Linee guida
per la qualità dei processi statistici che utilizzano dati amministrativi

Versione 1.1

Agosto 2016

Istat

La stesura del manuale è stata coordinata da G. Brancato

Hanno contribuito alla stesura delle Linee Guida:
Brancato G., Boggia A., Barbalace F., Cerroni F., Cozzi S., Di Bella G., Di Zio M., Filipponi D., Luzi O., Righi P., Scanu M.

Il manuale è stato rivisto da:
Signore M., Di Bella G., Di Consiglio L., Filipponi D., Luzi O., Pallara A., Rapiti F., Simeoni G.

Si ringrazia per la collaborazione il Comitato Qualità.

INDICE

Introduzione

Il crescente uso di dati amministrativi per finalità statistiche pone la necessità di adeguare gli standard e gli strumenti per la valutazione della qualità in modo da tenere conto delle relative peculiarità. Ciò ha orientato verso la stesura delle presenti ***Linee guida per la Qualità dei processi statistici che utilizzano dati amministrativi***, che si propongono di integrare per gli aspetti inerenti l'uso dei dati amministrativi, quelle già prodotte per i processi statistici di tipo rilevazione e già pubblicate sul sito web dell'Istat in italiano[1] e in inglese.

L'obiettivo principale è quello di dotare l'Istituto di un riferimento per la conduzione dei processi che utilizzano dati di fonte amministrativa e di uno strumento per la loro valutazione in un'ottica di audit o di autovalutazione.

Le presenti linee guida sono strutturate in tre parti riguardanti rispettivamente la qualità dell'input, la qualità del processo produttivo che utilizza dati di fonte amministrativa e la qualità dell'output (o di prodotto) e seguono il modello concettuale descritto nel paragrafo 1. La Parte A riguarda i principi e le linee guida relativi all'acquisizione dei dati di input al processo produttivo statistico e la misurazione della loro qualità.

Nella Parte B relativa alla qualità di un tipico processo produttivo statistico che utilizza dati di fonte amministrativa, per ciascuna fase in cui si articola il processo stesso, viene enunciato il principio che rappresenta l'obiettivo da perseguire e vengono fornite le linee guida metodologiche per il suo conseguimento.

La Parte C, qualità di prodotto, riguarda la qualità dell'output derivante da un processo produttivo che ha utilizzato dati amministrativi, circoscrivendo l'attenzione soltanto ai prodotti di natura "macrodati". Come per le precedenti linee guida, questa parte non è organizzata in princìpi, ma ripropone le definizioni Eurostat della qualità, e descrive come l'utilizzo dei dati amministrativi può condizionare il loro significato. Infine, sono fornite le definizioni degli errori che si generano nelle diverse fasi del processo produttivo che utilizza dati di fonte amministrativa.

Il crescente sfruttamento degli archivi amministrativi per finalità statistiche, ha portato molti Istituti Nazionali di Statistica, compreso l'Istat, alla identificazione di strutture permanenti o temporanee (Direzioni, Comitati e Commissioni), strategiche e operative, con diverse funzioni e dotate di diversi livelli decisionali, di supporto alle attività gestionali e metodologiche connesse all'uso di tali dati. Queste funzioni vanno: dalla sensibilizzazione degli enti titolari dei dati amministrativi, al coordinamento e monitoraggio delle variazioni della modulistica, alla ricognizione sulle esigenze di utilizzo interno, alla gestione dei rapporti con gli enti e acquisizione degli archivi, alla documentazione e valutazione della qualità, al monitoraggio sugli utilizzi interni. I principi e le linee guida relativi all'acquisizione centralizzata di archivi amministrativi sono sviluppati all'interno dell'appendice allegata.

Le presenti linee guida sono indirizzate ai responsabili dei processi produttivi statistici, che devono comprendere a fondo gli aspetti relativi alla qualità di un processo e del prodotto statistico che utilizza dati di fonte amministrativa, ma che devono anche avere consapevolezza delle problematiche che l'acquisizione di archivi amministrativi pone per un Istituto Nazionale di Statistica.

Queste linee guida presentano alcuni elementi di sovrapposizione con quelle sviluppate per le rilevazioni statistiche, elementi che è stato valutato utile tenere per dare completezza al volume.

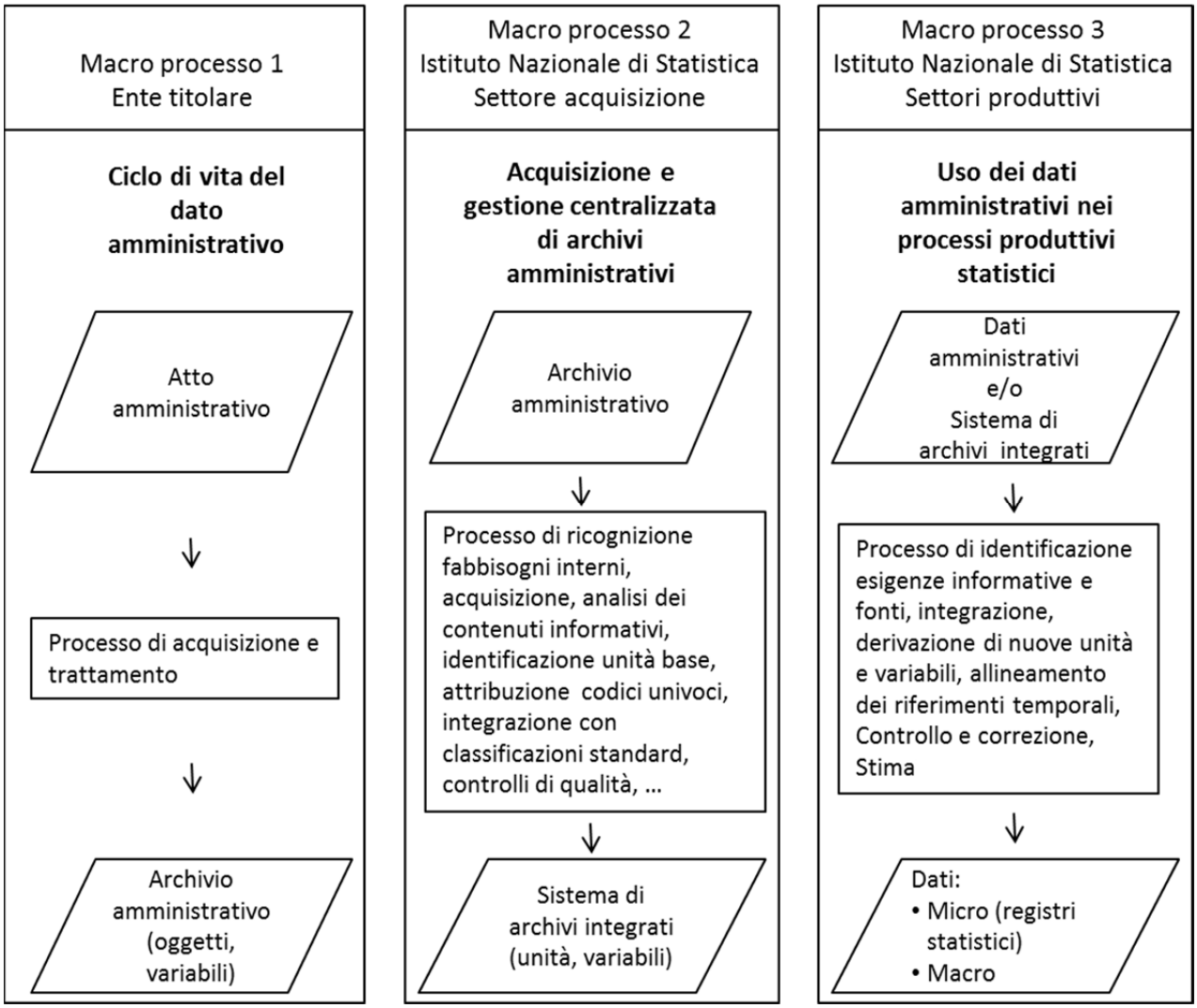
[1] Versione italiana disponibile al seguente link http://www.istat.it/it/files/2010/09/Linee-Guida-Qualit%C3%A0-_v.1.1_IT.pdf

1. Quadro di riferimento per la qualità dei processi statistici che utilizzano dati amministrativi

L'uso dei dati di fonte amministrativa nei processi produttivi statistici introduce specificità che richiedono una revisione nella rappresentazione dei processi produttivi e degli errori che si generano, rispetto ai tradizionali processi produttivi basati sulla rilevazione diretta.

Il modello sottostante queste linee guida identifica tre principali macro processi riguardanti i dati amministrativi, rappresentati sinteticamente nella Figura 1, ove per ogni macro processo sono riportati: il soggetto responsabile, l'input informativo, il processo di trasformazione e l'output.

Figura 1. Macro processi sui dati amministrativi



Il macro processo 1, di responsabilità dell'Ente titolare, può essere assimilato ad una rilevazione di dati e, come tale, i relativi errori possono essere modellati seguendo l'approccio classico di indagine (Groves *et al*., 2004; Zhang L.C., 2012). Questo macro processo è generalmente fuori dal controllo degli Istituti Nazionali di Statistica e non è pertanto specificatamente oggetto di queste linee guida. Tuttavia, è bene sottolineare che gli stessi istituti possono avere un ruolo di coordinamento e di armonizzazione della modulistica, come avviene nel nostro paese, allo scopo di aumentare l'acquisizione di dati amministrativi e la loro utilizzabilità

per la produzione di statistiche. Per tale motivo, spesso in questo ambito più che alla qualità dei dati ci si riferisce alla loro utilizzabilità per finalità statistiche.

Il macro processo 2 riflette una tendenza, riscontrabile in ambito internazionale (si veda per esempio Wallgren & Wallgren, 2014 e Statistics Finland, 2004), verso l'acquisizione centralizzata di archivi amministrativi, orientata alla costruzione di un sistema di archivi integrati o integrabili, che risponda ai fabbisogni interni. Questo si configura come un processo vero e proprio di acquisizione, trattamento e rilascio di un output, che rappresenta un prodotto intermedio di natura statistica, e a sua volta un possibile input per altri processi produttivi statistici (macro processo 3 nella Figura 1). I dati amministrativi acquisiti in questa fase sono sottoposti ad una serie di procedure prevalentemente volte ad identificare oggetti statistici a partire da oggetti amministrativi e a corredarli con le informazioni necessarie per un loro utilizzo successivo. Assumono in questa fase estrema rilevanza gli aspetti di natura gestionale che attengono ai rapporti con gli enti titolari dei dati amministrativi e quelli relativi alla documentazione di supporto all'utilizzo dei dati acquisiti. In relazione a questi elementi, gli indicatori di qualità inclusi nelle iperdimensioni "Fonte" e "Metadati" della checklist sviluppata da Statistics Netherlands (Daas *et al*. 2009) rappresentano un esempio di misurazioni per il monitoraggio e la valutazione della qualità. Principi e linee guida relativi a questo macro processo sono trattati nella Appendice.

Infine, il macro processo 3 riguarda i processi produttivi statistici che utilizzano dati di fonte amministrativa che possono essere stati già centralmente acquisiti e pre-trattati (e quindi derivano dal macro processo 2), oppure possono essere acquisiti direttamente dai titolari del dato amministrativo (e quindi derivano dal macro processo 1). I processi appartenenti alla terza tipologia sono finalizzati alla produzione di statistiche di tipo macro (produzione di stime) o di tipo micro (costruzione di registri statistici). Per ciascun singolo processo di questo tipo, il modello di qualità sottostante identifica:

i) la qualità dell'input (dati amministrativi e/o sistema di archivi integrati nella Figura 1);
ii) la qualità del processo che utilizza dati amministrativi;
iii) la qualità dell'output (micro e macro dati).

Qualità dell'input
Poiché l'input proviene da un processo di raccolta esogeno all'Istituto, è fondamentale il controllo della qualità dei dati di fonte amministrativa acquisiti, alla luce dello specifico obiettivo statistico (cfr. Parte A). Dimensioni della qualità e indicatori di qualità dell'input, in un'ottica già orientata all'output sono stati sviluppati nell'ambito del progetto internazionale Blue – ETS (Daas P., Ossen S., 2011).

Qualità del processo (o qualità del through-put)
Per quello che riguarda la qualità del processo, nella Figura 2 sono rappresentati le fasi del processo stesso, i relativi sottoprocessi, gli oggetti informativi e i potenziali errori sulle unità e sulle variabili. La rappresentazione utilizza gli standard attuali relativi ai metadati, quali il *Generic Statistic Business Process Model* - GSBPM (Unece, 2013a) e il *Generic Statistical Information Model* - GSIM (Unece, 2013b), personalizzandoli per l'uso dei dati amministrativi, mentre per l'identificazione degli errori si è fatto parzialmente riferimento al lavoro di Zhang (2012). In GSIM gli oggetti informativi vengono identificati in forma generica e si caratterizzano in base alla fase del processo di cui sono "output". Per maggiore chiarezza, nello schema gli oggetti generici sono specificati rispetto alla tipologia che assumono. Relativamente agli errori è importante notare che a volte si può verificare una loro propagazione sia in sottoprocessi differenti da quelli in cui si generano sia tra le diverse entità (unità, variabili).

Figura 2. Principali fasi, sottoprocessi, oggetti informativi e potenziali errori per le variabili e unità

Principali fasi, processi e sottoprocessi (GSBPM)	Oggetti Informativi - GSIM (unità)	Errori Potenziali (unità)		Oggetti Informativi - GSIM (variabili)	Errori Potenziali (variabili)
1. Definizione delle esigenze 1.1. Identificazione delle esigenze 1.4. Identificazione dei concetti 1.5. Verifica della disponibilità dei dati 2. Pianificazione	Popolazione (statistica obiettivo) ↓ Dati amministrativi o sistema di archivi integrati osservabili	Non corrispondenza concettuale (sottocopertura, sovracopertura)	Errore di Copertura	Concetti obiettivo ↓ Variabili (rappresentate e/o derivate)	Errori di specificazione Non stabilità concettuale
4. Raccolta* 4.2. Impostazione 4.3. Avvio 4.4. Finalizzazione	Dati amministrativi o sistema di archivi integrati osservabili ↓ Dati amministrativi o sistema di archivi integrati acquisiti	Errori di selezione (missing, duplicazioni, ritardi)		Variabili (rappresentate e/o derivate) ↓ Variabili rilevate**	Errore di misurazione
5. Trattamento 5.1. Integrazione 5.5. Derivazione di nuove unità e variabili 5.9. Allineamento dei riferimenti temporali*** 5.4. Controllo e correzione	Dati amministrativi o sistema di archivi integrati acquisiti ↓ Popolazione (statistica di analisi)	Errori di linkage, Errori di derivazione delle unità		Variabili rilevate ↓ Data set (Microdati validati)	Errore di processo (errori di classificazione, errori di dato mancante, errori di coerenza intra-fonte e inter-fonte, errore di specificazione da modello implicito o esplicito)
5.7. Calcolo di stime				Data set (Microdati validati) ↓ Stime e/o macrodati	Errore di specificazione da modello implicito o esplicito

* Può consistere in una acquisizione presso l'ente titolare oppure in una trasmissione interna da parte di una struttura centralizzata che acquisisce e pre-tratta

** Si è utilizzata la traduzione variabili rilevate per il termine inglese *instance variable*.

*** Questo sottoprocesso non è presente in GSBPM.

Una delle prime attività da considerarsi nell'utilizzare i dati amministrativi per specifici scopi statistici, è l'attenta analisi della corrispondenza tra concetti amministrativi e obiettivi statistici di misurazione, insieme alla pianificazione degli aspetti collegati all'acquisizione dell'archivio o del sottoinsieme dei dati di interesse (sottoprocessi 1 e 2 nella Figura 2). La raccolta dei dati si caratterizza in modo diverso se si assume un procedimento di trasmissione interno all'Istituto oppure un'acquisizione diretta dall'ente titolare del dato amministrativo. La non corrispondenza concettuale con l'obiettivo statistico (in termini di popolazione e di oggetto d'osservazione) limita la validità e l'utilizzabilità per le finalità statistiche del dato amministrativo. Relativamente alle colonne sulle unità, si osserva che la ricerca della popolazione di interesse statistico potrebbe condurre all'acquisizione di popolazioni derivabili da più archivi amministrativi.

Passando alle colonne sulle variabili nella Figura 2, quando i dati di fonte amministrativa non corrispondono ai concetti sottostanti le variabili obiettivo, si parla di errore di specificazione, così come per le indagini con questionario (Biemer e Lyberg, 2003). Inoltre, eventuali variazioni della legislazione o delle procedure che regolano l'atto amministrativo nel tempo o sul territorio limitano la confrontabilità temporale o territoriale dei concetti e di conseguenza la comparabilità delle serie storiche dei dati.

Se i dati amministrativi sono utilizzati per sostituire in parte o completamente le unità da rilevare mediante indagine diretta, si può ipotizzare una fase assimilabile a quella di raccolta/acquisizione durante la quale i concetti sottostanti all'archivio amministrativo diventano dati osservati e interpretati in termini di variabili statistiche. Così come nelle rilevazioni dirette l'osservazione può generare errori di misura, questo stesso tipo di errore può presentarsi nei dati amministrativi utilizzati per finalità statistiche (colonna relativa alle variabili nella Figura 2). Sul fronte delle unità, l'acquisizione dei dati è legata ad un potenziale errore, cui è attribuito il nome di errore di selezione. Si tratta tuttavia di un errore che di fatto incide sulla copertura, infatti eventuali problemi nell'acquisizione dei dati ed eventuali ritardi di aggiornamento dell'archivio, possono portare a errori di selezione e conseguentemente a errori di copertura.

La maggior parte dei tipi di utilizzo (costruzione di registri statistici, integrazione di dati di indagine relativamente a sottopopolazioni o a insiemi di variabili) prevede, all'interno del processo "5.Trattamento dei dati", un sottoprocesso di integrazione, che può realizzarsi attraverso procedure di record linkage o procedure di matching. In tale fase possono generarsi, a livello di unità, errori quali falsi abbinamenti e falsi non abbinamenti. I primi possono a loro volta generare sottocopertura nelle unità e errori di coerenza inter-fonte nelle variabili, per esempio dovuti all'attribuzione errata di variabili relative ad una unità, ad unità diverse. I falsi non abbinamenti, invece, possono dare luogo a sovracopertura e, se l'integrazione è finalizzata a colmare valori nelle variabili, anche a dati mancanti sulle variabili.

Nell'attività di derivazione di nuove unità si parla di possibile errore di derivazione che può riguardare l'unità stessa o le sue relazioni con altre unità. Errori di derivazione possono causare errori di copertura. Le attività di "Derivazione di nuove variabili" e di "Allineamento dei riferimenti temporali" (sottoprocessi 5.5. e 5.9.) riguardano le possibili ricodifiche di tipo semplice o ricostruzioni più complesse effettuate sulle variabili e possono causare errori che sono stati denominati "di classificazione". Un'attenzione particolare è richiesta per minimizzare i potenziali errori che si possono generare in fase di trasformazione necessaria per riportare il riferimento temporale dei dati a quello obiettivo.

Il processo di controllo e correzione ha l'obiettivo di migliorare la qualità dei dati e di risolvere problemi di coerenza interna alle fonti e, in genere, non costituisce una fonte rilevante di errore. Tuttavia è ipotizzabile la generazione in questa fase di un potenziale errore di controllo e correzione dovuto ad un'erronea specificazione delle regole, l'utilizzo di un modello inappropriato o ad errori nelle attività di correzione condotte manualmente.

A conclusione della fase di trattamento viene ricostruita la popolazione di interesse che se confrontata con quella obiettivo, consentirebbe la valutazione finale della copertura. Tuttavia è opportuno specificare che la popolazione obiettivo è raramente disponibile.

Va comunque sottolineato che nella pratica corrente non è sempre facile distinguere le singole fonti di errore attribuibili ai diversi sottoprocessi di integrazione, derivazione delle unità e delle variabili e di controllo e correzione, in quanto essi spesso coinvolgono procedure che sono condotte contestualmente.

Infine, durante la fase di stima gli errori più rilevanti potranno essere quelli da assunzione del modello in quanto i processi produttivi statistici che utilizzano dati amministrativi fanno ampio ricorso all'applicazione di modelli.

I principi e le linee guida per limitare e controllare gli errori che si generano durante il processo produttivo statistico sono oggetto della Parte B di queste linee guida.

Qualità dell'output
Relativamente alla qualità del prodotto di tipo macro (o output), l'approccio seguito si basa sull'adozione delle dimensioni della qualità stabilite in ambito europeo (Eurostat, 2009) e adottate dall'Istat (Parte C).

Relativamente all'output di tipo micro, il modello qui sviluppato prevede la valutazione della qualità durante il processo produttivo statistico, e in particolare dopo la fase di validazione, e in occasione dell'archiviazione del file dei microdati validati, aspetto trattato negli ultimi principi della Parte B.

Alcuni riferimenti bibliografici

Biemer P.P., Lyberg L. (2003). Introduction to survey quality. Wiley, New York.

Groves R M, Fowler F.J.Jr, Couper M, Lepkowsky J.M, Singer E., Tourangeau R. (2004). Survey Methodology. Wiley, New York.

Statistics Finland (2004). Use of Register and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland.

Unece (2013a). The Generic Statistical Business Process Model GSBPM v5.0 http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

Unece (2013b). Generic Statistical Information Model GSIM v1.1. http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model

Wallgren A. and Wallgren B. (2014). Register-based Statistics: Administrative Data for Statistical Purposes. Second Edition. John Wiley & Sons, Chichester, UK.

Zhang L.C. (2012). Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica, Vol 66, nr.1, pp. 41-63.

Parte A. Qualità dei dati di input (*Input quality*)

Principio A.1. Acquisizione di dati amministrativi

I dati di fonte amministrativa necessari al processo produttivo statistico dovrebbero essere direttamente ottenuti dalle strutture dell'Istituto preposte all'acquisizione centralizzata, quando siano già nella disponibilità di tali strutture. Se i dati amministrativi da utilizzare non rientrano nel piano di acquisizione centralizzata, allora la loro acquisizione presso gli enti titolari dovrebbe seguire il più possibile le procedure standard in uso all'Istituto.

Linee guida

In fase di pianificazione, le strutture di produzione statistica rendono manifeste le proprie esigenze di dati amministrativi presso la struttura preposta all'acquisizione centralizzata. Quest'ultima provvede a raccogliere e analizzare tali richieste, stabilendo nel caso le priorità. Infatti, non sempre si è nella condizione di acquisire centralmente tutti gli archivi amministrativi segnalati, ma può essere necessario dover concentrare le risorse su un sottoinsieme di archivi di rilevanza strategica.

Nel caso in cui i dati amministrativi necessari al processo produttivo statistico siano già a disposizione dell'Istituto, essi devono essere acquisiti esclusivamente attraverso trasmissione interna. A tale scopo, deve rigorosamente essere seguito l'iter di riferimento, in genere espresso attraverso protocolli/accordi interni, osservando le opportune regole in materia di sicurezza e privacy.

Se i dati amministrativi necessari al processo produttivo statistico non rientrano nella pianificazione delle acquisizioni centralizzate, ma è consentita la loro acquisizione direttamente dall'ente titolare, è opportuno seguire per quanto possibile le procedure standard di acquisizione adottate dall'Istituto. Nel dettaglio assumono particolare rilevanza i seguenti elementi tra quelli estensivamente definiti per l'acquisizione centralizzata (si veda Principio 3 dell'Appendice):

- identificare un referente presso l'ente amministrativo per il trasferimento dei dati;
- formalizzare accordi che fissino i tempi di trasmissione dei dati, i livelli attesi di qualità dell'archivio, la documentazione di supporto alla trasmissione dell'archivio;
- seguire le regole e le procedure che garantiscono le modalità di trasmissione sicura e trattamento di dati sensibili e la prevenzione del rischio di violazione della riservatezza.

Più in generale è opportuno stabilire e mantenere buoni rapporti con l'ente fornitore dell'archivio amministrativo e collaborare con l'ente per il miglioramento continuo della qualità dei dati attraverso il ritorno dell'informazione all'ente stesso.

Alcuni riferimenti bibliografici

Statistics Canada (2009). *Statistics Canada Quality Guidelines*, Fifth Edition – October 2009

Statistics Canada (2009). *Statistics Canada, Use of administrative data (website)* http://www.statcan.gc.ca/pub/12-539-x/2009001/administrative-administratives-eng.htm (ultimo accesso: Dicembre 2013)

United Nations Economic Commission for Europe - Unece (2011). *Using administrative and secondary source for official statistics: a handbook of principles and practice*. United Nations, New York and Geneva, 2011.

Principio A.2. Valutazione della qualità dei dati di input

La qualità dei dati di input, sia se acquisiti dalla struttura centralizzata che direttamente dall'ente titolare dei dati amministrativi, deve essere misurata e valutata in funzione dello specifico obiettivo statistico, prima del loro trattamento e integrazione nel processo produttivo statistico.

Linee guida

Il primo controllo della qualità dei dati, rispetto all'obiettivo statistico prefissato, deve essere effettuato sui dati di input, che siano essi stati acquisiti dall'ente titolare o trasmessi dalla struttura interna.

Gli elementi di qualità da prendere in considerazione sono molteplici e solo per alcuni sarà possibile calcolare delle misurazioni quantitative.

Alcuni aspetti che possono influire sulla qualità dei dati di input attengono al contesto amministrativo, per esempio in relazione alla normativa e alla stabilità delle procedure che ne regolano la produzione (legislazione, modulistica, ...) nel tempo e sul territorio. Possono avere impatto sulla qualità anche le procedure di registrazione degli eventi amministrativi da parte dell'ente titolare e la tempistica di fornitura dell'archivio all'Istat, nonché la qualità della documentazione fornita a supporto del dato amministrativo.

La copertura dell'archivio o degli archivi amministrativi (o dei data set estratti da archivi acquisiti centralmente) deve essere misurata rispetto alle specifiche popolazioni statistiche obiettivo di ciascun singolo archivio.

Rispetto alle variabili contenute negli archivi o nei data set estratti da archivi acquisiti centralmente, deve essere condotta in primo luogo un'analisi concettuale sulla corrispondenza con le variabili obiettivo statistico, per prevenire errori di specificazione. In secondo luogo sarà opportuno calcolare misurazioni di qualità sui dati, ad esempio l'entità dei dati mancanti nelle variabili di interesse.

Nel caso di dati acquisiti dalla struttura centralizzata preposta all'acquisizione, alcuni indicatori di interesse per l'obiettivo statistico specifico potrebbero essere già calcolati e resi disponibili.

Alcuni indicatori per la qualità dei dati amministrativi di input in un'ottica già orientata all'output sono disponibili in Daas e Ossen S. (2011).

Alcuni riferimenti bibliografici

Daas P., Ossen S., Vis-Visschers R., and Arends-Tóth J. (2009). *Checklist for the Quality evaluation of Administrative Data Sources*, Statistics Netherlands, The Hague /Heerlen, 2009

Daas P., Ossen S. (2011). *Report on methods preferred for the quality indicators of administrative data sources*, Blue – ETS Project, Deliverable 4.2.

Parte B. Qualità del processo (*Through-put quality)*

Principio B.1. Esigenze informative e scelta delle fonti amministrative

Le esigenze informative da soddisfare devono essere ben definite. Ogni decisione sull'uso dei dati amministrativi, in relazione agli obiettivi statistici individuati, deve essere preceduta da una valutazione generale delle caratteristiche e della qualità dei dati contenuti nella fonte. Nel caso di disponibilità di più fonti, la scelta deve essere condotta in base ad un'analisi comparata.

Linee Guida

L'identificazione delle esigenze informative richiede, così come comunemente avviene per le indagini dirette, la conoscenza approfondita degli utenti e dei loro bisogni informativi, conoscenza da acquisire attraverso l'istituzione di tavoli di confronto utenti-produttore che abbiano carattere continuativo e stabile. In generale, gli utenti sono molteplici e spesso portatori di interessi contrastanti: è quindi importante non solo conoscerne le diverse tipologie, ma anche essere in grado di associare loro un diverso grado di importanza rispetto ai risultati del processo. Di conseguenza i principali utenti devono essere chiaramente identificati, coinvolti nella definizione degli obiettivi e nella (ri)progettazione del processo e ne deve essere misurata la soddisfazione, con livelli diversi di formalizzazione e di coinvolgimento. È utile tenere una documentazione aggiornata dei principali utenti e delle loro caratteristiche.

Gli obiettivi conoscitivi e gli utilizzi dei dati amministrativi che si prevedono all'interno del processo produttivo statistico (produzione diretta, costruzione di registri statistici, supporto alla qualità dei processi di tipo indagine) devono essere identificati e pianificati in anticipo al fine di individuare i requisiti di qualità dei dati amministrativi e le metodologie di trattamento più idonee.

Nel caso di produzione diretta di statistiche, mediante la sostituzione di unità e/o variabili di indagine con dati amministrativi assumono particolare importanza: la popolazione e unità statistiche obiettivo, le variabili e le classificazioni di interesse, le dimensioni territoriale e temporale. Rispetto a tali elementi sarà opportuno procedere ad un'analisi oggettiva, prima prevalentemente concettuale poi maggiormente orientata ai dati, sull'utilizzabilità dei dati amministrativi nel processo produttivo statistico, per le finalità specifiche.
In primo luogo, a partire dalla definizione della popolazione statistica obiettivo, e cioè dalla specificazione delle unità che la compongono, dei riferimenti temporali e della delimitazione geografica (per esempio, popolazione residente in Italia ad una certa data), è opportuno analizzare attentamente la capacità degli archivi amministrativi a disposizione di cogliere completamente le unità della popolazione. Analogamente, dovrà essere valutata la corrispondenza tra concetti e definizioni che riguardano le variabili statistiche d'interesse e quelli relativi alle variabili desumibili dai dati amministrativi. In questo ambito, gli aspetti di qualità su cui concentrare l'attenzione riguardano i livelli di copertura attesi per la popolazione di interesse, la validità delle variabili utilizzate e quindi la valutazione di possibili distorsioni derivanti da un uso inappropriato dei dati amministrativi. Questi aspetti saranno ulteriormente approfonditi nei successivi principi B3 e B4.

Nel caso di costruzione di registri statistici, quando più fonti amministrative sono utilizzate congiuntamente al fine di assicurare la maggiore copertura possibile della popolazione di interesse, assume particolare rilevanza la dimensione di qualità relativa all'integrabilità dei dati, che deve essere valutata in primo luogo rispetto all'esistenza e qualità di chiavi di collegamento e\o codici identificativi univoci, e quindi in relazione agli errori che si possono generare nelle procedure di *record linkage* (si veda Principio B2 sull'Integrazione).

Quando i dati di fonte amministrativa sono utilizzati anche a supporto del processo produttivo statistico, come: liste di estrazione, fonte di informazioni per migliorare l'efficienza del disegno campionario e della fase di stima, ausilio per la fase di controllo e correzione, valutazione della qualità e di specifiche fonti di errore (per esempio nella fase di validazione dei risultati o per la stima degli errori di copertura), la loro qualità ha un impatto importante sulla qualità dei dati prodotti e, a seconda dello specifico utilizzo potranno avere maggiore rilevanza aspetti legati alla tempestività dei dati amministrativi, all'accuratezza delle informazioni contenute, alla coerenza e comparabilità, alla integrabilità degli archivi.

In generale, per statistiche e/o registri statistici da produrre con regolarità e continuità, assumono particolare rilevanza gli aspetti relativi alla frequenza, tempestività e stabilità dei dati di fonte amministrativa utilizzati.

Laddove vi sia disponibilità di più fonti di dati amministrativi utilizzabili, salvo il caso in cui non si scelga di acquisirle tutte per integrarle, la scelta della fonte migliore deve essere condotta sulla base di un'analisi comparata sui vantaggi e gli svantaggi dell'utilizzo di una fonte rispetto all'altra e sul possibile impatto in termini di qualità dei dati prodotti. In tal senso può essere utile condurre analisi di scenario, e soprattutto valutare attentamente i rischi connessi alle situazioni di indisponibilità della fonte, identificando strategie alternative per garantire la produzione statistica.

La valutazione sulla qualità dei dati di fonte amministrativa da inserire nel processo produttivo e l'analisi comparata tra le fonti deve essere basata sull'insieme degli indicatori sulla qualità dell'input (cfr. Principio A2 and Daas P. e Ossen S., 2011).

Le ipotesi e le motivazioni che hanno portato alla scelta dei dati di fonte amministrative da utilizzare e il tipo di utilizzo all'interno del processo produttivo statistico devono essere opportunamente documentate.

Alcuni riferimenti bibliografici

Daas P., Ossen S. (2011). Report on methods preferred for the quality indicators of administrative data sources, Blue – ETS Project, Deliverable 4.2.

Lavallée, P. (2000). "Combining Survey And Administrative Data: Discussion Paper." ICES-II, Proceedings of the Second International Conference on Establishment Surveys, Survey Methods for Businesses, Farms, and Institutions. Buffalo, New York. June 17-21, 2000. p. 841-844.

Statistics Canada (2009). Statistics Canada Quality Guidelines, Fifth edition (Chapter on the Use of Administrative Data) http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.htm

Principio B.2. Metodi per l'integrazione dei dati
L'integrazione tra le unità contenute in archivi amministrativi deve essere condotta in accordo con gli obiettivi conoscitivi e/o produttivi e deve essere basata su metodologie consolidate e condivise. La procedura di integrazione deve essere definita con chiarezza, evidenziandone tutte le ipotesi. Tutte le fasi devono essere testate e documentate. La validità dei risultati del processo di integrazione deve essere valutata e documentata calcolando opportuni indicatori.

Linee Guida

Nell'ambito dell'uso di dati amministrativi per la produzione statistica, l'integrazione può avere obiettivi e caratteristiche diverse. Semplificando l'integrazione può essere finalizzata:

- alla costruzione di un archivio di microdati utilizzabile per la produzione diretta di dati statistici e/o come lista di riferimento;
- per completare popolazioni e/o variabili di indagine.

L'integrazione può avvenire con modalità diverse. Quando le fonti condividono una chiave identificativa che si possa assumere registrata senza errori, allora si ricorre principalmente a tecniche di *matching* esatto. Quando tale chiave identificativa non è presente, in genere si ricorre alla selezione di un insieme di variabili che congiuntamente consentono di identificare il più univocamente possibile le unità (nome, cognome, indirizzo, …) e si ricorre all'applicazione di procedure di *record linkage*, basate sul confronto di queste chiavi nelle diverse fonti che si vogliono integrare. La presenza di possibili errori nelle chiavi identificative aumenta la possibilità di generare errori durante questa fase così rilevante nell'utilizzo dei dati amministrativi.

Sia essa tra fonti amministrative, che tra fonti amministrative e dati di indagine, la procedura di *record linkage* deve seguire una serie di passi, di seguito delineati con i necessari riferimenti bibliografici per un approfondimento (per una descrizione dettagliata, anche se non completa degli ultimi sviluppi, si veda Scanu, 2003 o Herzog *et al*., 2007; per una trattazione più legata agli aspetti di *information technology*, si veda Batini e Scannapieco, 2006). Tutti i passi eseguiti e le ipotesi sottostanti devono essere adeguatamente documentati. Inoltre, il processo di integrazione deve avvenire nel rispetto delle normative per la tutela della riservatezza.

Scelta del metodo deterministico o probabilistico. Chi conduce il *record linkage* trova davanti a sé due strategie alternative: quella del metodo deterministico e quella del metodo probabilistico. Nel suo complesso, la ricostruzione di un data set integrato tramite *record linkage* può essere ottenuta applicando successivamente procedure di *record linkage* diverse, compreso quello di partire con un approccio deterministico e poi recuperare le coppie di record più difficili da abbinare con una procedura probabilistica. I due approcci, deterministico e probabilistico, sono sostanzialmente differenti.
Metodo deterministico: il meccanismo del *record linkage* deterministico è molto semplice e si basa sulla definizione di regole che stabiliscono in modo univoco, sulla base del confronto tra le variabili chiave, quali coppie di record vengono considerate dei *match* e quali no. Un esempio di *record linkage* deterministico molto semplice e frequentemente applicato consiste nel dichiarare come *match* quelle coppie di record che hanno lo stesso codice fiscale. Se si basa l'intero *record linkage* su questa regola, deve essere chiara l'assunzione che la variabile (o le variabili) usata per definire la regola viene considerata priva di errore.
Metodo probabilistico: se nel metodo deterministico chi sta conducendo il *record linkage* sceglie le regole e quindi la determinazione delle coppie che sono dichiarate *match*, nel caso del *record linkage* probabilistico si tenta di stimare qual è la probabilità che sulla base dei confronti delle variabili chiave, due record facciano riferimento alla stessa unità (sono dei *match*) o meno (non sono dei *match*). Ciò corrisponde ad assumere che i risultati dei confronti fra le variabili di match siano derivanti da due modelli distinti: uno per le coppie che

sono dei *match* e l'altro per le coppie che non sono dei *match*. Questi due modelli sono molto differenti fra loro e tenderanno ad assumere valori su spazi disgiunti al diminuire delle quantità di errori sulle variabili di *match*. Al limite, quando le variabili di *match* non sono affette da errore, il modello dei confronti per i *match* assumerà con probabilità 1 l'uguaglianza dei valori delle variabili di *match*, mentre per i non *match* lo stesso valore congiuntamente per tutte le variabili di *match* viene assunto con probabilità zero. Se invece si ammette la possibilità di errori nelle variabili di *match*, i due modelli tenderanno a sovrapporsi e questo sarà la causa di possibili errori: falsi *match* e falsi non *match*. L'obiettivo delle procedure probabilistiche di *record linkage* consiste nel determinare un insieme di coppie dichiarate come *match* tenendo sotto controllo (definendo a priori) la possibilità dei due tipi di errore prima descritti (falsi *match* e falsi non *match*) e in modo tale che il ricorso a tecniche costose per capire lo status di quelle coppie per le quali non è semplice prendere una decisione sia il più possibile minimo. Il modello di riferimento dei dati e la procedura sono state definite in Fellegi e Sunter (1969). La decisione se la coppia sia un *match* o meno viene assunta sulla base del test rapporto delle verosimiglianze, ossia tra la verosimiglianza che la coppia di record sia un *match*, dato i confronti osservati, e la verosimiglianza di essere un non *match* sempre condizionatamente al risultato del confronto che si è osservato. Quanto maggiore è il rapporto, chiamato spesso "peso", tanto più è verosimile che la coppia sia un *match*, mentre valori bassi del rapporto suggeriscono di favorire l'ipotesi che la coppia sia un non *match*. Valori intermedi indicano situazioni di difficile valutazione che di solito devono essere sottoposti ad un esame manuale da parte di operatori opportunamente addestrati. Generalmente l'individuazione di *match*, non *match* e casi dubbi avviene in seguito alla definizione di due soglie, una superiore e una inferiore; le coppie con peso più grande della soglia superiore sono considerate *match*, quelle con peso più piccolo della soglia inferiore sono non *match* mentre le coppie con peso compreso tra le due soglie sono destinate a una revisione manuale, attività che può essere molto costosa. La definizione delle coppie è legata ai limiti di tolleranza fissati per le probabilità di commettere un errore: ossia considerare un *match* una coppia che non lo è (errato *link*) e, viceversa non considerare come *match*, una coppia che invece lo è (errato non *link*).

Scelta della variabile di abbinamento. La selezione delle variabili di abbinamento deve essere svolta in modo che le variabili selezionate siano congiuntamente in grado di identificare le unità della popolazione di interesse. Si raccomanda di considerare variabili possibilmente non affette da errori, mancate risposte, mancanza di stabilità nel corso del tempo, problemi legati alla privacy (National Statistics 2004a, 2004b). Per i casi di *record linkage* relativi a popolazioni di famiglie o individui, Gill (2001) fornisce dei suggerimenti validi nel caso di *linkage* di data set socio-demografici. Statistics New Zealand (2006) suggerisce di evitare di scegliere come variabili di *matching* variabili altamente correlate o dipendenti. L'obiettivo finale delle variabili di *matching* consiste comunque nel discriminare nel modo più efficace possibile quelle coppie di record che sono *match* da quelle che non sono dei *match*. Questo avviene tanto più quanto maggiore è il numero di categorie, nel caso di variabili qualitative: casi speciali si hanno quando solo alcune delle modalità di una variabile sono altamente discriminanti, si veda in proposito Winkler (1989, 1993, 1995, 2000).

Scelta di eventuali operazioni di blocking e di ordinamento dei file. Le procedure di *record linkage* mettono a confronto i record disponibili su due archivi, il cui numero è il prodotto delle dimensioni degli archivi stessi. Questo numero può essere troppo elevato sia per motivi computazionali che statistici (Scanu, 2003), e in tal caso, prima di effettuare il *record linkage*, si partizionano i due file secondo una o più variabili categoriali considerate particolarmente affidabili e si limita il confronto ai record appartenenti alla stessa partizione nei due data set (Baxter *et al.,* 2003). L'elevata qualità delle variabili di blocco evita la perdita di potenziali *match* in disaccordo sulla variabile di blocco stessa. Se variabili di elevata qualità non sono disponibili, si può ricorrere al metodo del *sorted neighborhood* ordinando le osservazioni in funzione di una o più variabili (ad esempio con l'ordine alfabetico) e confrontando gruppi di unità che occupano posizioni equivalenti negli ordinamenti dei due file (si veda Batini e Scannapieco, 2006).

Scelta della funzione di confronto. La scelta delle variabili di *matching* viene fatta fra le variabili comuni ai due file, dopo aver tolto le variabili scelte per le operazioni di *blocking* e ordinamento. Queste variabili vengono scelte in base al loro potere predittivo sullo status della coppia di record come *match* o non *match*, ossia devono essere tali che le coppie di record che presentano poche differenze nei valori di queste variabili tendono ad essere considerate come dei *match*, mentre al crescere delle differenze fra i valori osservati dei record si è più sicuri nell'affermare che la coppia di record sia un non *match*. Esistono diversi tipi di funzioni di confronto, con l'obiettivo di esaltare alcuni tipi di risultati nei confronti fra le variabili di *match,* in proposito si veda Gu *et al.* (2003). In genere la funzione di confronto maggiormente usata è quella che verifica l'uguaglianza (1) o la diversità (0) del valore della variabile di *matching* nei due record posti a confronto.

Modifica dei risultati al fine di soddisfare eventuali vincoli (abbinamento 1:1). Sia la procedura deterministica che quella probabilistica giungono alla definizione di uno status (*match* o non *match*) per ogni singola coppia di record, a prescindere dalle altre. Questa procedura può portare a risultati in contraddizione con alcune ipotesi: ad esempio si può ipotizzare che ogni record di un data set può agganciarsi al più con un record di un secondo data set, mentre la procedura di *record linkage* può portare alla definizione di coppie di record che condividono la stessa unità. Per risolvere questo problema si può utilizzare l'algoritmo del trasporto (Schrijver, 2003) cercando di massimizzare la somma dei pesi delle coppie che vengono definitivamente dichiarate *match* sotto il vincolo che un record di un data set può abbinarsi al massimo a un record dell'altro data set.

Valutazione della qualità. La procedura di integrazione, sia essa deterministica o probabilistica o mista, dovrebbe essere valutata attraverso la stima di indicatori sulle due principali tipologie di errore: falsi *link* e falsi non *link*. Spesso non si hanno a disposizione strumenti per condurre tale valutazione, ossia fonti di confronto più accurate rispetto alle quali verificare la correttezza del *match*, e si deve ricorrere a un'analisi manuale dei risultati ottenuti, almeno su un campione di coppie. Un altro aspetto che influisce sulla validità della procedura riguarda le ipotesi assunte: le regole applicate nel caso di procedure deterministiche e le ipotesi per la costruzione della funzione di verosimiglianza nel caso di procedure probabilistiche, ipotesi spesso non adatte ai dati disponibili (ad esempio non è vero che il confronto di una variabile chiave sia una variabile indipendente e identicamente distribuita sull'insieme delle coppie disponibili). Si veda comunque il lavoro di Belin e Rubin (1995) per una discussione su questi aspetti. Negli ultimi anni Tancredi e Liseo (2011) hanno definito un metodo *bayesiano* di *record linkage* che tende a risolvere le criticità relative all'adeguatezza delle ipotesi sottostanti il metodo di Fellegi e Sunter.

Alcuni riferimenti bibliografici

Batini C., Scannapieco, M. (2006). Data Quality: Concepts, Methodologies and Techniques. Springer Verlag, Heidelberg.

Baxter R., Christen P. and Churches T, " A comparison of fast blocking methods for record linkage", Proceedings of 9th ACM SIGKDD Workshop on data cleaning, Record linkage and Object Consolidation, 2003

Belin T.R, Rubin D.B. (1995). A Method for Calibrating False-Match Rates in Record Linkage. Journal of the American Statistical Association, 90, 694-707.

Fellegi, I. P., and A. B. Sunter (1969). A theory for record linkage. Journal of the American Statistical Association, Volume 64, pp. 1183-1210.

Gill L. (2001). Methods for automatic record matching and linkage and their use in national statistics, National Statistics Methodological Series No. 25, London (HMSO)

Gu L., Baxter R., Vickers D., and Rainsford C. (2003). Record linkage: Current practice and future directions. Technical Report 03/83, CSIRO Mathematical and Information Sciences, Canberra, Australia, April 2003.

Herzog, T.N. Scheuren, F.J. a Winkler, W.E. (2007). Data Quality and Record Linkage Tehniques. Springer Science+Business Media, New York.

National Statistics (2004a) National Statistics code of practice – protocol on Data Matching. Office for National Statistics, London.

National Statistics (2004b) National Statistics code of practice – protocol on Statistical Integration. Office for National Statistics, London.

Scanu, M. (2003). Metodi statistici per il record linkage. Istat, Collana Metodi e Norme, n. 13.

Schrijver, Alexander (2003). Combinatorial Optimization - Polyhedra and Efficiency. Springer Verlag

Statistics New Zealand (2006). Data integration manual; Statistics New Zealand publication, Wellington, August 2006.

Tancredi A., Liseo B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. Ann. Appl. Stat, Volume 5, Number 2B (2011), 1127-1698

Thibaudeau, Y. (1993). The discrimination power of dependency structures in record linkage. Survey Methodology, Volume 19, pp. 31-38.

Winkler W.E. (1989). Frequency-based matching in the Fellegi-Sunter model of record linkage, Proceedings of the Section on Survey Research Methods, American Statistical Association, 778-783.

Winkler, W.E. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage. Proceedings of the Survey Research Methods Section, American Statistical Association, 274-279.

Winkler, W.E. (1995). Matching and record linkage. Business Survey Methods, Cox, Binder, Chinappa, Christianson, Colledge, Kott (eds.). John Wiley & Sons, New York.

Winkler W. E. (2000). Frequency based matching in the Fellegi-Sunter model of record linkage (long version). Statistical Research Report Series no. RR 2000/06, U.S. Bureau of the Census. At http://www.census.gov/srd/papers/pdf/rr2000-06.pdf (last view, 21/04/2008).

Principio B.3. Individuazione e derivazione delle unità e valutazione della copertura
Il procedimento di individuazione e derivazione delle unità statistiche deve seguire pratiche consolidate. Tutte le ipotesi devono essere esplicitate e i passaggi devono essere documentati. La qualità in termini di copertura deve essere opportunamente valutata.

Linee Guida

L'individuazione e la ricostruzione dell'unità statistica è una problematica raramente menzionata nella letteratura classica relativa alle tecniche d'indagine. Ciò trova giustificazione nel fatto che nella progettazione di un'indagine la popolazione, l'unità statistica e l'unità di rilevazione trovano corrispondenza nell'osservazione pianificata. Questo non è più necessariamente valido nel momento in cui la popolazione obiettivo viene osservata da fonti di natura amministrativa. In tal caso, le unità statistiche di interesse non sempre esistono come tali negli archivi amministrativi e la popolazione amministrativa non sempre coincide con quella statistica.

É necessario, quindi, in particolare in fase di (ri)progettazione: lo studio degli oggetti contenuti nell'archivio amministrativo e delle loro relazioni con le unità che sono rilevanti a fini statistici, nonché la valutazione della rappresentatività della popolazione statistica da parte di quella amministrativa.

Individuazione e derivazione delle unità. Gli oggetti di un archivio amministrativo possono essere eventi o unità amministrative e la loro relazione con le unità statistiche non è sempre di immediata individuazione.
Le unità statistiche vengono create per derivazione[2] dagli oggetti amministrativi (Wallgren & Wallgren, 2014) per mezzo di una funzione di trasformazione, che consente di allineare il dato amministrativo a quello statistico prima a livello di metadati (tramite confronto e raccordo delle definizioni) e poi a livello di dati attraverso l'esplicitazione del trattamento cui sottoporre il dato amministrativo per poterlo usare a fini statistici. Purtroppo, vista l'eterogeneità degli archivi amministrativi, spesso il loro trattamento non segue metodi e procedure standardizzate, ma è specifico per l'obiettivo statistico e la fonte acquisita.
È importante che nel processo di derivazione delle nuove unità si valuti attentamente l'applicabilità delle tecniche disponibili in letteratura (Wallgren A. and Wallgren B., 2014).
In ordine crescente di complessità, la ricostruzione dell'unità statistica a partire dall'unità amministrativa può essere: (1) semplice, (2) assistita da esperto, (3) assistita da integrazione con altri archivi oppure (4) mista.

La ricostruzione dell'unità statistica è detta semplice quando l'unità amministrativa coincide o è facilmente riconducibile a quella statistica per aggregazione delle unità amministrative in base a definiti criteri. Per esempio: (a) gli individui iscritti nelle liste dell'Anagrafe sono unità statistiche della popolazione degli individui (relazione 1:1); (b) gli archivi amministrativi usati per la costruzione dell'Archivio Statistico delle Imprese Attive (ASIA) in Italia rilevano le unità legali di impresa che sono talvolta in relazione n:1 con l'unità statistica impresa; (c) nell'archivio Emens, prodotto mensilmente dall'Inps, l'unità statistica lavoratore è ottenuta aggregando diversi profili contributivi (relazione 1: n).

Altre volte la ricostruzione dell'unità statistica è più complessa e richiede l'ausilio di esperti di settore che sono spesso gli enti titolari della fonte, i soggetti della dichiarazione amministrativa oppure i sostituti del dichiarante nella comunicazione all'ente stesso. Un esempio di tale casistica si ritrova nell'Archivio delle società quotate in Borsa gestito dalla Consob, attualmente utilizzato all'Istat per la creazione del Registro statistico dei Gruppi di impresa: per individuare i legami di controllo tra imprese a partire dalle partecipazioni tra società quotate e non quotate, si ricorre alla conoscenza e all'esperienza sul campo dell'ente titolare della fonte (Consob).

[2] Da cui il termine di oggetti statistici derivati.

Talvolta la ricostruzione dell'unità statistica può essere assistita dall'integrazione con altri archivi, nel senso che alcune unità necessitano dell'integrazione tra archivi amministrativi per poter essere individuate. Per esempio (a) la ricostruzione dei nuclei familiari richiede non solo conoscere la lista degli individui e dei legami parentali ma anche sapere che essi coabitano nello stesso edificio residenziale, informazione derivabile da un archivio sulle residenze; (b) anche la ricostruzione delle unità locali richiede l'integrazione dell'archivio delle camere di commercio, che fornisce gli indirizzi in cui un'impresa opera, con archivi amministrativi/statistici che forniscono informazioni sul numero di dipendenti per luogo di lavoro per poter individuare l'unità statistica definita unità locale.

La casistica che si presenta più frequentemente per la ricostruzione dell'unità è l'approccio misto, che richiede l'ausilio congiunto sia delle conoscenze di esperti che dell'integrazione con altri. Ne è un esempio la ricostruzione dell'unità statistica "gruppo di impresa" che, configurandosi come un'associazione di imprese legate da relazioni di controllo decisionale, richiede da un lato l'ausilio di esperti sia di materia economica, giuridica e fiscale (i commercialisti) che amministrativa (l'ente titolare-Infocamere), per individuare i legami di controllo tra unità amministrative, e dall'altro lato sapere se le unità legali sono definite imprese, informazione derivabile solo per integrazione con il Registro delle Imprese attive (ASIA).

La complessità di ricostruzione dell'unità statistica a partire dall'unità amministrativa può essere funzione dell'unità stessa e crescere se l'unità statistica è composta, anziché semplice, nel senso che è una aggregazione di unità statistiche base (di natura diversa) legate tra loro da uno o più vincoli (relazioni). Ne è un esempio l'unità statistica gruppo di impresa che richiede, oltre alle unità statistiche base (le imprese), anche le relazioni di controllo tra imprese. Naturalmente, le unità statistiche composte vanno derivate a partire dalle unità statistiche base.

Nel ricostruire l'unità statistica si può incorrere nel cosiddetto errore di derivazione, per la cui valutazione gli strumenti a disposizione sono limitati. Una sua misura può essere fornita dal numero di unità che non possono essere attribuite univocamente alla popolazione di interesse. Tuttavia, qualora la ricostruzione dell'unità sia assistita dall'integrazione, gli errori di derivazione potrebbero derivare da errori di *linkage* (si veda Principio B.2.), per la cui valutazione è spesso necessario il ricorso al controllo manuale con operatori esperti.

Il processo di derivazione delle unità dovrebbe essere riproducibile e documentato. Le ipotesi sottostanti tale processo dovrebbero essere esplicitate e documentate.

Valutazione della copertura. Una volta ricostruita la popolazione obiettivo, è opportuno procedere ad una valutazione il più accurata possibile dell'errore di copertura e quindi dell'effettiva rappresentatività della popolazione derivata rispetto a quella obiettivo.
Infatti, la popolazione di una fonte amministrativa singola o integrata può differire dalla popolazione statistica di interesse. Questo determina una errata copertura della popolazione obiettivo. Una mancata corrispondenza tra la popolazione statistica di interesse e quella individuata dalla fonte amministrativa può dipendere da diversi fattori:

(a) *differenza definitoria tra il campo di osservazione target e quello oggetto della fonte amministrativa*. La differenza concettuale tra le due popolazioni determina un'inclusione di unità non appartenenti alla popolazione statistica (sovra-copertura) e una non inclusione di unità appartenenti alla popolazione (sotto-copertura);
(b) *errata individuazione delle unità statistiche*. La complessità di ricostruzione dell'unità statistica a partire dall'unità amministrativa può determinare errori nell'individuazione dell'unità e quindi errori di copertura. Se ad esempio l'unità amministrativa è in relazione n:1 con l'unità statistica, i mancati link nell'individuazione dell'unità generano un errore di sovra-copertura, mentre gli errati link tra unità amministrative causano errori di sotto–copertura.

(c) *ritardi e/o mancate registrazioni amministrative*. Mancate registrazioni determinano in genere errori di sotto-copertura. Tuttavia, quando gli oggetti amministrativi rappresentano eventi (ad esempio per alcuni eventi demografici), le mancate registrazioni possono determinare anche errori di sovra-copertura causati da mancate cancellazioni delle unità statistiche dalla popolazione di riferimento.

Dove possibile, si dovrebbe fare ricorso a più fonti amministrative, perché consente di ridurre sia l'errore di sotto-copertura, integrando fonti che coprono differenti porzioni della popolazione, sia l'errore di sovra-copertura, avendo la possibilità di disporre di più informazioni per stabilire quali unità appartengono correttamente al campo di osservazione desiderato. Per esempio, nell'ambito delle statistiche sulle imprese l'integrazione di diversi archivi amministrativi permette di stabilire se e quali unità appartengono al campo di osservazione delle imprese attive operanti nei settori economici (registro ASIA).

È opportuno tentare di stimare l'errore di copertura nei dati. La possibilità di valutare l'errore di copertura è legata alla disponibilità di un *benchmark* da utilizzare, possibilmente come *gold standard* (registro statistico, altre fonti amministrative). Esistono due principali approcci per misurare indicatori sul tasso di copertura: confronti aggregati rispetto a distribuzioni note e confronti unità per unità (*matching case-by-case*). I primi forniscono valutazioni più grossolane. Per una applicazione vi veda US Bureau of Census (2011). I secondi, maggiormente onerosi perché richiedono attività di *record linkage* tra archivi, si basano su tecniche note in letteratura come Cattura-Ricattura (Wolter, 1986), o evoluzioni di queste tecniche che consentono di rilassare alcune ipotesi (Biemer P.P., 2011, pp. 249-258).

In molti casi il *gold standard* di riferimento non è disponibile ed è quindi necessario individuare metodi di misura differenti. In mancanza di un *gold Standard* il classico metodo per la valutazione della copertura di una lista è l'indagine di copertura, generalmente condotta per la valutazione della copertura delle indagini censuarie.

Alcuni riferimenti bibliografici

Biemer P.P. (2011). Latent Class Analysis of Survey Error. John Wiley & Sons, Inc., Hoboken, New Jersey.

Cerroni F, Morganti E. (2003). La metodologia e il potenziale informativo dell'archivio sui gruppi di impresa: primi risultati. Contributi Istat 3/2003. http://www3.istat.it/dati/pubbsci/contributi/Contr_anno2003.htm

Cerroni, Di Bella, Galiè (2014). Evaluating administrative data quality as input of the statistical production process. Rivista di Statistica Ufficiale N. 1-2/2014.

Blue-Ets (2013). Guidelines on the use of the prototype of the computerized version of the QRCA, and Report on the overall evaluation results. Deliverable 8.2 of Workpackage 8 of the Blue-Ets project. http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable8.2.pdf

Eurostat (2010). Business Registers Recommendations Manual

ESSNet Consistency (2013). Disponibile a https://ec.europa.eu/eurostat/cros/content/consistency-0_en

Wallgren A. and Wallgren B. (2014). Register-based Statistics: Administrative Data for Statistical Purposes. Second Edition. John Wiley & Sons, Chichester, UK. ISBN: ISBN 978-1-119-94213-9

US Bureau of Census (2011). Source and Accuracy of Estimates for Income, Poverty, and Health Insurance Coverage in the United States: 2010 http://www.census.gov/hhes/www/p60_239sa.pdf

Viviano C., Garofalo G. (2000). The problem of links between legal units: statistical techniques for enterprise identification and the analysis of continuity. Istat. Rivista di Statistica Ufficiale 1/2000.

Wolter M.K. (1986). Some Coverage Error Models for Census Data. Journal of the American Statistical Association. Vol. 81, No. 394, pp. 338-346..

Zhang L.C. (2012). Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica (2012), Vol. 66, nr. 1, pp 41-66.

Principio B.4. Derivazione delle variabili e armonizzazione delle classificazioni

Il procedimento di derivazione delle variabili e armonizzazione delle classificazioni deve seguire pratiche consolidate. Tutte le ipotesi alla base del processo di derivazione delle variabili devono essere esplicitate e deve esserne valutata la correttezza. La validità delle variabili derivate deve essere valutata. L'intero processo di derivazione delle variabili e armonizzazione delle classificazioni deve essere documentato.

Linee Guida

Dopo l'unità statistica, l'individuazione delle variabili della fonte amministrativa rilevanti a fini statistici rappresenta il secondo passo per l'utilizzo statistico dell'archivio. Analogamente a quanto riscontrato per le unità, le tecniche di derivazione delle variabili sono di difficile standardizzazione e variano fortemente tra i diversi ambiti di applicazione. Quando possibile, si dovrebbe fare riferimento alle pratiche suggerite in letteratura.

Anche l'adeguatezza delle classificazioni adottate nell'archivio amministrativo deve essere valutata alla luce delle classificazioni necessarie in fase di elaborazione e per la diffusione delle variabili statistiche. In questo ambito assumono particolare rilevanza: la disponibilità delle definizioni relative alle modalità delle classificazioni e la "riconducibilità" tra classificazioni, ossia la capacità di ricostruire ogni modalità della classificazione statistica con una modalità o con l'unione di modalità disgiunte della classificazione amministrativa. Nel caso in cui i dati amministrativi per alcune variabili testuali utilizzino una classificazione che non è riconducibile a quella obiettivo, è opportuno assicurarsi l'accesso ai dati testuali (non codificati), che può consentire la piena riconducibilità alla classificazione di interesse.

Gli errori di classificazione, ossia gli errori che si commettono nell'allineare le classificazioni delle variabili incluse nell'archivio amministrativo con quelle obiettivo statistico, soprattutto se presenti in variabili determinanti per alcuni registri statistici (identificativi territoriali nei registri di popolazione, attività industriale nei registri di impresa) hanno impatto sull'identificazione delle popolazioni di interesse, e possono quindi causare a loro volta errori di copertura.

Individuazione e derivazione delle variabili. Come le unità, anche le variabili statistiche provenienti da un archivio amministrativo o dall'integrazione di più fonti sono create per derivazione da variabili amministrative (Wallgren & Wallgren, 2014) tramite una funzione di trasformazione/derivazione di tipo deterministico o tramite un modello casuale. Entrambi possono essere applicati sfruttando l'informazione interna ad una singola fonte oppure disponibile da più fonti amministrative. In quest'ultimo caso l'attività di derivazione delle variabili implica una attività preliminare di integrazione tra archivi, con gli eventuali errori che si possono generare nell'applicazione di procedure di *record linkage* (si veda Principio B2 per un approfondimento).
In generale, gli errori che si possono generare nel processo di derivazione delle variabili dipendono da una errata specificazione delle regole deterministiche o del modello casuale. È perciò opportuno esplicitare le ipotesi sottostanti il processo di derivazione delle variabili e armonizzazione delle classificazioni e valutarne la validità.

L'intero processo di derivazione delle variabili e armonizzazione delle classificazioni dovrebbe essere riproducibile e documentato.

Valutazione della validità delle variabili. Le differenze concettuali e nei dati (le prime rappresentate dall'errore di specificazione le seconde dall'errore di misura e processo) tra variabili amministrative e statistiche devono essere esplorate ed armonizzate. Gli approcci che possono essere seguiti variano a seconda della disponibilità o meno di variabili di *benchmark* o controllo, e in particolare dipendono da:

(a) la disponibilità di variabili di controllo diretto, ossia variabili statistiche con definizioni coincidenti o raccordabili provenienti da indagine o da Censimento e che abbiano funzione di *gold standard*, situazione che permette un confronto puntuale tra i dati, il calcolo di misure di scala e di forma distributiva degli errori e quello di funzioni di distanza;
(b) la disponibilità di variabili di controllo "funzionali", ossia non coincidenti da un punto di vista concettuale, ma funzionalmente collegate con le variabili oggetto di interesse, che permette l'applicazione di tecniche di *data mining* (per la ricerca degli *outlier*) e tecniche regressive anche multivariate per lo studio delle relazioni funzionali;
(c) la non disponibilità né di variabili di controllo né di variabili funzionali, che implica il ricorso ad approcci basati sullo studio della coerenza tra variabili (intra-fonte o tra fonti), all'analisi fattoriale o ai modelli a classi latenti che ipotizzano strutture di correlazione tra variabili interne all'archivio amministrativo.

Un esempio generalizzabile di validazione di variabili di tipo economico-contabile viene tracciato nel lavoro di Bernardi *et al.* (2013), dove si fa riferimento all'applicazione eseguita sull'archivio degli Studi di Settore. In esso si presenta una metodologia di validazione comprendente le tre situazioni tipo sopradescritte con i relativi metodi di validazione quantitativa associati.

È opportuno considerare che i metodi di validazione delle variabili descritti dovrebbero essere applicati anche in funzione del tipo di utilizzo del dato amministrativo. Infatti, se lo scopo è un utilizzo diretto dell'archivio amministrativo è auspicabile che la validazione della variabile di fonte amministrativa avvenga tramite l'utilizzo di variabili di controllo con funzione di *gold standard*: il confronto puntuale con variabili dalle definizioni coincidenti o raccordabili garantisce una elevata affidabilità a livello di microdato, requisito fondamentale se l'obiettivo è quello di sostituire la variabile statistica con quella proveniente da fonte amministrativa per produrre statistiche dirette.
Nel caso di micro-integrazione (Bakker, 2010; Zhang, 2012), ossia di integrazione al fine di ricostruire le variabili, se un *gold standard* non è disponibile sarà comunque possibile un controllo di coerenza su informazioni provenienti da più fonti amministrative o statistiche, per cui si può effettuare una validazione tramite variabili di controllo funzionali. Va ricordato che l'utilizzo di variabili di controllo funzionali può essere fatto con l'ausilio di modelli statistici - algoritmi per la micro-integrazione (Pannekoek, 2011) - oppure con l'ausilio di esperti. È un esempio di controllo assistito da esperto l'attività di *profiling* effettuata sulle imprese di grandi dimensioni nel Registro Statistico ASIA (Eurostat, 2010).

Ri-classificazione dei valori delle variabili in unità derivate. La derivazione di nuove unità (Principio B3) implica naturalmente un problema di codifica delle variabili relative alle nuove unità, che consiste nel calcolare il valore assunto dalle variabili, già presenti nell'archivio, sulle nuove unità. Ne è un esempio la ricodifica di tutte le variabili riferite all'impresa necessaria in fase di allineamento tra unità legali e imprese. Anche nella scelta del metodo di classificazione o codifica più appropriato si deve fare riferimento a diversi metodi possibili, quali:
(a) la scelta (deterministica o probabilistica) del dato contenuto nell'archivio amministrativo ritenuto più affidabile in termini della specifica variabile di interesse (ad esempio i caratteri identificativi quali ragione sociale, indirizzo, etc.), operazione frequente nel caso di integrazione dei dati da più archivi amministrativi;
(b) l'assegnazione di un nuovo valore alla variabile che è il risultato di un algoritmo specifico (per esempio l'attribuzione dell'attività economica principale all'impresa che avviene mediante il criterio dell'attività prevalente, oppure il calcolo degli addetti nelle unità locali contenute nel Registro Statistico delle unità locali).

Alcuni riferimenti bibliografici

Bakker B.F.M. (2010). Micro-integration: State of the Art. Note by Statistics Netherlands. UNECE Conference of European Statisticians. The Hague, The Netherlands, 10-11 May 2010

Bernardi A., Cerroni F. e De Giorgi V. (2013). Uno schema standardizzato per il trattamento statistico di un archivio amministrativo. Istat Working Papers 4/2013

Eurostat (2010). Business Registers Recommendations Manual.

Pannekoek, J. (2011). Models and algorithms for micro-integration. chapter 6. In Report on WP2: Methodological developments, ESSNET on Data Integration, available at https://ec.europa.eu/eurostat/cros/content/data-integration-finished_en

Wallgren A. and Wallgren B. (2014). Register-based Statistics: Administrative Data for Statistical Purposes. John Wiley & Sons, Chichester, UK. ISBN: ISBN 978-1-119-94213-9

ESSnet AdminData (2013). Final list of quality indicators and associated guidance. Deliverable 2011/6.5 of ESSnet on Admin Data https://ec.europa.eu/eurostat/cros/content/use-administrative-and-accounts-data-business-statistics_en

Zhang L-Chun (2012). Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica (2012) Vol 66, nr.1, pp. 41-63.

Principio B.5. Dimensione temporale e territoriale

I diversi aspetti inerenti la dimensione temporale e territoriale devono essere attentamente valutati alla luce delle caratteristiche dei dati amministrativi utilizzati, rispetto alla regolamentazione dell'atto amministrativo sul territorio, in relazione al ciclo di vita dell'archivio amministrativo.

Linee Guida

Una caratteristica rilevante dell'obiettivo di interesse nella produzione statistica riguarda la natura dei dati che si vuole produrre: la stima di interesse può essere trasversale (*cross-sectional*), longitudinale in un'ottica di microdati (indagini di natura panel) oppure longitudinale in un'ottica di macrodati, ossia per la produzione di serie storiche (indagini ripetute). Quindi definire il prodotto statistico di interesse implica identificare con precisione la dimensione temporale e spaziale dei dati che si utilizzano a tale scopo. Ciò significa attribuire delle caratteristiche "luogo" e "tempo" alla definizione della popolazione di interesse, e di conseguenza delle unità che la compongono. Analogamente ogni variabile è associata ad un riferimento temporale[3], sia esso un istante o un periodo di tempo. Le variabili di interesse statistico possono avere una natura di variabili di *stock* o di variabili di flusso (Wallgren & Wallgren, 2014):

- le variabili di *stock* forniscono la situazione in un punto di tempo specifico, per es.: età di un individuo ad una certa data, numero di addetti dell'impresa alla fine dell'anno;
- le variabili di flusso rappresentano somme in un dato periodo, per es: reddito percepito in un anno, ordini effettuati da una impresa in un mese.

Nell'utilizzare i dati di fonte amministrativa, è necessario l'allineamento del riferimento temporale e territoriale dei dati, sia qualora i dati in oggetto vengano utilizzati per la produzione diretta di statistiche che per la realizzazione di registri di microdati integrati. Dove sia necessario un processo di trasformazione delle informazioni per riportarle al riferimento temporale e territoriale desiderato, è inoltre rilevante verificarne le ipotesi di base.

Per l'aspetto relativo alla dimensione temporale possiamo identificare archivi che contengono:

- dati di stock riferibili ad un determinato istante di tempo, aggiornati in tempo reale o in un momento successivo alla data di riferimento;
- informazioni longitudinali sugli oggetti unità e eventi nati/cessati nel tempo e le loro caratteristiche. Non tutte le caratteristiche possono essere tracciate longitudinalmente.

Gli archivi possono contenere "variabili di riferimento temporale" ossia variabili che identificano l'istante di tempo relativo ad un evento che si verifica per un oggetto e in particolare su una unità. Per esempio sull'unità "individuo" e l'evento "ricovero ospedaliero", può essere associata la variabile di riferimento temporale "data del ricovero". Il formato di tali date varia in funzione delle procedure amministrative e può essere identificato da una data esatta, come per un evento caratteristico di un individuo della popolazione o da un periodo, per esempio un mese per un evento relativo ad un'impresa.

È evidente che le caratteristiche temporali dei dati contenuti nell'archivio amministrativo condizionano le possibilità di produzione statistica, laddove per analisi longitudinali sono necessari dati con natura longitudinale mentre per la produzione di stime puntuali sono sufficienti dati amministrativi di *stock* (le stesse possono anche essere derivate da archivi con dati longitudinali).

[3] Una variabile statistica è definita da: *i)* l'unità che possiede la caratteristica (esempio reddito per le persone e reddito per le famiglie sono due diverse variabili), *ii)* il metodo di misurazione, *iii)* la scala di misurazione e *iv)* il determinato istante o periodo di tempo cui si riferisce la misurazione (Wallgren & Wallgren, 2014, Cap 8, pag. 148).

Per quello che riguarda la dimensione geografica, è necessaria una valutazione attenta del riferimento territoriale dei dati amministrativi utilizzati, perché alcuni archivi amministrativi potrebbero avere una copertura territoriale diversa da quella obiettivo della statistica da produrre. Ciò può rendere necessaria un'integrazione tra dati di fonte amministrativa e/o con dati da indagine.

La costruzione di registri statistici di popolazione o di imprese, ossia archivi idealmente completi rispetto alle suddette popolazioni, richiede informazioni necessarie per tracciare e seguire le unità, le loro caratteristiche e relazioni nel tempo, ma anche informazioni sulle date delle occorrenze che si verificano in un dato periodo di tempo. Poiché seguire gli oggetti e le loro identità nel tempo è una attività onerosa, è opportuno distinguere gli eventi non rilevanti da quelli rilevanti ai fini degli obiettivi statistici, per i quali è indispensabile conoscere l'istante dell'occorrenza o la durata dell'evento.

Così come è importante conoscere le date di occorrenza degli eventi registrati nell'archivio amministrativo, è importante conoscere le loro date di registrazione. La conoscenza dei riferimenti temporali consente la costruzione di registri statistici che riflettono lo stato della popolazione in un dato istante o periodo di tempo. È importante condurre analisi di scenario per valutare come il tempo di registrazione degli eventi e di acquisizione dei dati amministrativi impatti su altre dimensioni della qualità, quali la copertura della popolazione obiettivo o l'accuratezza delle stime di frequenza.

Nel caso di dati integrati, le fonti coinvolte nel processo di integrazione possono contenere riferimenti temporali differenti, rappresentando un ulteriore elemento di complessità legato all'integrazione. La tempestività con la quale variazioni longitudinali nei dati vengono registrate negli archivi amministrativi influisce sulla qualità delle attività di integrazione e conseguentemente, può portare ad altri errori nei dati statistici prodotti. Le scelte e le ipotesi assunte quando si integrano dati con riferimenti temporali diversi, devono essere opportunamente motivate e documentate. Dovrà essere ben chiaro quale sia il riferimento temporale dei dati di output derivanti dal processo di integrazione.

Negli archivi che contengono informazioni longitudinali, la maggior parte degli oggetti non presenta difficoltà in fase di stima, in quanto si tratta di unità che sono presenti per l'intero periodo di riferimento di interesse, per esempio per l'intero anno di calendario. Alcune volte, gli oggetti entrano ed escono e possono richiedere l'utilizzo di appropriati pesi in fase di stima. Il tempo può essere usato come una variabile che genera i pesi, e usata per correggere le stime.

Le modifiche legislative e quelle procedurali, come per esempio nel caso di una variazione della classificazione adottata, possono impattare sulla comparabilità delle serie storiche prodotte e in particolare possono portare a cambiamenti nei livelli delle serie che non riflettono il reale contesto del fenomeno. È opportuno vigilare su tali modifiche, siano esse documentate nei metadati di supporto ai dati amministrativi (anche tardivamente rispetto al rilascio dei dati statistici) o conosciute in anticipo, per comprendere la reale natura di *shift* nelle serie temporali ed agire di conseguenza.

Le procedure e le tecniche utilizzate per combinare dati di fonte amministrativa caratterizzati da riferimenti temporali diversi e/o diversi da quelli delle stime o dei microdati da produrre, e i sistemi di ponderazione basati su variabili relative alla dimensione temporale, dovrebbero essere basate su metodologie consolidate ed essere riproducibili e opportunamente documentate.

Alcuni riferimenti bibliografici

Wallgren A. and Wallgren B. (2014). Register-based Statistics: Administrative Data for Statistical Purposes. John Wiley & Sons, Chichester

Principio B.6. Controllo e correzione
La strategia adottata per la fase del controllo e correzione, quando i dati provengano da fonti amministrative, deve tenere conto delle peculiarità delle tipologie di errore proprie di tali dati. L'impatto della procedura deve essere valutato attraverso idonei indicatori.

Linee Guida

Nella progettazione di un piano di controllo e correzione (C&C di seguito) bisogna tenere a mente quali sono gli obiettivi di questa fase, che non deve essere vista semplicemente come una procedura di 'pulizia' dei dati ma come un vero e proprio passo di validazione.

I principali obiettivi di una procedura di C&C dei dati sono così riassumibili:

- identificare le possibili fonti di errore al fine di migliorare il processo di produzione delle statistiche;
- fornire informazioni sulla qualità dei dati raccolti e rilasciati;
- rilevare e correggere gli errori influenti;
- fornire dati completi e coerenti.

Tali obiettivi sono validi anche nel contesto del C&C dei dati di origine amministrativa.

Le linee guida già sviluppate per un processo di produzione statistica di tipo "rilevazione diretta" rimangono generalmente valide anche in questo contesto[4]. Nel caso di dati da fonte amministrativa è necessario però tener conto delle ulteriori specificità che caratterizzano tali dati e che hanno un impatto sull'organizzazione di un piano di C&C.

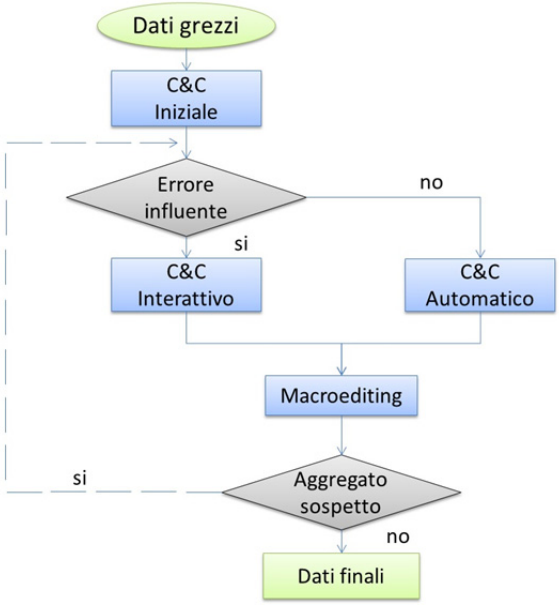
Nella progettazione di una procedura di C&C per dati amministrativi il primo elemento da tenere in considerazione è se le statistiche oggetto di stima sono ottenibili utilizzando un singola fonte oppure derivano dall'integrazione di più fonti.

Strategia di controllo e correzione nel caso di singolo archivio di dati di fonte amministrativa. Nel caso di utilizzo di un singolo archivio, la situazione è simile a quella in cui si debba progettare una procedura di C&C di singola indagine statistica, che tipicamente può essere rappresentata dal seguente grafico di flusso (Figura 1) tratto dal manuale sviluppato nell'ambito del progetto Edimbus (Luzi *et al*., 2008).

Nella prima fase (C&C Iniziale) vengono corretti tutti gli errori evidenti e di tipo deterministico. Successivamente l'insieme di dati viene suddiviso in due sottoinsiemi: uno generalmente con poche unità, caratterizzate da errori potenzialmente influenti e per le quali è richiesta un'operazione di revisione molto accurata (C&C Interattivo), ed un altro composto da molte unità potenzialmente affette da errori meno gravi e che possono essere trattate con metodi automatici, che sono generalmente più efficienti in termini di costi (C&C Automatico). Infine, vengono prodotte le stime e l'insieme iniziale di dati viene ricomposto e controllato in relazione a valori sospetti delle stime e rispetto a valori attesi o ad aggregati disponibili da altre indagini (cfr. Principio B.8.).

[4] Linee Guida per la qualità dei processi statistici (http://www.istat.it/it/files/2010/09/Linee-Guida-Qualit%C3%A0-_v.1.1_IT.pdf)

Figura 1. Grafico di flusso delle operazione di un processo di controllo e correzione dati



Strategia di controllo e correzione nel caso di più archivi di dati di fonte amministrativa. Nel caso di integrazione di più archivi, è necessario valutare in quale punto del processo applicare le procedure di C&C. In generale, ci si troverà di fronte a due alternative per la strategia complessiva del controllo e correzione, che configurano i due scenari seguenti:

1. C&C di ogni singolo archivio → Integrazione tra archivi → C&C di archivi integrati
2. Integrazione degli archivi → C&C degli archivi integrati

Un vantaggio del primo scenario rispetto al secondo, è che esso consente di rimuovere dalle singole fonti alcune tipologie di errori prima della fase di integrazione (come errori sistematici dovuti ad esempio a errori di unità di misura, errori di quadratura, ecc.), riducendo in tal modo la possibilità di errori di coerenza sui dati integrati. D'altra parte, nel primo scenario, non tutta l'informazione disponibile viene utilizzata congiuntamente: ad esempio, per l'imputazione di una variabile in un archivio potrebbe essere utile sfruttare variabili osservate in uno o più degli altri archivi e strettamente legate alla variabile da imputare. Inoltre, nel C&C post-integrazione, può essere necessario rimuovere anche eventuali inconsistenze inter-fonte, generate dal processo stesso di integrazione. Il primo scenario, inoltre, richiede costi elevati in termini di tempi e risorse. Una riduzione di tempi e costi di questa soluzione può essere realizzata minimizzando le risorse spese per l'editing dei singoli archivi, garantendo al contempo un livello accettabile di qualità dei microdati: una soluzione in questo senso potrebbe consistere ad esempio nell'effettuare sui singoli archivi solo C&C di errori sistematici ed errori influenti.

Il secondo scenario, rispetto al primo, è caratterizzato da un minore impegno di risorse, essendo necessario il disegno di una sola procedura di C&C. Tuttavia la complessità della procedura stessa può essere molto superiore rispetto al caso precedente. Inoltre, dal momento che i dati integrati non contengono generalmente tutte le variabili disponibili nei singoli archivi amministrativi, può accadere che alcune relazioni fra variabili esistenti negli archivi originari vengano perse.

In generale, la scelta dello scenario più appropriato è basata sul rapporto tra il livello di qualità attesa nei dati finali e le risorse effettivamente disponibili per ottenere tale livello.

Un elemento che in ogni caso riveste un'importanza cruciale ai fini dell'aumento dell'efficacia del processo di C&C è la disponibilità di esperti di modulistica amministrativa, che abbiano familiarità con i processi

amministrativi che hanno "generato" i dati e i loro specifici contenuti informativi, e che mantengano rapporti stretti e costanti con gli Enti fornitori dei dati.

É evidente che l'uso di informazioni longitudinali può aumentare l'efficienza della strategia complessiva di C&C.

Tipologie di errore e metodi di trattamento. Nelle applicazioni su dati di origine amministrativa, la tipologia di errore più rilevante e insidiosa è quella dell'errore di specificazione, ossia la non corrispondenza tra le definizioni obiettivo statistico (variabili) e quelle utilizzate per la produzione del dato amministrativo. Questo tipo di errore si riflette nelle variabili osservate in quello che è noto come errore di misurazione, e in particolare un errore di misurazione con una forte natura sistematica. La valutazione dell'esistenza o meno di questo tipo di errore, che richiede uno studio approfondito dei concetti e il coinvolgimento di esperti di settore, è il primo passo fondamentale di qualsiasi strategia di C&C sui dati di fonte amministrativa. Per la sua identificazione possono essere utili tecniche per l'individuazione di dati anomali applicate alla differenza fra i dati amministrativi e dati di indagine (ove siano disponibili). Ad esempio, può accadere che, sebbene le definizioni delle variabili statistiche obiettivo nelle fonti amministrative e nell'indagine siano state armonizzate, per qualche sottoinsieme di dati tale armonizzazione non sia sufficiente: questo può dar luogo ad un sottoinsieme di dati caratterizzato da una grossa differenza (relativamente alla distribuzione delle differenze fra valori) che può quindi suggerire la presenza di un problema su tale insieme di dati. Per la correzione di questa tipologia di errore risultano appropriati approcci di tipo deterministico, tipicamente adottati nel caso di errori di natura sistematica.

Per quanto riguarda la localizzazione e il trattamento degli errori influenti, l'applicazione di metodi di editing selettivo (*selective editing*), così come di editing interattivo, nei dati di fonte amministrativa è limitato dalle dimensioni dei dati e dalla difficoltà di ricontatto della "fonte" primaria (il rispondente che ha fornito il dato all'ente titolare). D'altra parte, non essendoci per questo tipo di dati un processo di selezione del campione, la selezione delle unità da ricontattare risulta semplificata dal fatto che non si deve tener conto dei pesi finali, come deve essere fatto nelle indagini classiche, e quindi può essere basata semplicemente sull'identificazione delle unità più influenti, ossia aventi maggior impatto sui risultati finali. É in ogni caso importante che nel processo di revisione interattiva dei dati influenti siano coinvolti anche esperti del dato amministrativo utilizzato, al fine di massimizzare la possibilità di corretta identificazione della causa dell'errore e il suo adeguato trattamento.

Relativamente all'uso di procedure automatiche di C&C, ad esempio quelle basate sul principio del minimo cambiamento dei dati, esso non pone questioni particolari nel caso di dati provenienti da una singola fonte amministrativa. Nel caso di dati provenienti da un processo di integrazione, laddove una stessa variabile sia disponibile da più fonti e si possono generare errori di coerenza inter-fonti, per la definizione della procedura automatica è necessario distinguere tra due situazioni:

- caso in cui una fonte sia considerabile un *gold standard*, quindi non affetta da errore.
- caso in cui tutte le fonti siano considerate ugualmente attendibili.

Nel primo caso attraverso procedure automatiche si verifica la correttezza del dato ritenuto *gold standard* relativamente alla sua consistenza all'interno dell'archivio considerato più affidabile. Nel caso in cui si verifichi una inconsistenza, per esempio una violazione di una regola di controllo (*edit*), il dato dell'archivio ritenuto meno affidabile può essere usato per ricostruire un dato consistente. Un esempio si ha quando il dato *gold standard* è affetto da un errore di misura, in questo caso il dato dell'archivio considerato ausiliario può aiutare a svelare questa tipologia di errore.

Nel caso invece in cui non vi sia una fonte ritenuta più affidabile, tutte le osservazioni possono contribuire alla ricostruzione di un dato finale consistente con i vincoli attesi. Le procedure utilizzate per la ricostruzioni

di variabili nel contesto dell'integrazione dei dati vanno sotto il nome di *microintegration*. Alcuni metodi automatici per trattare le due situazioni precedentemente esposte si possono trovare in Pannekoek (2014).

Così come nei dati da indagine, le tecniche di *macroediting*, ossia i metodi per l'individuazione degli errori partendo dai confronti di aggregati, possono essere utili per rivelare errori nei dati di un singolo archivio o nei dati integrati.

Per quanto riguarda il trattamento dei dati mancanti, che generalmente consiste nell'imputazione ovvero nella "previsione" del dato non disponibile o non utilizzabile nell'archivio amministrativo, è necessario sottolineare alcuni aspetti importanti.

Un elemento essenziale da tenere in considerazione è che la maggior parte delle tecniche di imputazione si basano sull'ipotesi che la probabilità di non risposta per una data variabile sia correlata con i valori osservati e non dipenda da fattori che causano la mancata risposta stessa (denominata MAR, *missing at random*). In sostanza questo vuol dire che le osservazioni disponibili sono sufficienti per stimare un modello di previsione dei dati mancanti, in termini meno rigorosi la popolazione osservata è rappresentativa di quella non osservata.

Con questa premessa, il primo elemento da tenere in considerazione è quindi la causa del dato mancante.

Il dato può essere mancante all'interno di un archivio perché la variabile in questione non è di principale interesse dell'ente amministrativo che raccoglie le informazioni, in questo caso può accadere che il soggetto dell'atto amministrativo non sia sufficientemente sollecitato a fornire informazioni su tale variabile. Questa situazione è simile a quella che si ha nelle indagini per le quali frequentemente viene ipotizzato un meccanismo MAR.

Un altro importante caso che si incontra nell'uso di dati di fonte amministrativa è quello che si ha quando si integrano diversi archivi che non rilevano tutti le stesse variabili. In questo caso si ha una mancata risposta che non può essere considerata casuale, perché questa corrisponde ai diversi segmenti di popolazione individuati dalle diverse fonti. Ipotizzare per tali situazioni un meccanismo MAR corrisponde quindi ad accettare l'idea che il segmento di popolazione in cui vengono osservate le variabili abbia un comportamento analogo in termini di struttura delle variabili con mancata risposta al segmento di popolazione non osservato.

Ad esempio, si supponga di voler stimare le voci del conto economico di tutte le imprese Italiane utilizzando la fonte Bilanci Civilistici (BC) e la fonte Modello Unico (Unico). Per la sottopopolazione delle Società di Capitale, la fonte BC fornisce tutte le informazioni necessarie alla stima delle voci di dettaglio del conto economico. Per la sottopopolazione dei Professionisti, non essendo possibile utilizzare la fonte BC, la fonte Unico fornisce informazioni utili alla stima di un sottoinsieme di voci del conto economico. Per la stima delle altre variabili obiettivo è necessario utilizzare le informazioni della fonte BC (uniche informazioni disponibili) e questo implica l'ipotesi che Professionisti e Società di Capitale siano caratterizzate da comportamenti omogenei in termini dei fenomeni oggetto di stima.

Come già accennato, quando si utilizzano dati amministrativi, poiché spesso le singole fonti non sono esaustive della popolazione in oggetto, è frequente l'utilizzo congiunto di più fonti che vanno quindi integrate.

L'integrazione di archivi comporta la possibilità di introdurre ulteriori tipi di errori. Nelle procedure di integrazione sono presenti *falsi link* e *falsi non link*. Gli errori introdotti dai *falsi non link* non possono essere trattati in fase di controllo e correzione perché non danno luogo ad osservazioni da controllare. I *falsi link* invece generano delle unità il cui contenuto informativo (per esempio le variabili provenienti dalle diverse fonti) è incoerente. L'incoerenza tra le variabili di un unità può essere riscontrata anche nel caso in cui l'abbinamento sia stato efficace, infatti sebbene si ipotizzi che le variabili siano state armonizzate, è sempre

possibile un disallineamento, anche se di lieve entità. Possiamo quindi riassumere affermando che errori specifici in questo contesto e non presenti nelle indagini classiche causano incoerenze intra-fonte, che sono naturalmente presenti nelle incoerenze inter-fonte o artificialmente introdotte da un falso abbinamento. Esistono diverse tecniche per trattare questa tipologia di errori che sono volte alla riconciliazione dei dati, anche in questo caso si fa riferimento alla cosiddetta *microintegration* per la quale si possono trovare importanti riferimenti in https://ec.europa.eu/eurostat/cros/content/data-integration-finished_en ed in Pannekoek (2014).

Valutazione dell'impatto della procedura. Indipendentemente dalla strategia adottata, come parte integrante della procedura di C&C, è necessario calcolare e monitorare insiemi di indicatori di qualità sui dati di input e di output, quali frequenza di attivazione di regole di controllo, tassi di imputazione, indicatori di impatto sulle distribuzioni (Luzi *et al*., 2008). Inoltre, poiché il processo di raccolta dei dati amministrativi è fuori dal controllo degli Istituti di Statistica, è importante che siano definiti strumenti per la segnalazione di eventuali variazioni nel processo di formazione del dato presso l'ente fornitore, per il monitoraggio e la riduzione degli effetti di tali variazioni sulle statistiche oggetto di diffusione.

Alcuni riferimenti bibliografici

de Waal, T., Pannekoek, J., Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*, Wiley.

Luzi O., Di Zio M., Guarnera U., Manzari A., De Waal T., Pannekoek J., Hoogland J., Tempelman C., Hulliger B., Kilchmann D. (2008) Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys. EDIMBUS project https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Pannekoek J., (2014). Method: Reconciling Conflicting Micro-Data, in the Handbook for Modern Business Methodology. http://www.cros-portal.eu/content/reconciling-conflicting-micro-data

Principio B.7. Processo di stima

Al fine della produzione delle stime, i dati acquisiti dalle fonti amministrative e opportunamente trattati devono essere elaborati in accordo a metodologie che tengano conto delle specificità del contesto informativo (presenza o assenza di dati di indagine) e produttivo. Quando è richiesto l'uso di modelli statistici di stima, le assunzioni sottostanti tali modelli devono essere opportunamente esplicitate e ne deve essere testata l'effettiva validità. Le stime prodotte devono essere accompagnate da stime dell'errore, in modo da permettere un corretto uso e interpretazione dei risultati.

Linee guida

La procedura per derivare le stime di interesse (stime di livelli, rapporti, tabelle di contingenza, ecc.) utilizzando fonti amministrative deve essere chiara, ben definita e possibilmente deve affidarsi a tecniche statistiche consolidate.

Il processo di stima mediante il dato amministrativo (base dati in cui sono state integrate le diverse fonti disponibili) dipende principalmente dalle seguenti assunzioni:

(1) la popolazione statistica obiettivo è coperta dalla o dalle fonti amministrative utilizzate;
(2) la variabile amministrativa osservata coincide nella definizione con la variabile di interesse ed eventuali differenze sono di tipo casuale e non sistematico;

e dai seguenti scenari informativi:

(a) esiste solo il dato amministrativo ed è nota la dimensione della popolazione di riferimento ed eventualmente la dimensione di sottopopolazioni di interesse;
(b) esiste una stima non distorta del parametro che si intende stimare con il dato amministrativo;
(c) esistono stime affidabili di parametri correlati al parametro di interesse.

Gli approcci utilizzabili dipendono dalle assunzioni e dagli scenari sopra elencati, ma variano anche in funzione dell'uso diretto o ausiliario dei dati amministrativi.

Uso diretto dei dati amministrativi per la stima. Se valgono le assunzione (1) e (2) e cioè se il dato amministrativo copre tutta la popolazione, la stima di un totale si ottiene come somma semplice dei valori.

In generale però l'assunzione (1) è raramente soddisfatta e il dato amministrativo offre una parziale copertura della popolazione di interesse, rappresentando pertanto un campione non probabilistico della popolazione stessa. In questo caso, è consigliato un trattamento dei valori mancanti che deve essere appropriato e coordinato tra le variabili da considerare (Wallgren e Wallgren, 2014). In generale quando la condizione (2) è verificata con parziale copertura della popolazione di interesse il processo si può avvalere dei due approcci alla stima che seguono:

- un modello statistico di predizione che deve essere chiaramente esplicitato e deve basarsi su assunzioni ragionevoli e possibilmente verificabili. Il principale obiettivo del modello statistico è la predizione dei valori della variabile di interesse nella parte non coperta dalle fonti amministrative integrate. Il modello statistico utilizza come covariate le variabili note sulla parte di popolazione coperta e non coperta dal dato amministrativo. Questo approccio, in generale, è appropriato quando si integrino dati da registri diversi per coprire la popolazione e si presentino valori mancanti per un insieme limitato di variabili di interesse;
- un metodo di calibrazione che calcola un peso per ciascuna unità che presenta il dato amministrativo secondo la logica dell'approccio alla calibrazione che riporta a totali noti per tutta la popolazione di riferimento (Wallgren e Wallgren, 2014). Il processo implicitamente considera un modello statistico in cui le covariate sono rappresentate dalle stesse variabili che definiscono i totali noti. Questo approccio è più appropriato quando non si hanno a disposizione un numero elevato di variabili di fonte amministrativa per l'unità.

In entrambi i processi di stima è fondamentale l'ipotesi che il modello stimato nella parte coperta dalla fonte sia valido per la parte non coperta. Questa ipotesi è verificata solo approssimativamente e ciò può generare distorsione nelle stime, il cui impatto può essere tuttavia ridotto in termini di MSE relativo.

Definendo un parallelismo con la teoria della stima da popolazione finita queste due procedure rappresentano un processo di imputazione e un processo di calibrazione. Nel primo caso si può utilizzare un'imputazione casuale (perturbando il valore predetto) o deterministica. Solo nel caso di variabili con un tasso di valori mancanti molto basso e distribuito uniformemente nei domini, cioè nelle sottopopolazioni di interesse, si può ipotizzare di produrre stime per somma, indicando il tasso di valori mancanti.

Nel caso dello scenario informativo di tipo (a) indicato precedentemente, i due approcci alla stima sfruttano le sole informazione da archivio amministrativo. In particolare, per il processo di stima che usa il secondo metodo (quello di calibrazione) è opportuno che la somma dei pesi riporti alle dimensioni note della popolazione/sottopopolazione.

In presenza dello scenario informativo (b) sono possibili diversi usi del dato amministrativo nel processo di stima:

- si può stimare il modello statistico utilizzando le unità del campione non coperte dalle fonti amministrative (Latilla e Holmberg, 2010). Si confrontano gli MSE delle stime campionarie e le stime con il dato amministrativo basate sul modello statistico/calibrazione e si sceglie lo stimatore con il minore MSE per i domini di maggiore dettaglio di interesse;
- si predicono i valori della variabile di interesse nella parte non coperta dal dato amministrativo mediante un modello statistico e si definisce uno stimatore composto che è dato dalla combinazione convessa[5] tra la stima campionaria e quella da dato amministrativo. I pesi delle due stime sono inversamente proporzionali ai rispettivi MSE (Moore *et al*., 2008). Questo approccio può condurre a processi di stima misti in cui parte delle stime sono totalmente campionarie e parte totalmente da dato amministrativo;
- si combinano attraverso tecniche di integrazione le fonti amministrative e i dati campionari a livello di unità. Si definisce una gerarchia sull'affidabilità del valore osservato tra dato amministrativo e campionario. Si ricostruisce l'universo integrando le due fonti e scegliendo in caso di sovrapposizione tra dato amministrativo e campionario quello gerarchicamente più affidabile. Si stima il totale della parte non coperta con la stima campionaria (Kuijvenhoven and Scholtus, 2010, 2011).

Nei tre processi di stima si assume che le dimensioni delle popolazioni /sottopopolazioni di interesse siano note.

Nel caso in cui lo scenario informativo sia di tipo (c) il processo di stima si può affidare ad un metodo di calibrazione che sfrutti i valori delle stime correlate al parametro di interesse.

Uso ausiliario dei dati amministrativi nel processo di stima. Quando la condizione (1) delle assunzioni iniziali non è valida, e cioè il dato amministrativo non rileva esattamente la variabile che definisce il parametro di interesse (errore sistematico), ma la condizione (b) è soddisfatta, allora il dato amministrativo può essere utilizzato come fonte ausiliaria. In particolare, questa strategia migliora l'efficienza delle stime campionarie se esiste una correlazione tra variabile di archivio e variabile di interesse. La stima si deve avvalere in questo caso di modelli statistici o del metodo di calibrazione che utilizzano i pesi diretti di riporto all'universo:

- si raffina il processo di stima campionaria sfruttando il dato amministrativo. In particolare, si individuano i pattern informativi omogenei in termini di variabili presenti nelle fonti amministrative. Si procede alla stima campionaria con lo stimatore di calibrazione per ciascuna sottopopolazione che

[5] Combinazione convessa: combinazione lineare di stimatori con coefficienti non negativi la cui somma è pari a 1.

presenta un dato pattern di variabili amministrative. Il peso finale di riporto all'universo è unico per ciascuna unità del campione;

- si possono utilizzare stimatori campionari specifici per ciascuna variabile di interesse. In questo caso si definisce uno stimatore proiezione (projection estimator) (Luzi *et al*., 2014) che sfrutta un modello statistico per ciascuna variabile di interesse. Lo specifico modello presenta come covariate quelle maggiormente correlate con la variabile di interesse presenti nel pattern di variabili amministrative. I valori predetti imputano i dati mancanti. Il modello è stimato con i pesi di riporto all'universo del campione probabilistico. Si produce una imputazione di massa su tutta la popolazione di riferimento per ciascuna variabile di interesse (Kim e Rao, 2011).

Indipendentemente dall'uso specifico dei dati amministrativi, dove possibile, si dovrebbe tentare di stimare la correttezza e precisione dei risultati, utilizzando anche metodologie avanzate, tra cui metodi Bayesiani e Modelli ad Equazioni Strutturali (Bryant J.R. e Graham P.J., 2013; Scholtus S. and Bakker B.F.M., 2013), purché applicati in un contesto di validità delle assunzioni che ne permettano l'applicazione. In ogni caso, il processo di stima e le assunzioni su cui si basa devono essere opportunamente documentati, e tentata una valutazione sui possibili errori che si generano da questa fase.

Alcuni riferimenti bibliografici

Bryant J.R. and Graham P.J. (2013). Bayesian Demographic Accounts: Subnational Population Estimation Using Multiple Data Sources. Bayesian Analysis (2013), 8, n. 3, pp.591-622.

ESSnet Admin Data (2013 Guidance on the Accuracy of Mixed-Source Statistics,). Deliverable 6.3/2011 USE OF ADMINISTRATIVE AND ACCOUNTS DATA IN BUSINESS STATISTICS March, 2013.

Kim, J. K. K., Rao, J. N. K. (2011). Combining data from two independent surveys: a model-assisted approach. Biometrika. No.8, pp. 1–16.

Kuijvenhoven, L. and Scholtus S. (2010). Estimating accuracy for statistics based on register andsurvey data. Discussion paper 10007. Statistics Netherlands, The Hague/Heerlen.

Kuijvenhoven, L. and Scholtus S. (2011). Bootstrapping combined estimator based on register and sample survey data. Discussion paper 201123. Statistics Netherlands, The Hague/Heerlen.

Laitila, T. and A. Holmberg, 2010. Comparison of sample and register survey estimators via MSEdecomposition. Paper for the European Conference on Quality in Official Statistics, 4–6 May,Helsinki.

Luzi, O., Guarnera, U., Righi, P. (2014). The new multiple-source system for Italian Structural Business Statistics based on administrative and survey data. European conference on quality in Official Statistics Q2014. Vienna, Austria, 2-5 June 2014.

Moore, K., G. Brown and T. Buccellato, 2008. Combining sources: a reprise. Paper for the CENEXISAD workshop Combination of surveys and admin data, 29–30 May, Vienna.

Office for National Statistics, Annual Business Survey. http://www.ons.gov.uk/ons/guidemethod/method-quality/specific/business-and-energy/annual-business-survey/index.html (accessed on 8/12/12).

Scholtus S. and Bakker B.F.M. (2013). Estimating the validity of administrative and survey variables through structural equation modeling. A simulation study on robustness. Discussion Paper (2013)

Wallgren A. and Wallgren B. (2014). Register-based Statistics: Administrative Data for Statistical Purposes. Second Edition. John Wiley & Sons, Chichester.

Principio B.8. Validazione dei risultati

I risultati delle analisi, prima della pubblicazione, dovrebbero essere valutati insieme ad esperti del settore per verificare se vi siano o meno delle anomalie. Quando possibile, i risultati devono essere confrontati con quelli ottenuti in precedenti replicazioni del processo stesso, dove sono stati utilizzati i dati della stessa fonte amministrativa. In alternativa, il confronto può essere effettuato con risultati simili ottenuti da altri processi dello stesso ente o di altri enti. Inoltre, dovrebbero essere calcolati ed analizzati, in modo rigoroso, gli indicatori di qualità di processo.

Linee guida

Prima di essere diffusi, i risultati del processo produttivo statistico, anche quando siano stati utilizzati i dati di fonte amministrativa, devono essere valutati mediante confronti con i risultati di precedenti edizioni, dove sono stati utilizzati i dati della stessa fonte amministrativa e mediante confronti con fonti statistiche interne, o esterne all'Istituto. Eventuali differenze riscontrate devono essere giustificate e documentate.

Se possibile, andrebbe controllata la coerenza dei risultati rispetto a rapporti che possono essere considerati pressoché costanti o soggetti a modifiche minime nel breve periodo, come ad esempio alcuni rapporti demografici. Anche in questo caso, eventuali differenze devono essere giustificate e documentate.

Inoltre, prima del rilascio dei dati, in caso di valori sospetti, i risultati devono essere controllati da esperti dell'Istituto o da esperti esterni, in primo luogo rappresentanti degli enti fornitori del dato amministrativo, ma anche eventualmente rappresentanti del mondo accademico o delle associazioni di categoria. Se il controllo viene effettuato da esperti esterni all'Istituto deve essere garantito il rispetto della confidenzialità dei dati. In ogni caso è preferibile coinvolgere, nella validazione, esperti, interni o esterni, che non siano direttamente impegnati nella produzione del dato.

Nella fase di validazione, gli indicatori di qualità disponibili sui dati di input e quelli calcolati durante il processo di trattamento del dato (come per es. indicatori sull'errore di abbinamento) andrebbero analizzati sistematicamente e confrontati con eventuali livelli attesi di tali indicatori o comunque con l'obiettivo di cogliere i punti di debolezza del processo e identificare possibili azioni correttive.

Il calcolo e l'analisi di misure di qualità e di indicatori di processo, sono finalizzati, in primo luogo, a garantire la qualità delle stime diffuse e, in secondo luogo, a valutare l'opportunità di adottare azioni di miglioramento per le successive edizioni del processo.

Laddove siano possibili margini di miglioramento agendo sulla fonte amministrativa, il risultato della valutazione dovrebbe concretizzarsi in informazione di ritorno per l'ente titolare del dato amministrativo, attraverso la struttura centralizzata che cura i rapporti con l'ente.

Principio B.9. Archiviazione, tutela della riservatezza, diffusione dei dati e documentazione
I microdati validati, opportunamente corredati da misure di qualità e metadati, devono essere archiviati secondo gli standard di Istituto prima del loro rilascio interno per ulteriori usi statistici e prima della loro diffusione all'esterno. I macrodati e i microdati diffusi devono essere preventivamente trattati per garantire una adeguata tutela della riservatezza. Il calendario di diffusione dei risultati statistici deve essere reso pubblico. Tutte le fasi del processo devono essere adeguatamente documentate.

Linee guida

I microdati validati devono essere archiviati insieme ai metadati necessari per la loro interpretazione (tracciati record, variabili e classificazioni associate) nei sistemi di Istituto[6], seguendo la procedura definita dall'Istituto, anche nel caso in cui vengano utilizzati dati di fonte amministrativa all'interno del processo produttivo. Essendo i microdati di fonte amministrativa frequentemente un prodotto intermedio, che viene utilizzato come input di altri processi produttivi statistici, è importante misurarne la qualità attraverso specifici indicatori, quali per esempio copertura, dati mancanti, tempestività e puntualità.

L'obiettivo della diffusione è quello di consentire un uso tempestivo ed efficace dell'informazione prodotta dall'Istituto, rispondendo così alle esigenze degli utenti. A tal fine è utile definire in anticipo un calendario di diffusione relativo ai vari tipi di rilasci, il quale dovrebbe essere reso pubblico agli utenti. L'accesso ai dati diffusi deve essere simultaneo per tutti gli utenti in modo da garantire l'imparzialità e l'indipendenza della statistica ufficiale.

Per consentire una migliore fruizione dei dati da parte degli utenti è importante diffondere dati che siano facilmente accessibili e comprensibili. L'accessibilità è legata al tipo di supporto utilizzato (diffusione on-line, CD-Rom, volume cartaceo) e alla facilità di reperimento dell'informazione. Date le attuali direttive nazionali, ed europee, Internet è diventata la modalità prevalente di diffusione, sia attraverso lo sviluppo di datawarehouse, sia attraverso la pubblicazione di documenti, comunicati e volumi on-line. La chiarezza, invece, è legata alla disponibilità di metadati relativi ai contenuti informativi e alle caratteristiche del processo di produzione, e di indicatori di qualità. I metadati di supporto devono essere integrati con gli elementi che consentano di comprendere come sono stati usati i dati amministrativi e la loro validità nel contesto produttivo specifico. Inoltre, devono essere comunicate eventuali limitazioni dei dati, quali ad esempio l'esistenza di interruzioni nelle serie storiche e l'eventuale carattere provvisorio dei dati rilasciati.

I vari tipi di rilascio, per esempio comunicati stampa ed annuari, devono rispettare gli standard editoriali, così come i file di microdati messi a disposizione degli utenti devono aderire alle varie tipologie rilasciabili (file ad uso pubblico, file standard, file per la ricerca, file per il Sistan).

La legge istitutiva del Sistema statistico nazionale, il D.to L.vo 322/89, prevede che debba essere tutelata la riservatezza dei rispondenti, e, in particolare, che i dati oggetto di diffusione debbano essere adeguatamente trattati a tal fine. Nel caso di diffusione di dati aggregati in tabelle possono essere utilizzati alcuni metodi come la regola della soglia, che viene posta come uguale o superiore a tre, e i metodi che consistono nel perturbare i dati in modo da ridurre la possibilità di identificazione ed acquisizione di informazioni sulle singole unità. Nel caso di diffusione di dati elementari si possono utilizzare metodi specifici, quali la ricodifica delle variabili per ridurre il dettaglio informativo, la soppressione di specifiche informazioni che

[6] Per i dati prodotti dalle indagini, gli standard di Istituto prevedono l'archiviazione nel repository ARMIDA (ARchivio dei MIcroDAti validati) nasce con l'obiettivo primario di conservare e documentare i dati, a cui si è successivamente affiancato l'obiettivo di diffondere i dati stessi. I dati archiviati in ARMIDA alimentano, infatti, i diversi canali di diffusione dei microdati (per usi interni all'Istituto attraverso il "Protocollo d'accesso ai microdati di ARMIDA" per gli utenti interni", per gli enti del Sistan, per i file per la ricerca, per i file standard, eccetera). I microdati archiviati in ARMIDA vengono, inoltre, utilizzati per rispondere alle richieste di utenti esterni presentate presso il laboratorio ADELE.

possono rendere identificabile un'unità, e metodi di perturbazione dei dati elementari. Per la tutela della riservatezza nella diffusione dei dati è opportuno usare software generalizzato.

Il processo di produzione deve essere opportunamente documentato, relativamente a tutte le fasi, dall'analisi e scelta della fonte amministrativa, all'integrazione dei dati nel processo produttivo statistico, al trattamento dei dati integrati, fino alla diffusione.

La documentazione deve includere indicatori di qualità, quali ad esempio indicatori di tempestività, di copertura e di dati mancanti, di coerenza e di confrontabilità nel tempo.

Alcuni riferimenti bibliografici

Hundepol A., Domingo-Ferre J., Franconi L., Giessing S., Lenz R., Naylor J., Nordholt E.S., Seri G., De Wolf P.P. (2010). Handbook on Statistical Disclosure Control. Version 1.2. ESSNet SDC – A network of excellence in the European Statistical System in the fields of Statistical Disclosure Control http://neon.vb.cbs.nl/casc/.%5CSDC_Handbook.pdf

Istat (2004). Metodologie e tecniche di tutela della riservatezza nel rilascio di informazione statistica. Metodi e Norme, n. 20 http://www3.istat.it/dati/catalogo/20040706_00/manuale-tutela_riservatezza.pdf

Istat (2008). Protocollo d'accesso ai microdati di Armida per gli utenti interni https://intranet.istat.it/MetadatiEQualita/Documents/Microdati%20validati/ProtocolloArmida.pdf

Istat (2009). Standard di documentazione e di memorizzazione dei file di microdati per la ricerca (MFR), Servizio SID, 26 giugno 2009 https://intranet.istat.it/Documentazione/Procedure/MFRStandard.pdf

Istat (2009). Procedura per il rilascio di file di dati elementari agli uffici Sistan, Ordine di servizio n. 148 del 17 novembre 2009 della Direzione Generale https://intranet.istat.it/Documentazione/Procedure/Procedurarilascio_sistan.pdf

OMB (2006). Standards and Guidelines for Statistical Surveys. Office for Management and Budget, The White House, Washington, USA. http://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/statpolicy/standards_stat_surveys.pdf

Parte C. Qualità del prodotto (*Output quality*)

1. Introduzione

Nelle parti A e B sono stati elencati i principi e le linee guida da seguire rispettivamente nella valutazione dei dati amministrativi di input e nella conduzione di un processo che li utilizza al fine di produrre, in modo efficiente, statistiche caratterizzate da un'elevata qualità. Tuttavia, aver impostato e condotto un processo di elevata qualità, non implica che la qualità delle statistiche prodotte non debba essere misurata. Così come fatto nelle Linee Guida per i processi produttivi statistici di tipo "rilevazione diretta", in questa sezione si riportano i criteri rispetto ai quali misurare la qualità delle statistiche prodotte e rispetto ai quali comunicarla agli utenti, tenendo conto della specificità della produzione del dato, ossia introducendo elementi relativi all'impatto che l'uso di dati di fonte amministrativa può avere nella misurazione o nei livelli delle misure di qualità. Non vengono, invece, fornite vere e proprie linee guida su come condurre la misurazione, che richiedono approfondimenti di carattere metodologico per i quali si rimanda alla letteratura specializzata.

2. La definizione e le dimensioni della qualità di prodotto

Per prima cosa è opportuno ribadire il significato e l'ambito relativamente al quale si definisce la qualità. Ai fini di questa trattazione, per prodotto si intende il prodotto finale derivante dal processo produttivo statistico o output (termine spesso utilizzato ormai anche nella terminologia italiana). Tale prodotto ha una natura statistica (distribuzioni, stime di livello, variazioni, etc.). Non si considerano qui prodotti che potremmo definire "intermedi" quali i registri statistici.

Ai fini della misurazione della qualità delle statistiche, l'Istat ha adottato la definizione della qualità rilasciata da Eurostat nel 2003 (ESS Working Group "Assessment of Quality in Statistics"), successivamente ripresa dal Codice di condotta delle statistiche europee (European Statistics Code of Practice, 2011) e dal Codice italiano delle statistiche ufficiali (Gazz. Uff. 13 ottobre 2010, n. 240)[7]. Queste definizioni sono state ulteriormente chiarite nei manuali Eurostat di riferimento per il *quality reporting* (Eurostat, 2009; Eurostat, 2014).

La qualità viene definita come "il complesso delle caratteristiche di un prodotto o di un servizio che gli conferiscono la capacità di soddisfare i bisogni impliciti o espressi" (Eurostat 2002, Eurostat, 2003a). In tal senso, la qualità delle statistiche prodotte e diffuse deve essere valutata con riferimento ai seguenti criteri (Eurostat, 2003a):

- pertinenza
- accuratezza
- tempestività e puntualità
- accessibilità e chiarezza
- comparabilità
- coerenza

Successivamente, all'accuratezza è stato aggiunta anche la componente dell'attendibilità. In queste linee guida sono state tradotte le definizioni dei criteri di qualità di Eurostat più aggiornate, contenute nel *ESS Handbook for quality reports* (Eurostat, 2014). In alcuni casi queste definizioni sono state riformulate e sintetizzate senza alterarne il significato.

[7] Tale definizione di qualità ha assunto una notevole importanza in quanto è stata inclusa nel quadro giuridico (regolamento (CE) No. 223/2009 del Parlamento Europeo e del Consiglio dell'11 marzo 2009[7]) che regola la produzione delle statistiche europee.

Definizione C.1. Pertinenza

La *pertinenza* è definita come il grado con cui l'informazione statistica soddisfa le esigenze attuali e potenziali degli utenti. Essa comprende la completezza dell'informazione prodotta (tutte le statistiche necessarie agli utenti devono essere prodotte) e il livello in cui i concetti utilizzati (definizioni, classificazioni…) riflettono le esigenze degli utenti.

Definizione C.2.1. Accuratezza

L'*accuratezza* di una statistica viene definite da un punto di vista statistico come il grado di vicinanza tra la stima e il valore vero che la statistica intende misurare.

Definizione C.2.2. Attendibilità

L'*attendibilità* si riferisce alla vicinanza del valore della stima iniziale diffusa ai valori successivi relativi alla stessa stima.

Definizione C.3. Tempestività e puntualità

La *tempestività* dei risultati è definita come il periodo di tempo che intercorre tra l'evento o il fenomeno che i risultati descrivono e il momento in cui gli stessi vengono resi disponibili.

La *puntualità* è il periodo di tempo tra la data del rilascio dei dati e quella pianificata da calendario, da Regolamento o da accordo preventivo tra partner.

Definizione C.4. Coerenza e comparabilità

La *coerenza* misura l'adeguatezza delle statistiche ad essere combinate in modo diverso e per diversi usi. Si parla di riconciliabilità tra statistiche all'interno di una stessa fonte relative a variabili diverse, calcolate su domini diversi, da fonti diverse o da processi con periodicità diverse.

La *comparabilità* nel tempo e geografica è una misura di quanto le differenze nel tempo e tra aree geografiche siano dovute a variazioni reali e non a differenze in: concetti statistici, strumenti di misurazione e procedure.

Definizione C.5. Accessibilità e chiarezza

L'*accessibilità* delle statistiche è la facilità con cui gli utenti possono ottenere i dati. Essa è determinata dalle condizioni attraverso cui gli utenti ottengono i dati: dove recarsi, come richiederli, tempi di consegna, politica dei prezzi, politica di diffusione, disponibilità di micro o macrodati, formati disponibili (carta, files, CD-ROM, Internet…).

La *chiarezza* delle statistiche è la facilità con cui gli utenti vengono messi in grado di capire i dati. Essa è determinata dal contesto informativo in cui vengono presentati i dati, se sono accompagnati da metadati appropriati, se vengono utilizzate illustrazioni quali grafici o mappe, se sono disponibili informazioni sull'accuratezza dei dati (incluse eventuali limitazioni d'uso) e fino a che punto viene fornita assistenza aggiuntiva dal produttore del dato.

3. La misurazione della qualità delle statistiche prodotte usando dati di fonte amministrativa

Ancor più per le statistiche che utilizzano dati di fonte amministrativa, misurarne la qualità in base alle componenti elencate nel precedente paragrafo non è affatto semplice. Infatti, solo alcune componenti si prestano ad una misurazione quantitativa diretta, in particolare tempestività e comparabilità nel tempo, mentre per le altre dimensioni spesso si possono solo formulare dei giudizi. Relativamente all'accuratezza, bisogna sottolineare che la componente dell'errore campionario in genere non si applica all'uso dei dati amministrativi e l'errore non campionario si caratterizza in modo differente rispetto alla situazione da indagine diretta.

Nel seguito verranno analizzate le componenti della qualità e come l'uso dei dati amministrativi può alterare la loro interpretazione e la misurazione degli indicatori relativi. Quindi, verranno elencati e definiti i principali errori non campionari che si generano nel processo che utilizza dati amministrativi, così come rappresentati nella Figura 2 del paragrafo 1 "Il quadro di riferimento per la qualità dei processi statistici che utilizzano dati di fonte amministrativa", utili come riferimento per l'identificazione di misure indirette dell'accuratezza del dato statistico che ne deriva.

3.1. Misurare le componenti della qualità per processi che utilizzano dati di fonte amministrativa

Possiamo affermare che l'uso dei dati amministrativi non altera il significato delle componenti della qualità: le statistiche prodotte saranno soggette a considerazioni sulla pertinenza, accuratezza e affidabilità, tempestività e puntualità, coerenza e comparabilità, accessibilità e chiarezza esattamente come le statistiche dirette da indagine. L'uso dei dati amministrativi per finalità statistiche influenza però particolarmente alcune componenti della qualità, e a volte limita la capacità del ricercatore di valutarle.

La Pertinenza e la Coerenza e Comparabilità, possono essere fortemente influenzate dall'uso di fonti amministrative direttamente per la produzione statistica (sostituzione di unità e/o variabili). In particolare, seguendo il modello introdotto nel Paragrafo 1 "Il quadro di riferimento per la qualità dei processi statistici che utilizzano dati di fonte amministrativa", errori di specificazione, ossia di non corrispondenza tra concetti obiettivo statistico e i relativi concetti sottostanti i dati amministrativi, possono introdurre nel prodotto statistico derivato, errori di pertinenza. Anche laddove tali errori non si presentino, variazioni legislative o procedurali che impattano sulle popolazioni e sulle variabili amministrative possono causare mancanza di comparabilità nei dati statistici prodotti a partire dal dato amministrativo. I processi di integrazione tra archivi possono introdurre errori di coerenza tra fonti, che si vanno ad aggiungere ai possibili problemi di coerenza interna alle fonti stesse, che non è stato possibile riconciliare attraverso le opportune procedure di armonizzazione e di controllo e correzione.

Gli indicatori di tempestività e puntualità rimangono perfettamente validi anche nel caso di statistiche prodotte utilizzando dati amministrativi. L'uso di dati amministrativi può impattare sulla tempestività in entrambe le direzioni: laddove la disponibilità e fornitura del dato amministrativo sia idonea alle esigenze di produzione statistica, il suo uso può portare a guadagni di tempestività. Viceversa, se il dato amministrativo è disponibile in ritardo rispetto alle esigenze produttive, ne può derivare una tempestività delle statistiche prodotte peggiore, rispetto alla situazione ipotetica di una indagine diretta.

Considerando l'accessibilità e chiarezza, si può affermare che l'uso dei dati amministrativi non ha impatto sulla misurazione di queste componenti. Per garantire la trasparenza nei confronti degli utenti, sarà opportuno documentare quali fonti amministrative sono state utilizzate e come.

Sicuramente, la componente che presenta maggiori difficoltà di misurazione nelle indagini dirette così come nei processi che utilizzano dati amministrativi è quella della accuratezza. Spesso i dati amministrativi sono utilizzati in combinazione con quelli rilevati da indagini dirette, e quindi anche gli stimatori utilizzati

possono avere forme diverse, che variano da situazioni in cui i dati sono integrati a livello micro fino a casi in cui lo stimatore è una media ponderata di stimatori da indagine e stimatori da dati amministrativi.

3.1.1 Accuratezza e attendibilità

Il livello di accuratezza dei risultati è legato alla quantità di errori che possono manifestarsi nel processo di produzione delle stime. Una misura dell'accuratezza è fornita dell'Errore Quadratico Medio (*Mean Square Error*, MSE nella letteratura anglosassone), che include variabilità e distorsione per tutte le componenti dell'errore campionario (non applicabile o non rilevante nel contesto dell'uso dei dati amministrativi) e non campionario.

Per le statistiche che utilizzano dati amministrativi, non sono disponibili in letteratura molte esperienze di stima dell'accuratezza delle statistiche prodotte. Laitila e Holmberg (2010) propongono una scomposizione dell'Errore Quadratico Medio che tiene anche conto della distorsione da rilevanza oltre a quelle classiche dello stimatore, da mancata risposta e da errore di misura. Sotto alcune ipotesi, essi forniscono un metodo di confronto dell'accuratezza delle stime da registro rispetto a quelle da indagine campionaria e mostrano che la maggior distorsione dello stimatore attribuibile all'errore di rilevanza può essere compensata dall'assenza di variabilità campionaria.

Altresì interessante risulta essere il lavoro condotto nell'ambito del WP 6 dell'ESSnet "Use of Administrative and Accounts Data in Business Statistics" (Deliverable 6.3., 2011) focalizzato sull'identificazione di misure di accuratezza per fonti "miste" in differenti situazioni di integrazione micro e macro. Stime della varianza e distorsione attraverso metodi di "bootstrap re-sampling" sono anche possibili. Stimatori e relativi intervalli di credibilità possono essere derivati attraverso il ricorso a metodi Bayesiani.

Interessante anche se non ancora applicabile a regime, l'uso di Modelli ad Equazioni Strutturali per la stima della validità delle variabili amministrative e di indagine (Scholtus S., Bakker B.F.M., 2013).

Statistiche prodotte utilizzando dati amministrativi sono potenzialmente valutabili attraverso il calcolo di misure di attendibilità, ossia attraverso indicatori utilizzati nell'ambito dell'analisi delle revisioni in statistiche congiunturali. Si ricorda che le revisioni sono aggiornamenti programmati e successivi di una stima dovuti a fattori di aggiornamento (utilizzo di fonti più aggiornate, aggiustamenti per stagionalità) o di miglioramento (nelle definizioni, nei metodi). Quindi laddove si rendano disponibili dati amministrativi con diverso livello di completezza e aggiornamento, la qualità delle statistiche prodotte con i diversi archivi è valutabile attraverso indicatori di revisione. Si veda il sito dell'OECD e quello di ONS per una trattazione esaustiva dell'argomento.

Così come fatto per statistiche di indagine, in condizioni di difficoltà nel derivare misure affidabili ed economiche dell'errore quadratico medio, spesso una valutazione della qualità dei risultati viene effettuata andando ad analizzare le diverse componenti di errore che impattano sull'accuratezza delle stime. Qui di seguito si riportano le descrizioni dei principali errori nel caso di statistiche prodotte utilizzando dati amministrativi, sempre facendo riferimento alla Figura 2 del paragrafo 1 "Il quadro di riferimento per la qualità dei processi statistici che utilizzano dati di fonte amministrativa".

3.1.2. Errori sulle unità o popolazione

<u>Non corrispondenza concettuale tra popolazione statistica e popolazione amministrativa</u>

Se a livello definitorio, le unità che compongono la popolazione obiettivo statistico non corrispondono con quelle sottostanti l'archivio amministrativo che le dovrebbe contenere, questa mancata corrispondenza può portare ad una sottocopertura o ad una sovracopertura. Tuttavia, i problemi di copertura possono non esaurirsi solo in questo stadio, ma possono originarsi dal processo di derivazione della popolazione statistica

a partire da quella amministrativa attraverso procedure di integrazione, derivazione, ricodifica. In questo caso la copertura includerà anche altre componenti che possono derivare da errori durante queste fasi (si veda oltre). Supponiamo di avere un obiettivo statistico sulla popolazione dei "permessi di costruire" definiti come: progetti di fabbricati nuovi (residenziali e non residenziali), di ampliamenti di fabbricati preesistenti, progetti esecutivi per la realizzazione di fabbricati o ampliamenti destinati a edilizia pubblica. Se la fonte amministrativa disponibile avesse come popolazione di riferimento i permessi di costruire o di ampliamento per i soli fabbricati residenziali e non residenziali, si avrebbe un errore di copertura. Nella terminologia utilizzata in Zhang (2012), l'insieme delle unità relative ai progetti di edilizia pubblica "non è accessibile".

Errori di selezione

Gli errori di selezione hanno origine dalla discrepanza tra l'insieme dei dati di interesse accessibile (o osservabile) e quello concretamente acquisito. Riprendendo l'esempio ipotetico sui permessi di costruire, supponiamo che tutti i permessi di costruire siano presentati ai comuni, alcuni in forma cartacea (residenziali e non residenziali) altri in forma elettronica (edilizia pubblica) e che siano quindi trasmessi dai comuni all'Istat. Se il sistema di trasmissione elettronica subisce un malfunzionamento, e quindi una parte dei permessi sfugge all'invio all'Istat, siamo in presenza dell'errore di selezione. Ancora una volta, nell'efficace terminologia di Zhang (2012), l'insieme delle unità relative ai progetti di edilizia pubblica sarebbe "accessibile ma non acceduto". Gli errori di selezioni si concretizzano in dati mancanti e duplicazioni. Anche i ritardi nella trasmissione dei dati vengono inclusi in questa categoria.

Errori di *linkage* o errore di abbinamento

Nell'integrare archivi diversi di dati amministrativi tra di loro e con dati di indagine, i principali errori che si commettono sono *i)* falsi non abbinamenti e *ii)* falsi abbinamenti. È evidente che la qualità dei risultati dell'abbinamento dipende fortemente dalla qualità delle chiavi di abbinamento utilizzate. Gli errori di abbinamento hanno un impatto su altre componenti dell'errore. Principalmente, i falsi non abbinamenti possono portare ad errori di copertura ed errori di dati mancanti, mentre i falsi abbinamenti possono portare a errori di coerenza nelle variabili.

Errori di derivazione

Questa categoria comprende varie tipologie di errore che si possono generare durante il processo derivazione delle unità. In particolare, errori nella creazione di unità ex-novo, ossia unità non esistenti come tali nelle fonti amministrative, come per esempio è il caso della unità statistica "azienda agricola", per la quale è necessario cercare segnali di esistenza nella varie fonti amministrative. Sono compresi anche errori nell'identificazione dell'unità statistica a partire da unità amministrative. Infine sono anche inclusi errori di allineamento tra unità "composte" e unità "base". Per esempio, se si hanno a disposizione dati a livello di individuo in diversi data set e si vuole formare una lista di individui residenti per residenza, se i diversi data set contengono indirizzi diversi, condizione che deve portare ad una decisione sull'indirizzo da assegnare ad ogni individuo, si possono commettere errori nell'identificazione del soggetto residente. Da sottolineare che le definizioni qui introdotte si discostano leggermente da quelle suggerite in Zhang (2012) e ne rappresentano una semplificazione.

Errori di copertura

Gli errori di copertura sono errori derivanti da discrepanze tra la popolazione statistica obiettivo e quella derivata nel processo produttivo che utilizza dati amministrativi. Quest'ultima, nella situazione più semplice, può coincidere con l'unità di riferimento di un archivio amministrativo, oppure può essere il risultato di procedure di integrazione e derivazione di unità che possono essere molto complesse.

All'interno dell'errore di copertura si possono identificare le usuali categorie di errori: *i)* sottocopertura, ovvero oggetti che appartengono alla popolazione obiettivo ma non sono elencati nell'archivio (o negli

archivi integrati), che rappresentano una potenziale fonte di distorsione; *ii)* sovracopertura, ossia oggetti presenti nell'archivio (o negli archivi integrati) ma non appartenenti alla popolazione obiettivo degli oggetti, che rappresentano una potenziale fonte di variabilità; *iii)* errori nelle variabili identificative degli oggetti, che possono dare luogo a successivi errori di integrazione e di coerenza.

Quando si utilizzano dati amministrativi, la copertura è il risultato di un insieme di possibili componenti quali: la copertura dell'archivio rispetto alla popolazione amministrativa (che a volte è riferita ad oggetti amministrativi del tipo evento), la copertura tra popolazione obiettivo statistico e popolazioni potenzialmente rilevate dall'archivio amministrativo, la copertura tra popolazione rilevata dall'archivio e popolazione acquisita (errore di selezione), la popolazione obiettivo statistico e quella derivata dal processo di integrazione e derivazione. Si ricorda infatti, che errori di abbinamento nella procedura di integrazione, ed in particolare falsi non abbinamenti possono causare errori di sottocopertura. Anche possibili ritardi nell'acquisizione dei dati e l'errore di derivazione dell'unità possono portare ad errori di copertura.

3.1.3. Errori sulle variabili

Errori di specificazione

Così come nel caso di indagini dirette gli errori di specificazione derivano da una non corrispondenza tra obiettivi conoscitivi di indagine e concetti rilevati attraverso i quesiti del questionario, nell'ambito dell'utilizzo dei dati amministrativi, questi errori sono riferiti a discrepanze tra il concetto obiettivo teorico statistico e quello amministrativo. È forse il tipo di errore che ha il maggior impatto sulla pertinenza delle statistiche prodotte, e può causare errori di accuratezza, in particolare sulla componente della distorsione.

Non stabilità concettuale

Variazioni legislative o procedurali che regolano l'atto amministrativo e che modificano i concetti sottostanti il dato amministrativo possono portare a errori di comparabilità nel tempo. Non omogeneità sul territorio della legislazione o delle procedure di acquisizione e trattamento dei dati amministrativi possono portare ad errori di comparabilità geografica. Gli errori di comparabilità temporale e geografica si propagano nelle statistiche prodotte con i dati amministrativi, tuttavia all'origine non vi è necessariamente un errore nelle fonti utilizzate, ma semplicemente delle variazioni (temporali e geografiche).

Errori di misurazione

Sono errori di osservazione che possono verificarsi nella fase di raccolta (*errori di misurazione in senso stretto*). In pratica il valore disponibile per una data variabile, al momento dell'acquisizione dell'archivio non corrisponde al valore reale. Tali errori possono essere sia fonte di distorsione che di incremento della variabilità associata alle stime.

Errori di processo

Sono errori che si generano nel trattamento del dato all'interno di un dato processo produttivo statistico (revisione, registrazione, codifica, controllo, elaborazione, ecc.). Una parte di questi errori potrebbe derivare anche da trattamenti precedenti all'utilizzo condotti dall'ente titolare o, più raramente dal settore centralizzato di acquisitone del dato. Infatti, il pre-trattamento in fase di acquisizione centralizzata, in genere, non modifica i dati acquisiti ma li integra con informazioni aggiuntive.

Errori di classificazione

Sono gli errori che si commettono nell'allineare le classificazioni usate nel dato amministrativo con quelle che si intende applicare alle variabili statistiche di interesse, ossia nel ricondurre ogni modalità della

classificazione amministrativa ad una modalità della classificazione statistica. Errori di classificazione relativi a variabili di stratificazione possono portare a problemi di copertura.

Errori di coerenza

Gli errori di coerenza derivano dalla violazione di una serie di regole di compatibilità all'interno di una singola fonte o tra più fonti integrate. Nel primo caso si parla di coerenza intra-fonte, nel secondo di coerenza inter-fonte. Quest'ultime possono o meno derivare da falsi abbinamenti nella procedure di integrazione.

Errori di dati mancanti

Si tratta di errori assimilabili a quelli che nelle statistiche da indagine vengono denominati mancate risposte parziali, in quanto la mancata risposta totale rientra, per dati amministrativi, negli errori di copertura. Tipicamente la mancata osservazione di dati si osserva in modo completo per alcune variabili che non sono di interesse per le finalità amministrative, mentre non risulta essere rilevante per le variabili di stretto interesse amministrativo. Questo tipo di errore può risultare anche quando si integrino variabili presenti in archivi differenti, come effetto del processo di integrazione tra archivi, laddove l'oggetto che si integra non sia presente in tutte le fonti. L'impatto sulle stime di questo tipo di errore è in termini di aumento della variabilità e possibile distorsione.

Errore di specificazione da modello implicito o esplicito

Gli errori da assunzione di modello si verificano tutte le volte che un modello viene introdotto, in genere, per aggiustamenti (dati mancanti, effetti stagionali) e per operazioni di stima. Le statistiche prodotte utilizzando dati amministrativi fanno tipicamente ampio ricorso ad approcci basati su modello, durante le attività di integrazione, derivazione di unità e variabili, controllo e correzione, stima. In genere un modello è un insieme di assunzioni sulle relazioni tra dati osservati e dati non osservati. La validità delle assunzioni del modello andrebbe verificata anche se non si hanno a disposizione strumenti per valutarla con certezza. Per valutare l'impatto delle assunzioni del modello si può ricorrere ad analisi di sensibilità, attraverso simulazioni.

3.2. Indicatori di qualità

Le difficoltà insite nella misurazione dell'accuratezza o dell'attendibilità e, più in generale, delle singole componenti della qualità fa sì che l'approccio alla misurazione della qualità maggiormente utilizzato consista in un compromesso: alle poche misurazioni dirette si affiancano delle misurazioni indirette. Nel caso dell'uso dei dati amministrativi, tali misurazioni possono essere definite sia sui dati di input che per le fasi di integrazione dei dati in specifici processi produttivi statistici. Gli indicatori sui dati di input, affiancati ad altri indicatori maggiormente di contesto sulla fonte, forniscono un quadro concreto sulla utilizzabilità del dato amministrativo per finalità statistiche.

Sui dati di input, la letteratura più rilevante è costituita dalle evidenze prodotte nell'ambito del progetto FP7 "Blue – Enterprise and Trade Statistics (Blu-Ets)", che coerentemente con l'approccio che definisce la qualità dell'input attraverso tre principali iperdimensioni (fonte, metadati e dati), ha sviluppato per la iperdimensione "dati" e per le sue diverse dimensioni, indicatori di qualità insieme a metodi di misura. Tali indicatori sono classificati come applicabili "indipendentemente dagli obiettivi di produzione statistica previsto" oppure orientati a valutare la qualità in relazione a specifici obiettivi statistici. A livello internazionale è da segnalare lo Statistical Network "Methodologies for an integrated use of administrative data in the statistical process - Administrative data (MIAD)", coordinato dall'Istat, che ha approfondito i diversi usi del dato amministrativo e definito un framework per la qualità dei dati amministrativi, insieme ad indicatori per la fase di scouting e per la fase di acquisizione e linee guida per il loro calcolo. L'ESSnet

AdminData ha prodotto una lista ragionata di indicatori di qualità, per ogni dimensione Eurostat della qualità, da adottare quando si usano dati amministrativi negli output statistici con riferimento alle statistiche economiche sia strutturali che congiunturali (ESSnet AdminData, 2013)[8].

Il calcolo sistematico di indicatori sulle fasi di integrazione e trattamento nel processo statistico, specificatamente quando queste coinvolgano l'uso di dati amministrativi, sono meno consolidati. Molti Istituti Nazionali di Statistica, tra cui l'Istat stanno lavorando nel definire indicatori che riflettono gli errori descritti nel paragrafo precedente.

Alcuni Riferimenti Bibliografici

Bakker B.F.M. (2010). Micro-integration: State of the Art. Note by Statistics Netherlands. UNECE Conference of European Statisticians. The Hague, The Netherlands, 10-11 May 2010

Bryant J.R., Graham P.J. (2013). Bayesian Demographic Accounts: subnational Population

Daas P., Ossen S. (2011). Report on methods preferred for the quality indicators of administrative data sources, Blue – ETS Project, Deliverable 4.2.

ESSnet Use of Administrative and Accounts Data in Business Statistics (2013). WP6 Quality Indicators when using Administrative Data in Statistical Outputs. Deliverable 6.3/2011: Guidance on the accuracy of mixed-sources statistics (disponibile sul sito https://ec.europa.eu/eurostat/cros/content/admindata-essnet-use-administrative-and-accounts-data-business-statistics_en)

Eurostat (2014). ESS Handbook for quality reports. 2014 Edition. http://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf

Eurostat (2011) "European Statistics Code of Practice – revised edition 2011". http://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF

Eurostat (2009). ESS Handbook for quality reports. 2009 Edition.

Eurostat (2003a) "Definition of quality in statistics". Working group "Assessment of quality in statistics", Luxembourg, 2-3 October 2003. http://ec.europa.eu/eurostat/documents/64157/4373735/02-ESS-quality-definition.pdf

Eurostat (2002). Quality in the European Statistical System – The way Forward, 2002 Edition (Leg on Quality) Luxembourg

Laitila T. Holmberg A. (2010). Comprison of Samle and Register Survey Estimators via MSE Decomposition (Q2014)

Methodologies for an Integrated Use of Administrative Data, MIAD (2015) https://ec.europa.eu/eurostat/cros/content/miad-methodologies-integrated-use-administrative-data-statistical-process_en

OECD. OECD / Eurostat Guidelines on Revisions Policy and Analysis. http://www.oecd.org/std/oecdeurostatguidelinesonrevisionspolicyandanalysis.htm

ONS. Revision and correction policy (last update: July 2011). http://www.ons.gov.uk/ons/guide-method/revisions/revisions-and-corrections-policy/index.html

Sholtus S., Bakker F.M. (2013). Estimating the validity of administrative and survey variables through structural equation modeling. A simulation study on robustness. Discussion paper (201302). Statistics Netherlands.

Zhang L.C. (2012). Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica (2012) Vol 66, nr.1, pp. 41-63.

8 Tailored list of quality indicators: Structural Business Statistics (Annex 1a) e Tailored list of quality indicators: Short Term Statistics (Annex 1b).

Appendice. Linee guida per l'acquisizione e gestione centralizzata di un archivio amministrativo

Principio 1. Scoperta di nuove fonti e loro conoscenza
É opportuno che vi sia un'attenzione particolare all'emergere di nuove fonti di dati amministrativi, rispetto alle quali è indispensabile la conoscenza approfondita di tutti gli aspetti legislativi e procedurali che regolano il ciclo di vita del dato amministrativo, precondizione per il loro utilizzo all'interno della produzione statistica.

Linee guida

Il Codice di condotta delle statistiche europee raccomanda di ridurre l'onere statistico sui rispondenti, e di migliorare il rapporto costi\efficacia nella raccolta dei dati (Eurostat, 2011). Ciò avvalora il pieno sfruttamento delle fonti amministrative esistenti e l'attenzione all'emergere di nuove fonti non ancora esplorate per l'utilizzo a fini statistici.
La legislazione nazionale in materia è in linea con tali indicazioni ed obbliga le amministrazioni pubbliche *"..che dispongano di archivi, anche informatizzati, contenenti dati…che siano utili ai fini di rilevazioni statistiche, a consentire all'Istat di accedere ai detti archivi e alle informazioni individuali ivi contenute... L'accesso avverrà secondo modalità concordate tra le parti*." (L.322/89 e successive modifiche, L.681/96, art.1 comma 8).

È opportuno condurre in modo sistematico, all'interno dell'Istituto, un'attività di "scouting" di nuove fonti, funzionalmente organizzata, che possa avvalersi anche dell'ausilio di organismi a ciò preposti e che preveda un processo di condivisione con i potenziali utilizzatori all'interno dell'Istituto. Un contributo importante può pervenire da parte delle comunità di esperti di settore che possono mettere le loro esperienze e conoscenze al servizio delle strutture che gestiscono l'attività di ricognizione delle fonti.

Per la conduzione di questa attività di scoperta di nuove fonti, è opportuno instaurare contatti con i diversi Enti operanti sul territorio e acquisire informazioni in forma strutturata[9] (D'Angiolini *et al*., 2014a; D'Angiolini *et al*., 2014b). Tali informazioni dovrebbero coprire prevalentemente le due iperdimensioni attinenti la qualità dell'input: quella della Fonte e quella dei Metadati (Daas *et al*., 2009, Iwig *et al*., 2013).

In particolare, è necessario reperire informazioni sull'ente fornitore, sulle finalità e usi attuali dell'archivio, sugli aspetti legati alla sicurezza e privacy, sulle condizioni legate alla fornitura dei dati, e su tutti gli aspetti procedurali e tecnici relativi all'acquisizione, trattamento e archiviazione dei dati da parte dell'ente titolare, per esempio: la modulistica utilizzata, le modalità di compilazione dell'atto amministrativo (se compilato dal diretto interessato o da un operatore; se cartaceo o informatizzato), gli standard di archiviazione.

È importante acquisire tutti i metadati concettuali che permettono la comprensione dei dati amministrativi contenuti nell'archivio, e cioè una documentazione completa e accurata degli oggetti (unità, eventi) e delle variabili contenute nell'archivio, che può richiedere un approfondimento della legislazione che regola il ciclo di vita del dato amministrativo; per questo è importante avvalersi del supporto della comunità di esperti di settore. Particolare importanza ricoprono le informazioni relative al periodo o data di riferimento del dato amministrativo e alla tempestività e periodicità di aggiornamento dell'archivio.

[9] L'Istat ha avviato una attività di coordinamento e di armonizzazione della modulistica, condotta attraverso istruttorie formalizzate sugli archivi amministrativi degli enti centrali. La documentazione raccolta viene archiviata in un apposito sistema denominato DARCAP (https://darcap.istat.it/darcap.php)

Alcuni riferimenti bibliografici

Daas P., Ossen S., Vis-Visschers R., and Arends-Tóth J. (2009). *Checklist for the Quality evaluation of Administrative Data Sources*, Statistics Netherlands, The Hague /Heerlen, 2009

D'Angiolini G., De Salvo P., Passacantilli A. (2014a) Istat's new strategy and tools for enhancing statistical utilization of the available administrative databases. European Conference on Quality in Official Statistics (Q2014). Vienna, Austria, 2-5 June 2014

D'Angiolini G., Patruno E., Saccoccio T., De Rosa C., Valente E.(2014b). DARCAP: A tool for documenting the information content and the quality of the available administrative data sources. European Conference on Quality in Official Statistics (Q2014). Vienna, Austria, 2-5 June 2014

Decreto Legislativo 6 Settembre 1989, N. 322 "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"

Eurostat – Sistema statistico europeo, Codice delle statistiche europee, 28 settembre 2011. http://www.istat.it/it/files/2011/01/statistics_code.pdf

Iwig W. Berning M., Marck P., Prell M. (2013). Data Quality Assessment Tool for Administrative Data. http://www.bls.gov/osmr/datatool.pdf

Legge 31 dicembre 1996, n. 681, " Finanziamento del censimento intermedio dell'industria e dei servizi nell'anno 1996"

Principio 2. Valutazione preliminare sull'opportunità dell'acquisizione

La valutazione preliminare sull'opportunità dell'acquisizione di un archivio amministrativo deve essere guidata da un'analisi approfondita della sua rilevanza attuale e potenziale, del rapporto costi/benefici, della sua stabilità e della sua qualità attesa.

Linee guida

La valutazione preliminare sull'opportunità o meno di acquisire un archivio amministrativo, o parte di esso, deve essere condotta sulla base di criteri oggettivi, siano essi fondati su informazioni provenienti dalle istruttorie cui si è accennato nel Principio 1, qualora disponibili, o *check-list* effettuate ad hoc.

L'istituzione di un tavolo di confronto, tra l'Istituto Nazionale di Statistica e l'ente titolare della fonte amministrativa, è un elemento di supporto al processo decisionale e alla definizione di eventuali accordi per le procedure di acquisizione.

La valutazione deve riguardare vari livelli: la rilevanza dei dati amministrativi contenuti nell'archivio rispetto agli obiettivi conoscitivi e/o produttivi, i costi di acquisizione, la qualità attesa, il rapporto costi/benefici.

L'analisi della rilevanza del dato amministrativo, ossia quanto esso possa essere utilizzato a fini statistici, va fatta tenendo conto non solo della sua utilità attuale, ma anche di quella potenziale. Pertanto, risulterà utile al riguardo sia l'acquisizione delle informazioni che riguardano i contenuti dell'archivio (normativa, collettivi di riferimento, concetti utilizzati, variabili e loro definizioni e classificazioni), sia la ricognizione dei possibili usi, presso i potenziali utilizzatori, da parte delle strutture preposte.

Tali usi non riguardano, come è noto, solo la produzione diretta di informazione statistica e la creazione di registri statistici, ma anche la possibilità di migliorare la qualità di alcuni processi produttivi condotti all'interno di un Istituto Nazionale di Statistica. Mentre potrebbe non essere possibile sostituire completamente una rilevazione diretta con i dati di fonte amministrativa, potrebbe essere comunque conveniente acquisire l'archivio per un uso indiretto, per esempio, per la creazione/integrazione di liste (o *frame*).

A tale scopo è bene tarare questa valutazione preliminare, basandola e differenziandola per tipologia di utilizzo del dato amministrativo, soprattutto quando se ne ha già conoscenza. In sintesi, gli usi generalmente possibili in un Istituto Nazionale di Statistica sono: *i)* creazione e manutenzione di registri; *ii)* supporto ai disegni di campionamento; *iii)* sostituzione o integrazione nella fase di raccolta dati; *iv)* supporto alle procedure di controllo e correzione; *v)* supporto al processo di stima; *vi)* produzione diretta di statistiche; *vii)* supporto alla validazione dei dati (Statistics Canada, 2009).

La valutazione sull'acquisizione di un archivio amministrativo può anche derivare esclusivamente dall'opportunità di ridurre il carico statistico sui rispondenti.

I costi totali di acquisizione e di uso nei processi produttivi dell'Istituto (finanziari e in termini di risorse strumentali e umane impiegate) possono essere difficili da prevedere. Tra questi vanno anche inclusi eventuali costi relativi alle infrastrutture tecnologiche, come per esempio lo sviluppo di piattaforme di scambio. É comunque importante avere una idea dell'impatto che l'acquisizione dell'archivio ha per l'Istituto rispetto ai costi sopracitati. I costi calcolati devono essere rapportati ai benefici ottenibili non solo in termini economici (riduzione costi di rilevazione diretta), ma anche di miglioramento della qualità (maggiore copertura, completezza delle liste, trattamento mancate risposte totali e parziali, nuove statistiche, stime più accurate, validazione, etc.).

Un ulteriore elemento che può influire sulla decisione di acquisizione dell'archivio attiene alla stabilità nel tempo dei dati amministrativi in esso contenuti, rispetto al contesto normativo e alle procedure che ne regolano la sua produzione. Frequenti e rilevanti cambiamenti nella struttura, contenuto, e formato dei dati dell'archivio potrebbero alterare il rapporto costi/benefici, con ricadute significative sulla produzione statistica, sulla qualità e sulla confrontabilità temporale dei dati. Tale aspetto assume una rilevanza determinante per le statistiche e i registri prodotti con regolarità e continuità.

Infine, la decisione sull'acquisizione dell'archivio deve anche basarsi su una valutazione della qualità attesa, ossia rispetto a dei livelli di qualità prestabiliti. É opportuno ricordare che, in questa fase, si fa riferimento alla qualità dell'archivio amministrativo da acquisire, denominata in letteratura come qualità dell'input, e non dell'archivio statistico derivato dopo il processo di trattamento dei dati in esso contenuti, né alla qualità delle stime desunte dalle successive elaborazioni. Inoltre, tale valutazione prescinde ancora dall'obiettivo statistico, che può non essere completamente identificato a questo stadio.

Pertanto, in questa fase, la valutazione della qualità attesa di un archivio da acquisire dovrà basarsi su aspettative comprovate in termini di:

- disponibilità da parte dell'ente a sottoscrivere un accordo formalizzato che stabilisca le modalità e i tempi di eventuale trasmissione dei dati, la sicurezza dei dati, la documentazione di supporto alla trasmissione dell'archivio;
- disponibilità di una completa documentazione dei metadati (definizioni per i collettivi e le principali variabili);
- conoscenza delle condizioni tecniche per l'acquisizione (accessibilità, leggibilità, conformità, convertibilità dei dati);
- conoscenza di eventuali carenze dell'archivio in termini di copertura/completezza in relazione ai principali collettivi e alle principali variabili di interesse per la produzione statistica.

È opportuno in questa fase acquisire un sottoinsieme di dati di prova dell'archivio amministrativo, per verificarne sperimentalmente la qualità.

Alcuni riferimenti bibliografici

Brackstone G.J.(1987). *Issues in the use of administrative records for statistical purposes*. Survey Methodology, June 1987

Calzaroni M. (2011). *Le fonti amministrative nei processi e nei prodotti della statistica ufficiale*, Istat http://www.istat.it/it/files/2011/02/Calzaroni.pdf (ultimo accesso: Dicembre 2013)

Daas P., Ossen S. (2011). *Report on methods preferred for the quality indicators of administrative data sources*, Blue – ETS Project, Deliverable 4.2.

Daas P., Ossen S., Vis-Visschers R., and Arends-Tóth J. (2009). *Checklist for the Quality evaluation of Administrative Data Sources*, Statistics Netherlands, The Hague /Heerlen, 2009

D'Angiolini G., De Salvo P., Passacantilli A. (2014). Istat's new strategy and tools for enhancing statistical utilization of the available administrative databases. European Conference on Quality in Official Statistics (Q2014). Vienna, Austria, 2-5 June 2014Statistics Canada (2009). *Statistics Canada Quality Guidelines*, Fifth Edition – October 2009

Tronti L. (2011) *I dati amministrativi per le statistiche sui mercati del lavoro locali: il progetto Guida*. Dipartimento Funzione Pubblica http://www.istat.it/it/files/2011/02/Tronti.pdf (ultimo accesso: Dicembre 2013)

Wallgren A. and Wallgren B. (2014). Register-based Statistics: Administrative Data for Statistical Purposes. Second Edition. John Wiley & Sons, Chichester, UK.

Principio 3. Acquisizione di un archivio amministrativo

L'acquisizione di un archivio amministrativo è regolamentata da accordi che fissano le condizioni relative alla trasmissione, alla documentazione e alla qualità. Deve essere assicurato un attento monitoraggio delle modifiche legislative o procedurali che hanno impatto sulla struttura o sui dati dell'archivio. Tali modifiche devono essere tempestivamente comunicate agli utilizzatori interni.

Linee guida

L'acquisizione dell'archivio amministrativo deve avvenire attraverso l'istituzione di accordi formalizzati con l'ente titolare (convenzioni, protocolli d'intesa, etc.). Tali accordi dovrebbero stabilire: le modalità e i tempi di trasmissione dei dati, la documentazione di supporto alla trasmissione e quella relativa ai contenuti dell'archivio, le regole per il rispetto della riservatezza e anche le modalità di ritorno dell'informazione statistica all'ente fornitore.

Nel dettaglio, come suggerito dall'Unece (2011), è opportuno che questi accordi:

- contengano un riferimento alla legislazione, se esistente, che consente l'accesso ai dati da parte dell'Istituto statistico;
- identifichino precisamente le persone e le strutture per il trasferimento e la ricezione dell'archivio;
- identifichino tutti i metadati e le informazioni sulla qualità che devono rappresentare la documentazione di base per il corretto uso dell'archivio. In particolare, informazioni importanti sono: le definizioni degli oggetti dell'archivio (unità, eventi) e delle variabili e classificazioni adottate, i riferimenti temporali dei dati dell'archivio, le descrizioni di eventuali trattamenti che i dati hanno subito prima di essere trasmessi all'Istituto;
- richiedano la descrizione dettagliata delle principali popolazioni e variabili derivabili dall'archivio e della qualità in termini di copertura delle popolazioni e la completezza delle variabili;
- fissino la tempistica (data prima fornitura, periodicità forniture successive) per la trasmissione;
- stabiliscano le regole e le procedure attinenti alla tutela della riservatezza che garantiscono le modalità di trasmissione e trattamento di dati sensibili e la prevenzione del rischio di violazione della riservatezza;
- fissino la validità temporale dell'accordo;
- stabiliscano gli eventuali costi della fornitura;
- determinino le condizioni per l'eventuale fornitura periodica dell'archivio;
- impegnino l'ente fornitore alla comunicazione tempestiva di eventuali variazioni nella struttura e/o nei contenuti dei dati per effetto di modifiche legislative o per altri motivi;
- impegnino l'Istituto ad un ritorno di informazione sotto forma di dati statistici come forma di interscambio, nell'ambito della collaborazione costruita nel tempo, tra la struttura di riferimento dell'amministrazione e quella dell'Istituto;
- prevedano un allegato tecnico, affinché la trasmissione dell'archivio amministrativo avvenga in modo sicuro e attraverso protocolli conformi agli standard dell'Istituto.

L'identificazione di una persona di riferimento presso l'ente titolare e il ritorno di informazioni all'ente fornitore sono elementi che pongono le basi per un maggiore sfruttamento dei dati amministrativi. Migliorare il clima di collaborazione tra le strutture coinvolte nel processo di acquisizione potrebbe consentire di avere un ruolo attivo nella fase di progettazione della modulistica utilizzata per la raccolta dei dati, anche al fine di limitare le conseguenze delle eventuali modifiche nella normativa di riferimento e, quindi, di continuare a garantire nel tempo la comparabilità delle statistiche prodotte dall'Istituto.

È importante assicurarsi la comunicazione tempestiva da parte degli enti titolari dei dati amministrativi sulle variazioni nella modulistica utilizzata e nelle definizioni relative ai concetti amministrativi (D'Angiolini *et*

al. 2014). Ciò rende possibile comprendere l'impatto di tali variazioni sulla produzione statistica anche con l'ausilio della comunità di esperti di settore.

Alcuni riferimenti bibliografici

Australian Bureau of Statistics (2011). Quality Management of Statistical Outputs Produced From Administrative Data. Information Paper. Australia. March 2011.

D'Angiolini G., Patruno E., Saccoccio T., De Rosa C., Valente E. (2014). DARCAP: A tool for documenting the information content and the quality of the available administrative data sources. European Conference on Quality in Official Statistics (Q2014). Vienna, Austria, 2-5 June 2014

Statistics Canada (2009). *Statistics Canada Quality Guidelines*, Fifth Edition – October 2009

Statistics Canada (2009). *Statistics Canada, Use of administrative data (website)* http://www.statcan.gc.ca/pub/12-539-x/2009001/administrative-administratives-eng.htm (ultimo accesso: Dicembre 2013)

United Nations Economic Commission for Europe - Unece (2011). *Using administrative and secondary source for official statistics: a handbook of principles and practice*. United Nations, New York and Geneva, 2011.

Principio 4. Pre-trattamento, controlli di qualità e messa a disposizione dell'archivio amministrativo

In fase di acquisizione dell'archivio amministrativo ne viene verificata la documentazione e la qualità e si procede all'eventuale integrazione con le informazioni che ne possano aumentare la chiarezza e l'utilizzabilità. L'archivio amministrativo pre-trattato deve essere reso disponibile per gli usi interni, corredato da tutti i metadati e gli indicatori di qualità precedenti e successivi al trattamento.

Linee guida

L'archivio viene acquisito dall'Istituto, secondo le modalità concordate con l'ente titolare, attraverso la struttura incaricata di gestire i rapporti con i fornitori, effettuare controlli di qualità generali, effettuare pre-trattamenti sui dati e predisporre gli archivi per eventuali successive integrazioni.

In generale, il pre-trattamento dei dati ricevuti potrà seguire le seguenti fasi: *i)* analisi concettuale degli oggetti contenuti nell'archivio; *ii)* caricamento dei dati in data-base interni; *iii)* individuazione delle unità e assegnazione di codici identificativi univoci e stabili nel tempo (chiavi univoche); *iv)* standardizzazione di alcuni dati; *v)* ricodifiche di alcune variabili. Infine, vengono calcolati indicatori di qualità e viene predisposta la documentazione sui metadati e sulla qualità (Di Bella G. e Ambroselli S., 2014; Kobus P. e Murawski P; Cetkovic P., *et al*., 2012).

L'archivio acquisito deve essere inserito nel sistema corrente di produzione dell'Istituto; ciò implica che il formato in cui l'archivio è fornito deve essere eventualmente reso compatibile con i formati in uso all'Istituto. In fase di acquisizione devono essere effettuati controlli relativi: alla corrispondenza tra il numero di record importati e quelli attesi (anche in base alle forniture precedenti); alla chiarezza del tracciato record; alla corrispondenza tra tracciato record e dati importati (per evitare possibili disallineamenti delle colonne); alla esistenza di chiavi univoche e comparabili con quelle in uso all'Istituto; alla corrispondenza tra tipologia di variabile (numerica, alfanumerica) e formato dei dati; all'adeguatezza della lunghezza dei campi assegnati per le variabili. Per le misure specifiche, relative a questa fase, si può far riferimento agli indicatori presenti nella dimensione dei *Technical checks* definiti nell'ambito del progetto BLUE-Ets (Daas *et al*., 2011).

Successivamente, è necessario procedere alla rimozione di eventuali duplicazioni di record. Questa operazione può richiedere che le variabili identificative del record siano precedentemente sottoposte a procedure di standardizzazione (attività di "*parsing*", ossia di separazione di una certa variabile in più variabili, come viene fatto per standardizzare i nomi e cognomi di individui o gli indirizzi).

Quindi deve essere condotta un'analisi delle classificazioni utilizzate, per comprendere il loro grado di aderenza alle classificazioni standard o in uso all'Istituto. In caso di non corrispondenza tra le classificazioni è necessario integrare i dati con le classificazioni pertinenti.

É bene sottolineare che, in questa fase i controlli applicati sono mirati a dare elementi per decidere se l'archivio fornito può essere considerato valido e rilasciabile agli utilizzatori interni in forma completa o parziale o con delle segnalazioni di cautela rispetto ad alcune variabili. In genere, non vengono applicate in questa fase le procedure di controllo e correzione tipiche del trattamento dei dati a fini statistici.

In caso di evidenze sulla non conformità della fornitura è auspicabile una verifica attraverso il ricontatto dell'ente titolare dell'archivio amministrativo.

É opportuno corredare l'archivio di un report di qualità che includa la descrizione dei metadati e del processo di trasformazione e integrazione subìto, e alcuni indicatori di qualità generali (copertura per le principali popolazioni di riferimento, tassi sui dati mancanti per le principali variabili, etc.).

L'informazione relativa alla versione dell'archivio disponibile, la documentazione sulle procedure di pre-trattamento e i controlli di qualità effettuati devono essere resi disponibili in modo da supportare l'utilizzo interno dell'archivio.

É opportuno che, una volta identificate le procedure per la validazione dei dati amministrativi in fase di ricezione, queste siano applicate con regolarità sulle forniture successive.

I metadati descrittivi dell'archivio amministrativo acquisito devono essere coerenti con quelli definiti dal Sistema Unitario dei Metadati (SUM). Per la reportistica della qualità si deve fare riferimento alla Quality Report Card, che contiene indicatori ampiamente testati e validati in ambito internazionale (Cerroni *et al*., 2014).

Alcuni riferimenti bibliografici

Cetkovic P., Humer S., Lenk M., Moser M., Schnetzer M., Schwerer E. (2012). *A quality monitoring system for statistics based on administrative data.* European conference on Quality in Official Statistics Q2012, 29 May – 1 June 2012, Athens, Greece.

Cerroni F., Di Bella G., Galiè L. (2014). *Evaluating administrative data quality as input of the statistical production process*. Rivista di Statistica Ufficiale, n. 1-2, 2014

Daas P., Ossen S. (2011). *Report on methods preferred for the quality indicators of administrative data sources*, Blue – ETS Project, Deliverable 4.2.

Di Bella G., Ambroselli S. (2014). Towards a more efficient system of administrative data management and quality evaluation to support statistics production in Istat. European Conference on Quality in Official Statistics. Q2014. Vienna, Austria 2-5 June 20014

Kobus P. e Murawski P (non noto). Transforming administrative data to statistical data using ETL tools. http://www.ine.es/e/essnetdi_ws2011/ppts/Murawski_Kobus.pdf

Statistics Canada (2009). *Statistics Canada Quality Guidelines*, Fifth Edition – October 2009

Principio 5. Monitoraggio e valutazione dell'uso dell'archivio e feedback all'ente fornitore
L'uso dell'archivio amministrativo acquisito deve essere monitorato e valutato attraverso un processo coordinato e condiviso, al fine di individuare sia le carenze in termini di qualità sia i trattamenti comuni a più processi. I risultati del monitoraggio e della valutazione devono essere condivisi con tutti gli utilizzatori attuali e potenziali ed eventualmente trasmessi, con opportune modalità, all'ente titolare dell'archivio.

Linee guida

Può accadere che uno stesso archivio amministrativo sia utilizzato come input in più processi di produzione. In particolare può presentarsi il caso in cui insiemi di variabili o sottopopolazioni diverse vengano utilizzate in più processi di produzione, oppure che le stesse variabili costituiscano l'input per la produzione di più statistiche (ad esempio statistiche di tipo congiunturale e statistiche di tipo strutturale). Anche il tipo di uso può variare laddove i dati amministrativi vengono utilizzati in modo diretto oppure a supporto del processo di indagine.

Pertanto, è necessario effettuare delle ricognizioni periodiche sui tipi di uso dei dati amministrativi nell'Istituto ovvero, per ciascun processo di produzione, conoscere quali dati amministrativi vengono utilizzati e come, preferibilmente attraverso l'interscambio di informazioni tra i vari sistemi informativi di gestione dei processi, in un'ottica di inter-operabilità che ottimizzi l'efficienza complessiva. In tal modo è possibile ottenere un quadro completo della rilevanza di ciascuna fonte sulla base della quale definire dei livelli di attenzione particolari necessari a gestire la "dipendenza" della produzione dell'istituto dai dati amministrativi.

Dal punto di vista gestionale è anche opportuno avvalersi delle valutazioni da parte delle comunità di esperti e utenti di fonti amministrative, in particolar modo per quelle fonti più utilizzate dall'Istituto, come ad esempio i dati di fonte previdenziale e di fonte fiscale. Il coordinamento deve essere gestito centralmente al fine di condividere le problematicità, favorire le sinergie ed evitare duplicazioni di attività in un'ottica di efficienza e di standardizzazione.

Infine, la condivisione di possibili problemi di qualità dei dati che possono limitarne l'utilizzo statistico, permette di concordare eventuali azioni con l'ente fornitore volte al loro superamento. Dal caso più semplice di incremento della tempestività di fornitura dei dati, al caso di miglioramento della chiarezza dei metadati o all'eventuale possibile riduzione della distanza tra concetti amministrativi e concetti statistici. Questo processo di feedback è particolarmente importante al fine di potenziare l'uso dei dati amministrativi nel tempo.

D'altra parte il coordinamento degli utenti interni della fonte rende possibile definire esattamente l'insieme dei dati richiesti in relazione alla periodicità e alla tempestività delle varie forniture evitando di appesantire inutilmente l'attività del titolare della fonte per l'approvvigionamento dei dati.

Alcuni riferimenti bibliografici

Di Bella G., Ambroselli S. (2014). Towards a more efficient system of administrative data management and quality evaluation to support statistics production in Istat. European Conference on Quality in Official Statistics. Q2014. Vienna, Austria 2-5 June 20014.