercatore può lavorare liberamente, l'unico limite è legato all'ampiezza dello spazio sul disco che negli anni è stato varie volte incrementato dall'Istituto proprio per migliorare la qualità del servizio offerto.
Tutti gli output che, adeguatamente aggregati, si vogliono salvare sul proprio computer, devono essere messi su un apposito file che successivamente viene inviato all'utente via e-mail. L'invio degli output è un processo continuo,che si ripete ogni cinque minuti, e tutte le e-mail inviate vengono immagazzinate per essere successivamente controllate dal personale dell'unità Drs. Nel caso in cui si presentino delle violazioni viene prontamente contattato l'utente.

Un sistema di accesso remoto del tutto analogo a quello appena descritto è stato implementato in Svezia (Söderberg L.G.,2005) ed è in funzione a partire dal 2005. Questo sistema prende il nome di MONA – Microdata ON-Line Access ed è un sistema basato su un sistema Citrix; le comunicazioni che avvengono tra utente e Istituto vengono criptate utilizzando le cosiddette "RSA SecurID card".

3. Olanda (Statistics Netherlands)

Già dai primi anni novanta Statistics Netherlands rilascia file di microdati protetti per scopo di ricerca. Tali file vengono rilasciati dietro la sottoscrizione di un rigido contratto che regola il comportamento dell'acquirente nel rispetto della riservatezza dei rispondenti e dopo che gli stessi siano stati protetti utilizzando una serie di tecniche come la soppressione locale e la ricodifica globale. Se in un primo periodo tali file appagavano le necessità dei ricercatori, con lo sviluppo di strumenti e di nuove metodologie di analisi, il livello di dettaglio informativo contenuto in detti file non permette, in alcuni casi, di effettuare delle analisi complete.
La legge statistica Olandese è comunque molto rigida e non permette l'analisi di dati con un maggiore contenuto informativo se non sotto lo stretto controllo dell'Istituto di Statistica. Per superare questi problemi è stato istituito all'interno dell'Istituto stesso un laboratorio di dati (**OnSite facilities**) dove vengono messi a disposizione di selezionati ricercatori file di microdati particolarmente dettagliati. I risultati delle elaborazioni vengono controllati da personale apposito per possibili violazioni sulla riservatezza e solo successivamente rilasciati al ricercatore. Quindi, tutti i risultati intermedi possono essere analizzati esclusivamente all'interno del laboratorio e solo l'output finale, previo controllo, può essere rilasciato all'utente. Il laboratorio dati (OnSite facilities) è stata una scelta di successo da parte dell'Istituto in quanto ha avuto un buon riscontro tra i ricercatori universitari. L'unico limite dell'OnSite facilities è legato agli spostamenti che i ricercatori devono affrontare per recarsi nei locali del Laboratorio. Questo ultimo aspetto e la sempre maggiore possibilità di utilizzare connessioni internet sicure hanno creato le premesse per la costruzione di un canale di accesso remoto.

Il progetto, che ha come obiettivo la costruzione di tale servizio prende il nome di "OnSite@home" (Hundepool *et al.,* 2005) ed è in fase di sperimentazione. L'idea, che è alla base del progetto, è quella di costruire un laboratorio sulla rete internet che rispecchi il più possibile le caratteristiche, in termini di riservatezza, del laboratorio presente all'interno dell'Istituto. Esclusivamente il punto di

accesso al servizio viene trasferito nella sede di lavoro dell'utente, mentre il resto della struttura organizzativa rimane invariato. Le caratteristiche dell'accesso sono le seguenti:

- solo utenti autorizzati possono accedere al servizio;
- i microdati rimangono all'interno dell'Istituto;
- gli output da rilasciare devono essere controllati per verificare il rispetto della riservatezza;
- per poter accedere al servizio è necessario sottoscrivere un contratto legale.

Per quanto riguarda il primo punto il servizio OnSite@Home utilizza l'identificazione biometrica. Per poter accedere al servizio i ricercatori autorizzati vengono identificati tramite le proprie impronte digitali; durante il progetto pilota le impronte vengono verificate ad ogni inizio di sessione di lavoro, mentre per il futuro ciò avverrà casualmente durante le varie sessioni.

La rete usata dal servizio "OnSite@Home" è separata dalla rete di produzione e la connessione che l'utente fa con il servizio è di tipo "Terminale". Ovvero, l'utente, usando un Citrix MetaFrame (Interfaccia web), può vedere i risultati intermedi delle proprie analisi sul video, ma non può ne stampare né salvare tali informazioni sul proprio computer.
Gli output che vengono rilasciati, infatti, vengono sempre prima controllati da personale preposto. Nel momento in cui un utente vuole salvare un particolare output, deve mettere lo stesso su di una specifica directory all'interno dell'ambiente di lavoro che gli è stato assegnato. Tale directory è costantemente controllata e, nel caso in cui contenga dei file, il personale dell'Istituto viene prontamente avvertito. In questo modo il file di output viene immediatamente controllato. La verifica della non violazione delle regole per la riservatezza è totalmente manuale. Anche se il controllo manuale risulta essere un processo che richiede molto impegno sia di risorse che di tempo, viene considerato necessario, in quanto, lo sviluppo di un software che possa svolgere tale lavoro e che riesca a prendere in considerazione output di diversa natura e di diverso formato è molto complesso e ancora non disponibile .

Come per il Laboratorio, prima di poter accedere al servizio "OnSite@Home", deve essere sottoscritto un contratto tra l'Istituto di Statistica e l'Istituto presso il quale lavora il ricercatore inoltre il ricercatore stesso deve sottoscrivere una dichiarazione di non violazione delle regole sulla riservatezza.

Durante la fase di sperimentazione sono state abilitate due postazioni all'interno dei locali dell'Università di Tilburg (Netspar), sulle quali sono state istallate due macchine per la lettura delle smartcard e delle impronte digitali. Cinque ricercatori di tale università sono stati autorizzati all'accesso ai dati dell'Istituto Olandese tramite il canale dell'OnSite@Home e sei ricercatori dell'Istituto sono stati assegnati al controllo degli output.

4. Altre Esperienze Europee

Germania (Federal Statistical Office of Germany)

A partire dai primi anni 80[2], con l'approvazione della prima legge Statistica Federale che regola la diffusione di statistiche ufficiali sotto forma di microdati, l'Istituto federale di statistica produce una serie di file di microdati anonimizzati per diversi scopi; file PUF (pubblic use file) e SUF (scientific use file). Tali file si differenziano sostanzialmente per il livello di protezione applicato e, conseguentemente, per la tipologia dell'utenza che ne fa richiesta.

[2] http://www.unece.org/stats/publications/statistical.confidentiality.pdf *"Research Data Centres of the official statistics"* Tom Wende and Markus Zwick. Proceedings of the Seminar Session of the 2003 Conferenze of European Statisticians.

Nel caso in cui un ricercatore necessiti comunque di maggiori informazioni di quelle contenute nei file SUF o di dati relativi ad indagini campionarie per le quali non sono stati creati dei file protetti, i Centri di Ricerca Dati (uno ogni stato federale) hanno sviluppato due ulteriori canali, il Controlled Remote Data Processing (CRDP) e lo Special Data Processing (SDP).
Il CRDP si struttura come un accesso remoto via e-mail. In un primo momento il ricercatore può condurre le sue analisi sui dati anonimizzati o su file di dati detti "strutturali". Questi ultimi sono dei file che risultano essere uguali a quelli contenenti i dati originali in termini di struttura ma non in termini di contenuti. In questo modo vengono prodotti programmi di analisi in formato SAS, SPSS, STATA, ecc... che successivamente vengono inviati ai Centri di Ricerca Dati. Qui personale apposito provvede all'applicazione degli stessi a dati protetti in misura inferiore rispetto ai SUF o, in alcuni casi, non protetti. Personale del centro provvede a controllare l'output dei programmi per la tutela della riservatezza.
Un' ulteriore forma di accesso remoto è il cosiddetto "Special Data Processing" che si prefigura come una forma di consulenza statistica da parte del Centro. In questo caso il ricercatore descrive il progetto di ricerca ad un rappresentante del Centro che successivamente svolge, in maniera autonoma, l'intero processo di analisi dei dati.

Gran Bretagna (ONS)
Anche l'ONS, a partire dal 2004, ha predisposto un sistema di accesso remoto, detto Virtual Microdata Laboratory (VML) (Ritchie, 2005), per l'accesso a file di microdati di impresa. Il sistema VML viene definito come un sistema "thin-client" ovvero i ricercatori possono connettersi attraverso l'uso di internet al sistema, applicare i propri programmi ai dati, che si trovano in un ambiente protetto all'interno dell'Istituto, e infine ottenere i risultati delle proprie elaborazioni. Questo sistema è simile a quello già in uso in Danimarca: permette l'interrogazione sia di collezioni di dati dell'ONS non protetti, sia di collezioni ottenute anche attraverso forme di link tra dati in possesso dell'utente e dati dell'ONS.
L'utilizzo di questo sistema di accesso è condizionato alla sottoscrizione di un contratto che, a partire dal 2002, è diventato più stringente, ovvero non solo coinvolge l'utente che elabora i dati ma anche l'istituzione a cui esso appartiene, sfruttando così una forma collettiva di responsabilità.
Gli scopi che hanno creato i presupposti per lo sviluppo del sistema VML sono essenzialmente due: quello di poter produrre risultati che generalmente non vengono prodotti dai canali di statistica ufficiale, e quello di rendere fruibili collezioni di dati composte in maniera innovativa (dati linkati) attraverso l'integrazione di dati prodotti dall'ONS e dati esterni.
Le ampie possibilità di cui si avvalgono gli utenti del VML conducono ad output molto variabili (sia per tipologia di analisi che di dati interrogati) e, conseguentemente, all'impossibilità di definire l'insieme degli output accettabili in termini di tutela della riservatezza
I controlli sull'output sono quindi totalmente manuali e gli utenti vengono coinvolti in questo processo attraverso un preventivo corso in cui vengono esposti i principi e le regole da seguire. In questo modo si cerca di incoraggiare gli utenti ad avere un approccio collaborativo con il personale dell' ONS producendo così output che non presenti particolari problemi in termini di riservatezza.
Attualmente l'accesso al sistema VML è predisposto solo per postazioni presenti all'interno di locali dell'ONS. In via sperimentale si è predisposto un collegamento con la sede dell'ONS di Southport.

5. Progetti LIS e LES

Il progetto LIS[3] (Luxembourg Income Study) (Schouten, 2003) è un progetto di ricerca non-profit nato dalla collaborazione di 30 paesi (tra i quali l'Italia, rappresentata dall'Istituto di ricerche sulla popolazione e le politiche sociale – IRPPS e dalla Banca d'Italia) nel 1983, con l'obiettivo di rendere accessibili microdati provenienti da più paesi per studi comparativi sui redditi familiari.
Il database del progetto LIS è quindi costituito da una collezione di dati campionari relativi alle indagini sui bilanci delle famiglie che fornisce informazioni demografiche ed economiche sia a livello individuale che familiare. Nel 1994 nasce un progetto del tutto analogo, LES (Luxembourg Employment Study), che, a differenza del progetto LIS, ha per oggetto microdati campionari relativi alle indagini sulle forze di lavoro provenienti da diversi paesi.
Per entrambi i progetti, i microdati vengono dapprima standardizzati e resi confrontabili in modo tale che sia possibile studiare mercati del lavoro e sistemi di reddito tra loro diversi. I dati contenuti nei due database sono protetti attraverso l'applicazione sia di tecniche di ricodifica che tecniche perturbative[4].
Entrambi i database sono consultabili tramite un sistema di accesso remoto che si basa sul software LISSY appositamente sviluppato.

Gli utenti possono accedere al sistema inviando ad un preciso indirizzo, detto mail server, un messaggio di posta elettronica in cui sono contenute le istruzioni da applicare ai dati e le informazioni di identificazione relative all'utente. I programmi che possono essere utilizzati sono SAS, SPSS e STATA. Gli utenti possono utilizzare file sintetici per poter verificare la correttezza della sintassi dei propri listati di istruzioni.
Il mail server è l'unica parte del sistema "visibile". Il sistema è costituito da diverse componenti software che interagiscono tra loro tramite una o più reti. Le varie componenti quindi ricevono le richieste degli utenti, le applicano ai dati e restituiscono i risultati statistici agli utenti tramite posta elettronica. Il ruolo fondamentale è quello svolto dal cosiddetto Post office, ovvero la componente che riceve le richieste. Tale componente ogni 5 secondi riceve le richieste, le analizza dal punto di vista della sicurezza, distribuisce le richieste alla componente che applica i programmi ai dati, invia ai vari utenti le elaborazioni richieste.

Dal punto di vista dei controlli, il post office prima verifica la correttezza delle informazioni relative all'utente e, in seguito, esamina la sintassi delle istruzioni di programma per appurare che il tipo di procedura statistica da applicare ai dati non violi le regole sulla tutela della riservatezza concordate. In particolare ricordiamo che vengono disabilitate le procedure tipo: PROC PRINT (SAS), LIST (SPSS e STATA) o FREQUENCIES. I risultati che rispettano tutti i principi di riservatezza previsti dal progetto (LIS LES) vengono automaticamente restituiti all'utente. Nel caso il sistema LISSY verifichi una qualche violazione automaticamente rimuove la richiesta dal sistema che viene così esaminata manualmente da personale apposito.
All'interno di una politica di miglioramento nella diffusione di microdati, nell'ambito del progetto LIS è stato sviluppato un nuovo sistema di interrogazione on-line per la costruzione di tabelle. Questo servizio, che prende il nome di **Online web tabulating services**[5], è disponibile solo per i dati relativi ai paesi che hanno rilasciato il permesso al trattamento dei propri dati attraverso un servizio via web (attualmente solo 9: Canada, Germania, Italia, Messico, Olanda, Svezia, Gran Bratagna e Stati Uniti). Solo un limitato insieme di variabili, che comunque ricopre tutti gli argomenti base del progetto LIS, è disponibile per l'interrogazione via web. Anche per accedere a

[3] http://www.lisproject.org, 1-2006
[4] http://www.lisproject.org/introduction/faq.htm, 1-2006

[5] http://www.lisproject.org/webtabulator.htm , 1-2006.

questo servizio è necessario sottoscrivere una richiesta correlata dalla presentazione di un dettagliato progetto di ricerca. Nel momento in cui il progetto viene accettato viene rilasciato all'utente un account con il quale è possibile accedere al servizio di tabulazione via web.

6. Progetto PiEP

Il progetto PiEP[6] (Pay Inequalities and Economic Performance Project) è condotto da una commissione internazionale di ricercatori universitari in stretta collaborazione con Eurostat e alcuni Istituti nazionali di statistica. Il progetto utilizza i microdati relativi all'indagine ESES (European Structure of Earnings Survey) del 1995 di 6 paesi (Belgio, Danimarca, Irlanda, Italia, Spagna, Gran Bretagna). I dati sono conservati nella sede di Eurostat a Lussemburgo e sono consultabili tramite un sistema di accesso in remoto gestito dalla London School of Economics. Il sistema di accesso utilizzato è un adattamento del sistema LISSY descritto nel paragrafo precedente, detto appunto PiEP-LISSY. Questa versione differisce da quella utilizzata per i progetti LIS/LES in quanto prevede dei controlli più restrittivi su comandi o combinazioni di essi che possono fornire informazioni confidenziali. In particolare vengono disabilitate le procedure per il calcolo delle tabelle di frequenza, la rappresentazione grafica di dati individuali, dei residui e dei valori estremi[7].

7. Australia (Australian bureau of Statistics – ABS)

A partire dal 1983 l'Istituto di Statistica Australiano (Australian Bureau of Statistics - ABS) diffonde informazione sotto forma di record individuali (microdati). Tali collezioni di dati possono essere rilasciate solo dopo aver stabilito che le informazioni contenute in esse siano tali da non permettere l'identificazione di persone o organizzazioni. I microdati vengono così rilasciati sotto forma dei cosiddetti Confidential Unit Record files (**CURF**)[8] ovvero file individuali che, attraverso l'applicazione di alcune tecniche di protezione, vengono anonimizzati.

Il rilascio di tali file è altamente controllato dall'Istituto il quale ha istituito una commissione apposita (**Microdata Review Panel**) che analizza le proposte di rilascio e verifica il tipo e il livello di protezione applicata ai file in modo tale da garantire la riservatezza dei rispondenti alle indagini.
Per garantire la riservatezza dei rispondenti l'ABS si avvale di tre strumenti: tecniche di protezione, contratto legale e differenziazione del canale di diffusione.

In generale le **tecniche di protezione** applicate sono:
- ricodifica globale;
- perturbazione casuale dei valori;
-data swapping;
-eliminazione dal file di recod che presentano caratteristiche "non usuali".
A seconda del livello di protezione applicata vengono prodotti tre tipi di file CURF: *base, esteso e speciale.*

Un **contratto legale** viene sottoscritto sia dai ricercatori che fanno richiesta dei file CURF che dalle organizzazioni a cui essi appartengono. Tale contratto regola il comportamento che deve essere tenuto dagli utenti sia durante l'utilizzo dei dati contenuti nei file che nella diffusione dei risultati ottenuti.

---

[6] http://cep.lse.ac.uk/piep/, 1-2006.
[7] http://cep.lse.ac.uk/piep/papers/Final_Report_V5.pdf, 1-2006.
[8] http://www.abs.gov.au/Websitedbs/D3110129.NSF/85255e31005a1918852558ac00697645/72d92417a0ba71b5ca256d01002c47a4!OpenDocument#What%20are%20CURFs%3F 2-2006

I canali di diffusione dei file CURF messi a disposizione degli utenti sono: CD-ROM, Accesso Remoto (Remote Access Data Laboratori-RADL), Laboratorio Dati (ABS Data Laboratory – ABSDL).

Il sistema **RADL**[9] messo a disposizione dall'ABS è un sistema di query on-line che permette l'accesso a file CURF di tipo base ed esteso. L'utente, utilizzando un account sicuro (username e password), può accedere ai dati attraverso un qualsiasi computer che prevede un accesso internet.

L'accesso è autorizzato previa presentazione di un progetto di ricerca in cui si descrive l'utilizzo dei dati che si vogliono consultare. Il contatto avviene a livello di organizzazione di appartenenza e la richiesta può riguardare più di un file CURF. Viene così definito un responsabile, all'interno dell'organizzazione, che si prende cura dei contatti con l'ufficio dell'ABS preposto al rilascio dei file CURF. Le autorizzazioni, una volta ottenute, devono essere rinnovate annualmente e una serie di informazioni come l'uso che viene fatto dei file CURF, eventuali pubblicazioni o stato di avanzamento del progetto devono essere fornite all'Istituto che monitora tali informazioni non solo per scopi di controllo ma anche per analizzare possibili cambiamenti nella tipologia delle richieste dei dati.
In generale l'Istituto non rilascia CURF a ricercatori che appartengono ad enti non australiani, anche se le autorizzazioni per accessi via RADL vengono analizzate caso per caso. Per poter accedere ai dati via RADL è necessario comunque appartenere ad organizzazioni affiliate ad Università Australiane, enti governativi o Istituti Nazionali di Statistica.
Una volta che la richiesta viene accettata e che il contratto legale viene firmato sia dall'organizzazione che dal singolo ricercatore che utilizzerà i dati, l'ABS fornisce l'account e la password necessaria per accedere via web (sito ABS) ai dati richiesti. E' il sistema stesso che fornisce il software di analisi necessario, in particolare vengono forniti i programmi SAS e SPSS.

L'utente, dopo aver specificato il file di dati a cui è interessato, può analizzare gli stessi scrivendo un file di istruzioni, sas o spss, in una directory appositamente dedicata. Tali istruzioni verranno successivamente applicate in modalità batch ai dati che restano comunque all'interno dell'ambiente informatico dell'istituto.

Nel caso in cui venga soddisfatta una serie di vincoli, imposti direttamente dal sistema RADL sia sulla natura della query che sull'ampiezza e natura dell'output richiesto, l'output prodotto viene reso automaticamente disponibile. I vincoli imposti dal sistema non vengono considerati nel caso di interrogazioni fatte sui file CURF di tipo base (resi disponibili anche su CD-ROM). L'output non viene rilasciato nel caso in cui non vengano soddisfatti i vincoli imposti dal sistema RADL o qualora ci siano errori nelle istruzioni inviate.
Nel primo caso l'utente può comunque richiedere un ulteriore controllo di tipo manuale per analizzare la possibilità di rilascio dell'output stesso. Nel secondo caso viene fornito all'utente il file di log. Per poter verificare la correttezza delle istruzioni e valutare i programmi di analisi che si vogliono applicare, l'ABS ha messo a disposizione, per tutti i file CURF disponibili via RADL a partire dal settembre 2003, un file di test scaricabile, con contenuti casuali e struttura identica al file CURF ad esso associato.

[9] http://www.abs.gov.au/websitedbs/D3110129.NSF/0/36bd92d8f488355dca256f4a0010c110/$FILE/ABS%20RADL%20User%20Guide_July2005.pdf 2-2006

8. Canada (Statistics Canada)

Statistics Canada rilascia, a partire da dati relativi ad indagini sociali, file di microdati protetti (Public Use Microdata File - PUMFs), per motivi di ricerca, già dai primi anni '70[10]. Sono raramente rilasciati, invece, i file di dati di impresa o longitudinali in quanto, per tali dati, si ritiene sia troppo elevato il rischio di violazione della riservatezza dei rispondenti.
I file PUMFs vengono protetti attraverso l'applicazione di diverse tecniche di protezione; le informazioni geografiche vengono rilasciate solo ad un livello piuttosto aggregato e molte delle informazioni circa il disegno campionario (strati, cluster e pesi campionari) vengono omesse.
Già nei primi anni 90', proprio per sopperire alle richieste di informazioni non soddisfatte dai PUMFs, come ad esempio quelle che necessitano di particolari variabili non contenute nei file o delle indicazioni campionarie per il calcolo della varianza, Statistics Canada propone l'accesso remoto come canale di accesso indiretto alle proprie collezioni di dati.
In un primo momento l'accesso remoto è stato offerto solo per un piccolo numero di indagini campionarie, per alcune delle quali non era prevista la predisposizione di file PUMFs e il sistema è direttamente e interamente gestito dalle divisioni che si occupano delle singole indagini. L'utilizzo di tale canale è ristretto ad un numero molto limitato di ricercatori, i quali devono presentare un dettagliato progetto di ricerca. Solo dopo l'accettazione di tale progetto e la sottoscrizione di un contratto che regoli l'uso dei dati e degli output, viene rilasciata l'autorizzazione all'accesso al servizio.
Nel caso in cui sia predisposto un accesso di tipo remoto, sono fornite all'utente tutta una serie di informazioni necessarie per poter formulare i propri programmi di analisi, ovvero:

- una documentazione che includa la descrizione della struttura del file di dati confidenziali (file di base o file master);
- il file di prova che segua la struttura del file master affinché l'utente possa testare i propri programmi su tale file.

Inoltre ogni indagine che decide di offrire un canale di accesso remoto ai propri dati deve provvedere all'adattamento del sistema che permette di ricevere programmi via e-mail e successivamente di rinviare all'utente l'output dell'elaborazione dopo che lo stesso abbia superato un accurato controllo manuale dal punto di vista della riservatezza. Un controllo di tipo automatico risulterebbe troppo difficile sia a causa della possibilità, da parte degli utenti, di utilizzare più pacchetti statistici quanto per la complessità dei dati trattati.
Di seguito descriviamo gli approcci seguiti per l'accesso remoto da due indagini differenti: National Population Health survey (NPHS) e Survey of Labour and Income Dynamics (SLID)[11](Tambay *et al*.,2003).

1) National Population Health survey (NPHS)

L'indagine longitudinale NPHS è l'indagine che maggiormente è stata richiesta attraverso il canale dell'accesso remoto. Il successo di tale indagine è sicuramente legato alla completezza delle informazioni disponibili. Infatti, è a disposizione degli utenti una documentazione dettagliata sul piano di campionamento, un dizionario sui dati, un file sintetico molto realistico su cui gli utenti possono testare i propri programmi e la possibilità di utilizzare diversi software come SPSS, SAS, Stata ecc... Inoltre risultano relativamente brevi i tempi di ritorno degli output.
Il file sintetico viene accuratamente predisposto per replicare la struttura del campione di base anche se contiene un numero inferiore di record.

[10] http://www.statcan.ca/english/Dli/continuumofaccess.htm, 1-2006 data dell'ultima visita al sito.
[11] http://www.statcan.ca/english/research/75F0002MIE/75F0002MIE1994014.pdf

2) Survey of Labour and Income Dynamics (SLID)

Anche questa indagine è di tipo longitudinale ed è caratterizzata da una struttura molto complessa. Tale complessità è legata alle particolari unità di analisi incluse nell'indagine. Per agevolare l'accesso e l'utilizzo di tali dati si è sviluppato un particolare sistema di reperimento dati (SLIDRET). Attraverso questo sistema l'utente può creare un dataset che corrisponde alle proprie necessità senza dover comprendere l'intera struttura dei dati.
Un utente che voglia procedere all'analisi di tali dati attraverso il canale dell'accesso remoto, deve per prima cosa creare un proprio file di analisi attraverso il sistema SLIDRET e, successivamente costruire i file di programma (SAS, SPSS, STATA) da mandare via e-mail a Statistics Canada che, tramite personale autorizzato, provvederà ad eseguirli sul file di analisi. Esiste una versione pubblica di SLIDRET alla quale è associato un database vuoto ma con la stessa struttura della versione non pubblica.
I file di output ottenuti dopo aver eseguito le analisi richieste dall'utente vengono infine analizzati per garantire la riservatezza e, se soddisfano questo requisito, vengono restituiti via e-mail all'utente.

In generale non vengono rilasciati:
-dati a livello individuale anche se si riferiscono ai residui di un modello di regressione,
-valori minimi o massimi. In particolare ciò avviene per quelle variabili sensibili che, nel rilascio del relativo file PUMF vengono trattate con top-coding o bottom-coding
- valori di celle o di statistiche che si basano su un numero di rispondenti molto esiguo (generalmente 5)
- informazioni che rivelano l'esatta locazione delle unità campionarie.
In generale tutte le regole che vengono adottate per verificare la riservatezza degli output da rilasciare vengono comunque implementate tenendo in considerazione le informazioni contenute nei file PUMFs.

9. National Center for Health Statistics (NCHS)

Il National Center for Health Statistics (USA) rilascia attualmente più di 500 file PUMS[12] (Pubblic Use Microdata File) che ricoprono la maggior parte dei campi di ricerca che dell'ente. Nonostante ciò, le restrizioni sul dettaglio di alcune variabili, imposte dai controlli sul rischio di violazione della riservatezza degli intervistati, hanno notevolmente limitato l'utilizzo di alcuni dati nell'ambito della ricerca scientifica. Per rispondere alle sempre più pressanti richieste della comunità scientifica il NCHS nel 1998 ha costituito un Laboratorio di dati elementari (Research Data Center-RDC) tramite il quale è possibile accedere a file di dati dettagliati in un ambiente sicuro senza compromettere la riservatezza dei rispondenti. Contemporaneamente al Laboratorio, nel 1998 è stato creato un sistema di accesso remoto detto ANDRE[13] (ANalytical Data Research by Email) che viene considerato come parte integrate del Laboratorio stesso.
L'obiettivo principale di ANDRE è quello di fornire uno strumento flessibile, economico e conveniente per l'analisi statistica su dati considerati riservati. Tutti i file di microdati messi a disposizione da ANDRE, anche se protetti attraverso diverse tecniche di protezione (*restricted data)*, sono dati che presentano un contenuto informativo maggiore rispetto a quelli diffusi via CD-Rom; non sono accessibili direttamente in quanto restano all'interno dell'ambiente informatico dell'ente.

[12] PUMS sono file di microdati protetti con diverse tecniche come: data swapping, perturbazione casuale dei valori di alcune variabili, ricodifica ecc.
[13] http://www.cdc.gov/nchs/r&d/rdc.htm
http://www.cdc.gov/nchs/data/GuidelinesRDC11-8-05.pdf 1-2006.

In alcuni casi il canale di accesso remoto ANDRE viene utilizzato dai ricercatori in maniera congiunta all' utilizzo del Laboratorio. Ciò può avvenire sia in una fase preliminare, al fine di condurre elaborazioni esplorative dei dati, che in una fase conclusiva o di rifinitura del lavoro di ricerca.

Il sistema ANDRE permette agli utenti di effettuare elaborazioni attraverso l'invio per posta elettronica di programmi scritti in SAS. Il SAS è stato scelto sia per la sua ampia diffusione nell'ambiente della ricerca statistica che per la possibilità di impostare delle query filtro per il controllo automatico del processo di elaborazione dei dati (*filtering software for setup)*. Il controllo automatico avviene sia sui comandi inviati dagli utenti che sugli output prodotti da rilasciare. Per quanto riguarda i comandi, alcuni di essi vengono disabilitati (come PRINT, ADD, OBS, PROC IML; la lista completa è consultabile sulla pagina web del laboratorio), mentre per altri viene modificato l'output (come per la PROC MEANS, N MEANS, STD). Il controllo automatico sugli output avviene attraverso la soppressione di valori estremi di variabili considerate particolarmente sensibili o identificative e la soppressione di celle corrispondenti ad un'ampiezza campionaria inferiore a imiti prefissati. Sebbene in ANDRE il controllo sia quasi completamente automatizzato (per il 90%), alcuni casi dubbi vengono sottoposti all'attenzione di personale del Laboratorio per una verifica più specifica.

Per accedere ad ANDRE è necessario sottoporre un progetto di ricerca al personale del Laboratorio del NCHS, che ne esamina la fattibilità e la conformità alle regole che vincolano il rilascio di dati elementari attraverso l'accesso remoto. Una volta approvato il progetto e sottoscritto un contratto legale, all'utente viene assegnato un identificativo ed una password. L'utente può inviare i suoi programmi da qualsiasi indirizzo e-mail, mentre gli output verranno inviati esclusivamente all'indirizzo fornito dall'utente nel modulo di richiesta, indirizzo valutato come credibile e sicuro da personale dell'istituto.

I ricercatori possono accedere sia a file di dati in possesso del laboratorio che a file costruiti ad hoc attraverso un processo di abbinamento con dati in possesso dell'utente stesso (file linkati). Ciascun file di dati messo a disposizione dell'utente viene appositamente preparato, da personale dell'Istituto, ed inoltre può essere consultato esclusivamente dall'utente che ne abbia fatto specifica richiesta.
Il sistema di accesso remoto ANDRE, che ha avuto 45 utenti in 5 anni, presenta comunque dei limiti e dei vincoli; per questo motivo il NCHS sta progettando un nuovo sistema di accesso remoto denominato ANDREW (ANalitic Data Reasearch by Email and Web) (Gambhir *et al.,* 2005). Tale sistema sarà costruito sulla base dell'architettura già collaudata per andre, ma presenterà miglioramenti dal punto di visto degli algoritmi utilizzati per il controllo della riservatezza e nuove caratteristiche che renderanno più flessibile ed efficiente la sua utilizzazione:

- ANDREW sarà totalmente automatizzato;
- sarà possibile utilizzare oltre al SAS i programmi Sudaan e Stata;
- la piattaforma tecnica sarà costruita utilizzando l'ambiente di sviluppo MS Visual Studio;
- gli algoritmi per il controllo della riservatezza già utilizzati in ANDRE saranno potenziati. Ad esempio:
  1- verranno oscurati i valori estremi ottenuti con la Proc Univariate e utilizzati opportuni programmi statistici, presenti in commercio, per oscurare valori particolarmente sensibili in tabelle ad una o due dimensioni.
  2- Il controllo non sarà più effettuato direttamente sui file di output del SAS (file .lst). Verranno, infatti, sviluppati degli algoritmi che in un primo momento trasformeranno i file di output di SAS, Sudaan e Stata in formati compatibili con i programmi statistici

adottati per il controllo della riservatezza e che successivamente si occuperanno di riconvertire l'output nel formato di origine (ovvero quello del programma di analisi utilizzato).

Nonostante non sia stato ancora risolto il problema metodologico del controllo della tutela della riservatezza nel caso di una molteplicità di richieste da parte di un singolo utente, ANDREW affronterà tale problema attraverso un controllo della dimensione e del tipo di risultati richiesti in relazione a particolari variabili considerate a priori più o meno rischiose. Ovvero, una commissione di esperti valuterà ogni file di dati consultabile tramite ANDREW identificando il potenziale di pericolosità, dal punto di vista della riservatezza, delle variabili presenti. Per ognuna di esse verrà così fissato un livello di tolleranza. Nel processo di controllo degli output ANDREW esaminerà tutte le variabili considerate a rischio e, nel caso in cui vengano superati i livelli di tolleranza prefissati, ANDREW informerà automaticamente il responsabile del sistema generando inoltre un report in cui si evidenzieranno tutti i risultati rilasciati per la variabile in questione.

L'interfaccia utente sarà una componente Web. La possibilità di utilizzare ANDREW senza dover specificare un particolare linguaggio statistico sarà implementata nella cosiddetta Interfaccia grafica per utenti (GUI Graphic User Interface) e permetterà di specificare variabili e vincoli attraverso il solo utilizzo del mouse. La GUI sarà particolarmente utile a chi non conosce bene programmi statistici (come SAS Stata ecc.) e a chi è interessato a risultati molto veloci. L'uso di una tecnologia come quella della GUI permette di effettuare uno stretto controllo su ciò che l'utente può richiedere, a differenza di quanto avviene nel caso in cui l'utente invia un proprio programma.

10. Bureau of Census (BOC)

Il Bureau of Census (USA) ha aperto diversi Laboratori di microdati dove ricercatori autorizzati possono accedere a dati per i propri progetti di ricerca. Tuttavia, per rispondere alle sempre maggiori richieste di informazione e per migliorare la qualità di tale servizio, il BOC ha predisposto due canali di accesso remoto: l'Advanced Query System, per svolgere interrogazioni su tabelle, e il Microdata Analysis System (in fase di test) per l'analisi attraverso modelli statistici.

Entrambi i sistemi di accesso remoto previsti dal BOC sono definiti come web-based system. Tali sistemi prevedono un accesso ai dati attraverso una pagina web e un insieme di interrogazioni possibili sui dati previste direttamente in fase di progettazione del sistema stesso. Questo tipo di sistema viene definito anche sistema "abilitante" ovvero un sistema dove l'utente non si vedrà rifiutare il rilascio di un output da lui prodotto in quanto l'utente potrà produrre esclusivamente risultati che rispettano i vincoli, legati al rispetto della riservatezza, impostati direttamente durante la fase di progettazione del sistema stesso.

Advanced Query System – AQ[14]

Il Sistema Avanzato di Interrogazione AQ è stato sviluppato per poter permettere agli utenti di richiedere informazioni sotto forma di tabelle non standard a partire dai dati del Censimento del 2000 (dati relativi all'indagine *short-form* e dati campionari relativi alla *long-form*).
In un primo momento tale sistema era stato progettato come parte integrante di *American FactFinder*; un più ampio impianto finalizzato alla diffusione via internet di tabelle standard predefinite. Nell'aprile del 2003, dopo una fase di test condotta sui dati relativi Censimento del

[14] http://64.233.183.104/u/census?q=cache:Ua3zN4dcg1YJ:www.census.gov/srd/sdc/AdvancedQuerySystem.pdf+remote+access&hl=en&ie=UTF-8 (2-2006)

2000, l'accesso al sistema AQ è stato reso disponibile ad alcune agenzie e centri di ricerca federali che hanno una grande esperienza nell'utilizzo di dati censuari (State Data Center, Census Information Center e State Legislatures).

I dati contenuti nel database sul quale avvengono le interrogazioni sono file di microdati precedentemente protetti con tecniche di data swapping e ricodifica globale.

AQ prevede un sistema di filtri che limita le richieste che l'utente può fare. In particolare possono essere selezionate solo tre variabili per tabella, il più piccolo dettaglio geografico disponibile è il "*census block"* (sezione di censimento) per i dati relativi all'indagine completa mentre per l'indagine campionaria è il "*census tract"* (suddivisione geografica definita sulla base dell'omogeneità nel tempo di alcune caratteristiche demo-sociali); entrambi sono dettagli territoriali più informativi rispetto a quelli disponibili su *American FactFinder*. Ogni area selezionata deve comunque avere un minimo di 200 persone.
Il sistema è completamente automatizzato e non prevede l'intervento umano. I log delle richieste dei vari utenti sono conservati e periodicamente analizzati sia per verificare possibili violazioni alla riservatezza dei rispondenti, sia per determinare le richieste più frequenti.

L'utente, per poter utilizzare AQ, deve fare una registrazione gratuita con il BOC il quale assegna all'utente login e password per poter entrare nel sistema.

I risultati dell'interrogazione, generati in tempo reale, vengono restituiti in pochi minuti sotto forma di tabelle direttamente via web; possono quindi essere salvate in diversi formati o direttamente stampate. Il sistema è disponile ogni giorno per 24 ore.

Microdata Analysis System

All'interno dei laboratori di dati elementari del BOC, dislocati in diversi punti del paese, vengono in particolare promosse ricerche basate su applicazioni di modelli e il risultato di tali analisi viene rilasciato previo controllo da parte di personale del laboratorio. Solo in casi particolari il controllo viene effettuato dalla Commissione di Controllo per la Riservatezza dei dati del Bureau of Census. Nel corso di più di dieci anni di attività dei laboratori, i controlli fatti sugli output da modelli hanno messo in evidenza la mancanza di problemi dal punto di vista della riservatezza di detti output. L'obiettivo quindi è diventato quello di riprodurre una tale situazione in un sistema automatico di interrogazione dei dati.

Il Bureau of Census ha così finanziato un progetto per la realizzazione del prototipo di un sistema di interrogazione remota del tipo web-server che abbia come richieste applicazioni di modelli statistici ai microdati: *Microdata Analysis System* (Steel *et al.,* 2005). In particolare, come dati di test vengono utilizzati i microdati relativi ai pubblic use file del Current Population Survey, quindi dati protetti con sistemi di perturbazione, e come programma statistico viene utilizzato il programma SAS.

Nella realizzazione di tale sistema un ruolo fondamentale è svolto dalla definizione della strategia per la tutela della riservatezza da adottare, che interessa 5 fasi differenti: fase di preparazione, fase esplorativa, fase di definizione dell'universo di interesse, fase di definizione del modello e fase di definizione dei risultati.

Durante la *fase esplorativa* è molto importante per l'utente poter richiedere le distribuzioni univariate e bivariate prima di dover specificare il modello che si vuole analizzare. Inoltre deve

poter richiedere tabelle per quasi tutte le variabili categoriche ad un determinato livello geografico. Questa fase è molto importante per l'utente perché può ottenere informazioni utili per ben definire il modello statistico da richiedere.
I requisiti di riservatezza in questa fase sono definiti proprio nella fase di *preparazione dei dati*.
In particolare, per le indagini CPS è possibile analizzare tabelle bivariate che non presentino una frequenza inferiore a 100,000 unità. Per le variabili numeriche è possibile ottenere degli indicatori o delle rappresentazioni grafiche.

Anche la fase di *definizione dell'universo di riferimento* comporta problemi di riservatezza. In questa fase si definisce l'ampiezza della popolazione di riferimento attraverso un insieme di condizioni; il che equivale a determinare il valore della cella di una tabella multipla. E' possibile che venga così definita una cella che presenti frequenza unitaria e che quindi, attraverso l'applicazione di un modello, sia possibile ricostruire interamente il record ad esso associato. L'utilizzo di tecniche di protezione (es. tecniche di rounding o oscuramento di celle) generalmente applicate a problemi di riservatezza per tabelle non è, in questo caso, molto appropriato. Nella definizione dell'universo si è quindi assunto che esso non possa essere costituito da meno di 75 osservazioni; tale valore garantirebbe la possibilità di costruire tabelle, a partire dall'universo così definito, con celle che non assumono mai un valore inferiore a quattro.

Problemi di riservatezza possono derivare dall'uso di modelli di regressione; in particolare da modelli di regressioni lineare, modelli logit, probit ed alcuni modelli lineari generalizzati. Nel caso in cui le variabili utilizzate nel modello siano variabili continue, il rischio di violazione della riservatezza è basso in particolare se il campione di riferimento è sufficientemente ampio. I rischi maggiori si hanno quando vengono utilizzate variabili dummy come variabili esplicative in quanto, in alcuni casi, dal punto di vista della riservatezza il modello è equivalente ad una tabella. Per questi motivi vengono impedite interazioni che coinvolgono 4 o più variabili e modelli di più di tre variabili.
Le limitazioni imposte in fase di definizione del modello fanno sì che i coefficienti dei modelli possano essere diffusi senza particolari restrizioni così come molte informazioni sui residui. Una particolare attenzione è stata posta proprio sul come rilasciare informazioni sui residui in quanto essi rappresentano uno strumento per misurare la validità dei modelli e svolgono un ruolo particolarmente importante nelle prime fasi di studio del modello da applicare. Vengono così diffusi i valori dei residui sintetici (Reiter, 2003) in quanto forniscono lo stesso tipo di informazione dei residui effettivi e possono quindi essere considerati dei validi sostituti in fase di analisi.
Per la stima dei residui sintetici viene utilizzata la routine SAS KDE (Kernel Density Estimation). Durante la fase di testing, accanto ai valori effettivi dei residui, sono stati presentati i valori sintetici proprio per verificare l'adeguatezza di quest'ultimi.

I limiti di sistemi del tipo web-server sono fortemente legati ai problemi che si incontrano anche nei sistemi in cui si diffondono tabelle, in quanto la stima dei coefficienti in alcuni modelli è equivalente alla determinazione del valore di celle di tabelle.

Il tipo di architettura su cui si basa il prototipo del sistema di accesso remoto finanziato dal Bureau of Census, in questa fase di sviluppo, è strettamente legato al tipo di dati considerati. Gli obiettivi futuri sono quindi quello di rendere il sistema applicabile anche a dati diversi utilizzando anche dati non campionari, ridurre le condizioni di topcoding, implementare altre procedure statistiche di analisi.

11. National Center for Education Statistics (NCES)

Per soddisfare le richieste di accesso a microdati, oltre ad una serie di file PUMS, il National Center for Education Statistics (USA), a partire dal 1987, fornisce agli utenti la possibilità di accedere alle proprie collezioni di dati attraverso il Disclosure Avoidance System[15] (DAS). L'esigenza di fornire tale opportunità ad utenti esterni nasce dalla considerazione che l'insieme delle protezioni applicate ai dati (top e bottom-coding, data swapping ecc.), necessarie per la creazione di file PUMS, spesso rende i dati non utilizzabili per progetti di ricerca di livello medio alto.
DAS è un programma che genera istruzioni raccolte in un file (detto das file), consentendo di specificare le informazioni che si vogliono raccogliere in tabelle. Inoltre si possono generare proporzioni, medie e coefficienti di correlazione. Le matrici di correlazione possono essere utilizzate per modelli di regressione lineari. I dati utilizzati dal sistema DAS, per la creazione di tabelle, sono dati criptati in modo tale che non siano leggibili senza l'utilizzo del programma che genera le tabelle stesse (DAS). Per ottenere una maggiore protezione vengono applicate ai dati, anche se con un impatto minore, le tecniche di protezione utilizzate anche per la creazione del file PUMS.

Il primo prototipo di DAS è stato sviluppato sotto forma di software distribuito su CD. Gli utenti potevano utilizzare il programma sul proprio computer creando le tabelle necessarie per la propria ricerca senza però poter accedere direttamente ai microdati. L'ouput previsto da questa versione del DAS era solo sotto forma di tabelle. Nel 1997 è stata sviluppata una seconda versione di DAS sotto forma di applicazione Web. Gli utenti potevano scaricare il software DAS direttamente da web. Con tale programma potevano definire le proprie richieste, creando i cosiddetti DAS-file, che venivano successivamente inviate via internet. Le richieste venivano così elaborate e successivamente restituite all'utente in meno di sei ore lavorative. Anche in questa versione di DAS l'output previsto era solo sotto forma di tabelle.

La versione corrente di DAS è stata sviluppata nel 2003. Questa ultima è disponibile sia come versione web (DAS on-line) [16] che come applicativo windows.
La versione su web permette all'utente di creare DAS-file (file di programma) specificando così le informazioni che si vogliono ottenere. Esistono sistemi DAS per ogni collezione campionaria, ma tutti hanno la stessa interfaccia grafica e la stessa struttura dei comandi. La collezione dei dati che viene interrogata dal sistema DAS, oltre ad essere criptata, è protetta attraverso l'applicazione di diverse tecniche (data swapping ecc...). I livelli di protezione adottati differiscono a seconda del livello di informazione contenuta nel file dei dati e comunque risultano essere inferiore a quelli adottati per la creazione di file PUMS. Ulteriori controlli e protezioni vengono applicati infine agli output richiesti; ad esempio per le tabelle vengono soppresse le celle che presentano meno di 30 casi.

La corrente versione di DAS, oltre ad essere su WEB, presenta dei miglioramenti anche in relazione agli output prodotti. In particolare, in modalità "Table" è possibile richiedere tabelle di stime e la corrispondente varianza della stima che viene calcolata tenendo in considerazione i complessi disegni campionari utilizzati dall'Istituto. In modalità "Correlation", invece, DAS produce matrici di correlazione che possono essere successivamente utilizzate come input per l'analisi di modelli di regressione lineare.
Il processo di elaborazione dei risultati avviene in tempo reale e il vantaggio del sistema basato su un'applicazione web è che gli output vengono restituiti all'utente in pochi secondi attraverso la rete internet.

---

[15] http://nces.ed.gov/das/ 1-2006.
[16] http://nces.ed.gov/dasol/ 1-2006.

12. Una visione analitica del sistema di accesso remoto

Come si può desumere dalla definizione stessa di accesso remoto, una caratteristica importante di questo canale è la sua *flessibilità*. In effetti, può essere implementato in diversi modi in relazione alle esigenze degli Istituti, legate soprattutto alla legge vigente sul rispetto della riservatezza, alle caratteristiche dei dati e all'intera gamma di prodotti offerti dall'Istituto stesso in termini di microdati.

In questa sezione si analizza il sistema di accesso remoto, evidenziando le fasi più significative e caratterizzanti del sistema stesso e le particolarità dell'output richiedibile dall'utente; nella tabella 1 si distinguono varie fasi e si confrontano i sistemi fin qui descritti.
Una prima fase che definisce il sistema è sicuramente la fase di protezione, caratterizzata dall'insieme degli accorgimenti legali, amministrativi, tecnologici e delle metodologie statistiche messi in atto per tutelare la riservatezza dei rispondenti. Una seconda fase è invece quella della comunicazione tra l'utente e l'Istituto o l'Agenzia Statistica che detiene i microdati. Tale comunicazione avviene in due momenti distinti del processo, ovvero nel momento dell'accesso al sistema (o interrogazione dei dati) e nel momento della restituzione dell'output. Infine nella tabella 1 si identifica il tipo di output disponibile.

In sintesi, le principali differenze tra le possibili implementazioni sono quindi legate a:

- tipo e livello di protezione che si vuole applicare ai dati sia in fase di input che di output;
- tipologia del canale di comunicazione (e-mail o via web) che si vuole attivare tra l'utente e l'Istituto
- tipo di output che può essere richiesto.

Come già accennato, nella determinazione di un sistema di accesso remoto, per quanto riguarda la fase di protezione, possono essere coinvolte una serie di limitazioni, legali o tecniche, sulle quali l'Istituto o l'Agenzia di statistica può agire per meglio raggiungere gli obbiettivi con l'implementazione del sistema stesso. In particolare abbiamo: la presenza o meno di un contratto legale da sottoscrivere per poter ottenere l'accesso al sistema, possibili limitazioni sul numero e sulla collocazione delle postazioni di lavoro abilitate a tale accesso, limitazione sul tipo di utenza abilitata.
In generale possiamo dire che maggiori sono le restrizioni tecniche e legali adottate dal sistema, maggiore è l'informazione messa a disposizione nel database di interrogazione, in quanto tali limitazioni in qualche modo sopperiscono al minor livello di protezione applicato ai dati.

Per quanto riguarda il sistema di dati su cui avvengono le interrogazioni abbiamo due possibili situazioni: un file di dati reali non protetti, come nel caso dell'accesso remoto messo a disposizione da Statistics Canada, CBS e Statistics Denmark, oppure un file di dati protetti come nel caso dei file CURF per l'Australian Bureau of Statistics e dei file messi a disposizione dalle varie agenzie Americane (BOC, CES e NCHS). Tali differenze sono da inserirsi nell'ambito delle politiche di tutela della riservatezza perseguite dagli Istituti.
Per quanto riguarda gli output ottenibili abbiamo casi in cui il sistema di accesso remoto prevede esclusivamente il rilascio di tabelle (AQ, DAS), casi in cui gli output rilasciati possono essere esclusivamente il risultato dell'applicazione di modelli statistici (PiEP, Microdata Analysis System) o casi in cui sono disponibili entrambi le tipologie di output. Ovviamente anche limitazioni sul tipo di output sono una protezione nel senso di un maggiore controllo sul processo.
Il controllo che viene effettuato sugli output può essere di tipo automatico o manuale. La tipologia del controllo è molto legata al tipo di dati messi a disposizione, all'output ottenibile e al canale informatico utilizzato per la comunicazione con l'utente. Nel caso in cui i dati di base siano stati

preventivamente protetti e/o gli output ottenibili siano tabelle, il controllo è prevalentemente di tipo automatico, come nel caso di AQ, ANDRE, DAS e del sistema dell'Australian Bureau of Statistics. Tale controllo avviene generalmente attraverso filtri sul tipo d'istruzioni che è possibile eseguire. Se invece i dati di base non sono preventivamente protetti, come per Statistics Canada e Statistics Netherlands, il controllo è generalmente di tipo manuale. Anche il canale utilizzato nella comunicazione tra Istituto e utente in qualche modo condiziona il tipo di controllo sull'output; in generale se il sistema utilizza un accesso via internet sia per l'interrogazione del database che per la restituzione dell'output (web-based system- AQ, DAS) allora il controllo dell'output è di tipo automatico e i tempi di ritorno sono generalmente brevissimi (pochi secondi). Se invece il mezzo utilizzato per la comunicazione dell'input è la posta elettronica (mail-based system) il controllo degli output è in parte di tipo automatico (attraverso filtri su programmi di input) e in parte di tipo manuale. Quest'ultimo può essere previsto su tutti i file di output prima della restituzione degli stessi all'utente (Statistics Canada) o su richiesta dell'utente stesso nei casi in cui l'output non soddisfi i filtri di sistema e quindi non venga automaticamente rilasciato (LIS LES PieP).
In alcuni casi all'interno della gamma dei possibili canali di accesso ai microdati uno stesso Istituto può prevedere due diversi accessi in remoto. In generale si diversificano tra loro per quantità e qualità di informazione preservata nel database di interrogazione, per tipo di output rilasciato e per velocità di rilascio dell'output stesso. In questo modo il canale di accesso remoto viene utilizzato per rispondere alle esigenze di due tipologie diverse di utenza: una più "veloce", che richiede informazioni meno dettagliate ma risposte più immediate, l'altra più "esperta", che richiede informazioni molto dettagliate per sostenere progetti di ricerca più ampi. Un'offerta di questo tipo è stata realizzata recentemente all'interno del progetto LIS con la creazione, accanto al classico canale di accesso remoto via mail, del cosiddetto web-tabulator. Altri esempi di Istituti che stanno promuovendo una simile strategia di offerta di accesso a microdati sono il Bureau of Census (AQ, America Fact Finder) e il National Center for Health Statistics (Andrew GUI).

Tab. 1

Paesi	Fase di protezione			Fase di comunicazione		Software, disponibilità file di prova
	Utenza, Contratto e Postazioni di lavoro	**1) Tipo input 2) Protezione dati input**	**1) Tipo output 2) Protezione dati output**	**Fase di accesso**	**Fase di restituzione modalità e tempi**	
Danimarca	-Canale offerto a ricercatori Danesi. -Abilitate solo postazioni appositamente assegnate. -Presentazione progetto di ricerca. -Sottoscrizione contratto a livello di Istituzione.	1) Dati sociali e alcuni dati di impresa. 2) Nessuna protezione dei dati: -le collezioni messe a disposizione soddisfano il criterio del *need to know*.	1) Sia tabelle che modelli. 2) -Tutti gli output inviati all'utente vengono immagazzinati ed eventualmente controllati dalla DRS. -Nessuna protezione dell'output che comunque deve soddisfare regole stabilite nel contratto.	-Accesso attraverso internet (smart card) con assegnazione di account e password. -Postazioni assegnate in quanto il pc deve essere appositamente predisposto.	-L'output viene restituito via mail all'indirizzo prefissato. -tempi di ritorno ogni 5 minuti.	- SAS SPSS STATA.
Australia	-Canale offerto a ricercatori di enti Australiani. -Presentazione progetto di ricerca. - Sottoscrizione contratto a livello di Istituzione.	1) Dati sociali; esiste una lista costantemente aggiornata di CURF disponibili via RADL. 2) Dati protetti con tecniche di data swapping, perturbazione casuale valori, ricodifica ed eliminazione di record.	1) Sia tabelle che modelli. 2) -Controllo automatico dei risultati attraverso filtri sui programmi di analisi (**filtering software for set-up**) -Nel caso di output non rilasciati l'utente può richiedere una verifica manuale.	-Accesso attraverso internet (account e password).	-Risultati su pagine web in tempo reale se soddisfatti i vincoli del sistema ; altrimenti possibilità di richiesta di ulteriore controllo manuale.	- SAS SPSS. -Per i file a partire dal 2003 esistono file di prova con dati simulati.
Canada	- Canale offerto solo ad un ristretto numero di ricercatori. - Presentazione progetto di ricerca. - Sottoscrizione contratto.	1) Solo alcune indagine di tipo sociale rendono disponibile i dati via Accesso remoto (SLID, NPHS). 2) Nessuna protezione dei dati sui quali vengono effettuate le analisi richieste dall'utente.	1) Sia tabelle che modelli anche se c'è una preferenza verso i modelli. 2)Controllo **manuale** dell'output anche se presenti alcuni filtri sui programmi utilizzati.	-Accesso via e-mail (il programma viene inviato via e-mail e successivamente applicato ai dati da personale dell'Istituto).	-I risultati vengono restituiti all'utente via e-mail. - I tempi di restituzione dell'output sono di circa 2 o 3 giorni lavorativi.	- SAS SPSS STATA. -Disponibili file di dati simulati su cui testare i programmi.

Paesi	Fase di protezione			Fase di comunicazione		Software, disponibilità file di prova
	Utenza Contratto e Postazioni di lavoro	**1) Tipo input 2) Protezione dati input**	**1) Tipo output 2) Protezione dati output**	**Fase di accesso**	**Fase di restituzione modalità e tempi**	
Olanda	-Canale offerto solo a ricercatori appositamente selezionati. - Abilitate solo postazioni appositamente assegnate ed identificazioni biometriche dell'utente. - Presentazione progetto di ricerca. - Sottoscrizione contratto.	1) Solo dati sociali 2) Nessuna protezione dei dati sui quali vengono effettuate le analisi richieste dall'utente.	1) Sia tabelle che modelli. 2) Controllo esclusivamente **manuale** dell'output.	-Accesso attraverso internet (smartcard). -Postazioni assegnate -Possibilità di vedere su schermo risultati intermedi anche se non è possibile salvarli sul proprio computer o stamparli.	-I risultati vengono restituiti all'utente via e-mail. - I tempi di restituzione dell'output sono di qualche giorno.	- SAS SPSS STATA.
NCHS: ANDRE	-Presentazione progetto di ricerca. - Sottoscrizione contratto.	1) Dati dell'agenzia. 2) Dati protetti (*restricted data)* con tecniche di: -data swapping perturbazione casuale valori, ricodifica ed eliminazione di record.	1) Sia tabelle che modelli. 2) Controllo automatico dei risultati attraverso filtri sui programmi di analisi per il 90% dei casi. Alcuni casi dubbi vengono analizzati manualmente.	Accesso via e-mail (il programma viene inviato via e-mail e successivamente applicato ai dati da personale dell'Istituto).	I risultati vengono restituiti all'utente via e-mail esclusivamente all'indirizzo indicato al momento del contratto.	- SAS.
NCHS: ANDREW (in fase di sperimentazione)	-Presentazione progetto di ricerca. - Sottoscrizione contratto.	1) Dati sociali. 2) Dati protetti (*restricted data)* con tecniche di data swapping, perturbazione casuale valori, ricodifica ed eliminazione di record.	1) Sia tabelle che modelli. 2) Controllo totalmente automatizzato e potenziato rispetto a quello di ANDRE.	-Accesso anche via web. -Possibilità di utilizzare interfaccia GUI (Graphic User Interface).	I risultati vengono restituiti all'utente via e-mail.	- SAS Spss SUDAN STATA.

Paesi	Fase di protezione			Fase di comunicazione		Software, disponibilità file di prova
	Utenza Contratto e Postazioni di lavoro	**1) Tipo input 2) Protezione dati input**	**1) Tipo output 2) Protezione dati output**	**Fase di accesso**	**Fase di restituzione modalità e tempi**	
BOC: ***Advanced Query System*** **AQ**	-Registrazione gratuita sul portale del BOC il quale assegna login e password.	1) Dati relativi al Censimento 2000. 2) Dati protetti con tecniche di data swapping, perturbazione casuale valori, ricodifica ed eliminazione di record.	1) Esclusivamente tabelle non standard. 2) Controllo totalmente automatizzato (non è previsto intervento umano).	-Accesso via web.	-I risultati vengono restituiti all'utente via web generalmente in tempo reale.	-SAS.
BOC: ***Microdata Analysis System*** (in fase di sperimentazione**)**	-FASE DI SPERIMENTAZIONE.	1) Dati relativi a Current population survey (CPS). 2) Dati protetti con tecniche di data swapping, perturbazione casuale valori, ricodifica ed eliminazione di record.	1) Esclusivamente modelli 2)Controllo totalmente automatizzato (non è previsto intervento umano).	-Accesso via web.	-I risultati vengono restituiti all'utente via web generalmente in tempo reale.	-SAS.
NCES: ***Disclosure Avoidance System*** **DAS**	-Registrazione gratuita sul portale del NCES il quale assegna login e password.	1) Dati dell'agenzia. 2) Dati protetti con tecniche di data swapping, perturbazione casuale valori, ricodifica ed eliminazione di record.	1) Tabelle e matrici di correlazione. 2) Controllo automatico dei risultati attraverso filtri sugli output richiesti.	-Accesso via web.	-I risultati vengono restituiti all'utente via web generalmente in tempo reale.	-Software DAS Specifico.

Paesi	Fase di protezione			Fase di comunicazione		Software, disponibilità file di prova
	Utenza Contratto e Postazioni di lavoro	**1) Tipo input 2) Protezione dati input**	**1) Tipo output 2) Protezione dati output**	**Fase di accesso**	**Fase di restituzione modalità e tempi**	
LIS/LES	-Sottoscrizione regole e breve descrizione del progetto di ricerca con successiva assegnazione di login e password. - Il servizio è offerto gratuitamente ai ricercatori appartenenti ai paesi membri del sistema Lis; in caso contrario il servizio è a pagamento.	1) In LIS si interrogano dati campionari relativi ad indagini sui bilanci familiari mentre in LES dati campionari sulle forze di lavoro relativi a 25 paesi. 2) Dati protetti con tecniche perturbative e di ricodifica.	1) Sia tabelle che modelli 2) Controllo automatico dei risultati attraverso filtri sui programmi di analisi (**filtering software for set-up**) e sulle dimensioni dell'output. Il sistema utilizzato prende nome di LISSY.	-Accesso via mail.	I risultati vengono restituiti all'utente via e-mail.	- SAS SPSS STATA.
LIS/LES Web Tabulator	-Sottoscrizione regole e breve descrizione del progetto di ricerca con successiva assegnazione di login e password.	1) Si possono interrogare i dati relativi al database LIS/LES relativi ai soli paesi che hanno rilasciato l'autorizzazione a tale servizio (10 paesi). 2) Dati protetti con tecniche perturbative e di ricodifica.	1) Solo tabelle. 2) Limitazione sull'insieme delle variabili rese disponibili per la creazione di tabelle.	-Accesso via web	-I risultati vengono restituiti all'utente via web generalmente in tempo reale.	
PIEP	-Sottoscrizione regole e breve descrizione del progetto di ricerca con successiva assegnazione di login e password.	1) In PIEP si possono interrogare i microdati relativi all'indagine ESES relativamente a 6 paesi. 2) Dati protetti con tecniche perturbative e di ricodifica.	1) Solo tabelle. 2) Controllo automatico dei risultati con filtri sui programmi di analisi (**filtering software for set-up**) e sulle dimensioni dell'output. Il sistema PiEP-LISSY differisce dal sistema LISSY in quanto prevede controlli più restrittivi.	-Accesso via mail.	I risultati vengono restituiti all'utente via e-mail.	- SAS SPSS STATA

13. Conclusioni

Negli ultimi anni sempre più Istituti hanno implementato o stanno progettando l'implementazione del canale di "*accesso remoto"* come canale di accesso a microdati. Le motivazioni sono in genere legate alla volontà, da parte degli Istituti e delle agenzie Statistiche, di soddisfare le sempre maggiori richieste da parte del mondo scientifico di file di dati sempre più informativi e più facilmente accessibili.

La diffusione del canale di "accesso remoto" ai microdati è strettamente legata alla sua duttilità di implementazione, ovvero alla facilità di adattarsi alle diverse esigenze che un Istituto o un'Agenzia di Statistica presenta in termini di prodotti da offrire ai propri utenti. Inoltre lo sviluppo di tecnologie informatiche che rendono sempre più sicure le comunicazioni via web, e la crescente utilizzazione di internet stanno rendendo questo sistema di accesso ai dati sempre più diffuso e apprezzato sia dagli Istituti di statistica che dagli utenti stessi.

E' da sottolineare che, comunque, per quanto riguarda l'Istituto, il rendere disponibile i dati tramite tale canale presuppone una notevole mole di lavoro. Infatti, oltre alla parte strettamente informatica relativa alla creazione di un sistema tale da garantire un accesso sicuro tramite internet, esiste una parte strettamente statistica di notevole importanza. E' infatti necessario predisporre dati e metadati di facile consultazione, strumenti che permettano un'analisi preventiva dei dati, data set di prova che permettano agli utenti di verificare i propri programmi di analisi, e, cosa molto onerosa, è necessario definire una strategia di protezione, fondamentale per garantire la riservatezza dei rispondenti. Tale strategia deve essere determinata sulla base delle finalità che l'Istituto vuole raggiungere con l'implementazione del sistema di accesso remoto. Essa coinvolge la definizione del tipo di dati input (es. dati sociali od economici), quella dei dati di output (modelli o tabelle), e la definizione di una serie di regole, determinate sulla base di specifiche tecniche di protezione, che i dati di output devono soddisfare per garantire la riservatezza dei rispondenti.

Riferimenti bibliografici

Andersen, O. (2003). From on-site to remote data access – The revolution of Danish system for access to microdata. *Joint ECE/Eurostat work session on statistical data confidentiality, Luxemburg 7-9 April 2003.*

Capobianchi, A. (2005). Alcune esperienze in ambito internazionale per l'accesso ai dati elementari *Documenti Istat n.8/2005.*

Gambhir, V., Harris, K.W. (2005). ANalytical Data Research by Email and Web (ANDREW). *Proceedings del Joint/Unece Eurostat worksession on Statistical Data Confidentiality (Geneva 2005)*

Hundepool, A., de Wolf P.P. (2005). OnSite@Home: Remote Access at Statistics Netherlands. *Proceedings del Joint/Unece Eurostat worksession on Statistical Data Confidentiality.* (Geneva 2005)

Istat (2004). Metodologie e tecniche di tutela della riservatezza nel rilascio di informazione statistica. *Metodi e Norme n.20.*

Reiter, J. (2003). Model Diagnostics for Remote Access Regression Servers. *Statistics and Computing* ,13,pp.371-380.

Ritchie, F.,(2005). Access to business micordata in the UK: dealing with the irreducible risks. *Proceedings del Joint/Unece Eurostat worksession on Statistical Data Confidentiality*. (Geneva 2005)

Rowland, S., (2003). An Examination of Monitored, Remote Microdata Access System. *Presented at NAS Workshop on Access to Research Data: Assessing Risks on Opportunities*. (16-17 October,2003)
http://www7.nationalacademies.org/cnstat/Rowland_Paper.pdf#search='microdata%20remote%20access%20rules'

Schouten, B., Cigrang, M.(2003). Remote access system for statistical analisys of microdata. *Statistics and Computing,* 13, 381-389.

Söderberg, L.G (Lars-Johan) (2005). Mona-Microdata on-line. Access at Statistics Sweden. *Proceedings del Joint/Unece Eurostat worksession on Statistical Data Confidentiality*. (Geneva 2005).

Steel, P., Reznek, A. (2005). Issues in designing a confidentiality preserving model server. *Proceedings del Joint/Unece Eurostat worksession on Statistical Data Confidentiality (Geneva 2005)*.

Tambay, J.L.,Goldmann, G., Potter, J. (2003). Providing Researcher Access to Data for Analysis at Statistics Canada. *Workshop on Microdata, Stockholm, 21-22 August, 2003*. Disponibile su http://www.micro2122.scb.se/papers.asp

Tranmer, M., Pickles, A., Fieldhouse, E., Elliot, M., Dale, A., Brown,M., Martin, D., Steel, D., Gardiner, C. (2005). The case for small area microdata. *Journal of the Royal Statistical Society, A,*168, part 1, 29-49.

Trewin, D. (24 gennaio 2006). Interim Guidelines-"Managing Statistical Confidentiality and Microdata Access-Principles and guidelines of good practice". *CES Task Force on Confidentiality and Microdata*.