Linee Guida
per la Qualità delle Statistiche
del Sistema Statistico Nazionale

ver. 1.0

Marzo 2018

Istituto Nazionale di Statistica – Istat

La stesura di questo manuale è stata coordinata da G. Brancato

Autori: G. Brancato, A. Boggia, G. Ascari

Si ringraziano per la collaborazione: S. Terracina (Parte II, sezione M), M. Fortini (Parte I, Capitoli 2 e 3), A. Sabbatini (Parte I, Capitolo 3), L. Tosco (Parte II, sezione M), P. De Salvo (Parte I, Capitolo 2, Parte II Sezione G), F. Barbalace (Parte II, sezione A), A. Villa (Parte I, Capitolo 1), P. Anzini (Parte II, sezioni J e K)

Si ringraziano i revisori interni all'Istat e gli Enti del Sistema Statistico Nazionale che hanno fornito suggerimenti per il miglioramento delle Linee Guida

Sommario

Introduzione

Negli ultimi anni l'Istat si è impegnato fortemente per migliorare l'efficienza del Sistema statistico nazionale (Sistan) e la qualità delle informazioni statistiche che esso produce e diffonde. Già con la direttiva sull'Adozione del Codice italiano delle statistiche ufficiali[1], che eredita lo schema concettuale e i principi del Codice delle statistiche europee, è stato intrapreso un percorso per allineare la produzione statistica nazionale agli standard qualitativi europei. Tale percorso ha visto un'intensa attività di monitoraggio da parte dell'Istat sull'ottemperanza al Codice nel periodo 2012-2015 sia attraverso un modulo ad hoc inserito nella "Rilevazione sugli elementi identificativi, risorse e attività degli uffici di statistica del Sistan" (EUP), sia attraverso delle *peer review* condotte presso un elevato numero di enti appartenenti al Sistan. Oltre alle iniziative citate, un forte stimolo a proseguire in questa direzione è arrivato nel 2015, anno in cui l'Istat e un sottoinsieme delle altre Autorità Nazionali (*Other National Authorities*, d'ora in poi ONA) che producono statistiche europee[2] sono stati sottoposti a *peer review* sull'attuazione del Codice delle statistiche europee. Ne sono scaturite un insieme di raccomandazioni e azioni di miglioramento per la ridefinizione del Sistema statistico nazionale e il rafforzamento del suo coordinamento. Tra queste, vi è l'adozione di un approccio di tipo audit sui alcuni processi produttivi delle ONA[3].

L'**audit statistico** sui processi è una procedura di valutazione da parte di un team di auditori che ha l'obiettivo di accompagnare i responsabili di tali processi in un percorso di identificazione delle criticità e delle attività più adeguate per il loro superamento. Le **Linee Guida** rappresentano lo standard di riferimento e, insieme al **questionario di valutazione** che ne rispecchia da vicino i contenuti, permettono la conduzione degli stessi audit statistici.

L'Istat ha una consolidata esperienza nella progettazione e conduzione di audit per i processi produttivi statistici, avendo adottato questo approccio, insieme all'autovalutazione, sui processi interni, già a partire dal 2010. L'autovalutazione consiste in una valutazione da parte del responsabile del processo, che sulla base delle stesse linee guida standard di riferimento, compila il questionario di valutazione e identifica le azioni per migliorare la qualità. Gli strumenti sviluppati, che all'inizio riguardavano solo le rilevazioni dirette, sono stati poi estesi ai processi che utilizzano dati di fonte amministrativa. Tutta la documentazione è disponibile sul sito dell'Istat.

Il principale **obiettivo** di queste Linee Guida è quindi fornire uno strumento di supporto alla valutazione della qualità delle statistiche prodotte dagli enti del Sistan. Il manuale riporta i **principi** da seguire per produrre le statistiche secondo gli standard metodologici più consolidati e in modo da assicurare che l'informazione prodotta sia di qualità, e fornisce **suggerimenti pratici** per garantire l'ottemperanza ai principi enunciati. Le Linee Guida, insieme ad un questionario di valutazione, saranno utilizzate per condurre audit da parte di team composti da personale dell'Istat e degli Enti. Il manuale può essere altresì un utile riferimento per la progettazione e realizzazione di un processo statistico, fornendo una guida per il corretto sviluppo delle fasi e attività del processo stesso.

[1] Direttiva n. 10 del 17 marzo 2010 pubblicata sulla Gazzetta Ufficiale della Repubblica Italiana n.240 del 13 ottobre 2010

[2] Statistiche europee sono quelle incluse nel Programma statistico europeo prodotte dall'Eurostat, dagli Istituti Nazionali di statistica e dalle altre Autorità Nazionali (ONA) e che seguono i principi statistici previsti nel Codice delle statistiche europee e nella legge statistica europea, Regolamento (CE) N. 223/2009 del Parlamento europeo e del Consiglio relativo alle statistiche europee modificato dal Regolamento (UE) 2015/759 del Parlamento europeo e del Consiglio del 29 aprile 2015. L'elenco delle ONA è pubblicato sul sito di Eurostat ai sensi dell'art. 5 del Regolamento (CE) 223/2009.

[3] Informazioni sulle peer review condotte da Eurostat e tutta la relativa documentazione sono disponibili al seguente link: http://ec.europa.eu/eurostat/web/quality/peer-reviews.

Queste Linee Guida sono rivolte ai responsabili dei processi statistici degli enti del Sistan, che in tal modo possono ripercorrere tutte le fasi di un tipico processo produttivo diretto o che utilizza dati di fonte amministrativa, per comprendere: quali siano le attività statistiche che possono essere svolte, come dovrebbero essere implementate per prevenire gli errori, quali indicatori possano essere calcolati per monitorare in corso d'opera e valutare a posteriori la qualità, quale sia l'impatto delle procedure sulla qualità finale dei dati.

Queste Linee Guida sono state sviluppate per la valutazione tramite audit delle statistiche europee delle ONA, così come richiesto dalla *peer review* europea. Tuttavia, esse insieme al questionario di valutazione hanno carattere generale, e potrebbero essere utilizzate anche da altri enti del Sistan e/o in un approccio di auto-valutazione.

Il manuale è così strutturato. Nella prima parte (Parte I) si descrive il quadro di riferimento adottato per la valutazione della qualità e si introducono i concetti relativi alla qualità delle statistiche, agli standard utilizzati per descrivere il processo produttivo e agli errori che si generano durante le fasi del processo stesso. La seconda parte (Parte II) segue le fasi[4] rilevanti in cui si può articolare un processo produttivo statistico, sia esso diretto o che utilizza dati di fonte amministrativa. Viene descritto l'obiettivo di ciascuna fase e le possibili scelte metodologiche, tra quelle maggiormente consolidate, per conseguirlo, sono enunciati uno o più principi di qualità ed elencati i suggerimenti per la loro ottemperanza. Inoltre, sono suggeriti alcuni indicatori di qualità e performance utili nel monitoraggio e nella valutazione della qualità, è riportata la mappatura con i sotto-processi di GSBPM e infine viene fornita una bibliografia tematica per ulteriori approfondimenti. Si è ritenuto utile anche fornire un Glossario per alcuni termini utilizzati nel manuale. Infine il manuale comprende delle appendici di approfondimento su tematiche specifiche.

Questa versione recepisce i commenti ricevuti nell'ambito di una consultazione sulle linee guida effettuata presso i principali enti del sistema statistico nazionale.

[4] Qui si parla genericamente di fase, ma possono essere anche sotto-processi di una fase o aggregazioni di sotto-processi.

Parte I

1. Il contesto Europeo per la qualità

Negli ultimi 20 anni, Eurostat, l'Ufficio di statistica dell'Unione Europea, in collaborazione con gli Istituti nazionali di statistica degli stati membri, ha tracciato un lungo e articolato percorso per la qualità, che ha rappresentato un orientamento e uno stimolo per le attività di sviluppo della qualità e per il suo miglioramento.

Da tempo la comunità degli statistici europei ha condiviso le definizioni della **qualità delle statistiche**, convergendo sull'idea che oltre all'accuratezza delle stime prodotte, che è una caratteristica prettamente statistica, siano importanti altri aspetti quali: la pertinenza, la tempestività e puntualità, l'accessibilità e chiarezza, la confrontabilità e la coerenza (Eurostat, 2003).

Già agli inizi degli anni 2000 (*Leg on Quality*, 2001), era emersa la consapevolezza che, l'introduzione di un approccio di **gestione della qualità** (*quality management)*, personalizzato per la statistica ufficiale, avrebbe portato benefici in termini di qualità delle statistiche prodotte. Si è andato quindi costruendo un sistema di gestione della qualità che ha tra i suoi pilastri il **Codice delle statistiche europee,** approvato nel 2005 e rivisitato nel 2011 (Eurostat, 2011). Sviluppato in tre aree (istituzionale, dei processi e dei prodotti), il Codice è uno strumento di auto-regolamentazione che ha l'obiettivo di aumentare la fiducia nella statistica ufficiale, contribuendo al rafforzamento dell'indipendenza, integrità e responsabilità delle Autorità statistiche[5] e migliorando la qualità delle statistiche europee. Esso è affiancato dal ***ESS Quality Assurance Framework***, o ESS QAF, sviluppato nel 2011 e successivamente aggiornato (Eurostat, 2015), che identifica metodi e strumenti a livello istituzionale o di processo utili per rendere operativa l'ottemperanza ai principi del Codice, contribuendo alla promozione di buone pratiche. Il Sistema statistico europeo si è anche dotato di un meccanismo concreto per il monitoraggio della qualità delle statistiche prodotte dagli Istituti nazionali di statistica, che prevede in una prima fase la compilazione di un questionario di autovalutazione e in seguito la conduzione di *peer review*. Gli strumenti adottati hanno l'obiettivo di verificare l'aderenza al Codice, la definizione di eventuali azioni di miglioramento in risposta all'esito della *peer review*, il monitoraggio sull'implementazione delle azioni di miglioramento definite.

Se in generale tutto l'impianto costruito intorno al Codice europeo definisce una cornice ampia, è in particolare nel Principio 4 dello stesso codice che si enuncia l'impegno delle Autorità statistiche per il miglioramento continuo della qualità dei prodotti e dei processi, attraverso la costruzione di una infrastruttura e di una strategia per la qualità, lo sviluppo di procedure per il monitoraggio della qualità, la misurazione e comunicazione della qualità secondo le dimensioni definite da Eurostat e, infine, l'adozione di meccanismi di valutazione dei processi produttivi statistici, quali per esempio l'audit e l'auto-valutazione.

In ambito Eurostat, per **audit statistico** si intende una procedura indipendente per ottenere evidenze verificabili e oggettive sull'aderenza a standard stabiliti. Può essere realizzata attraverso auditori esterni o interni all'organizzazione (ma non coinvolti nel processo auditato), può richiedere la compilazione di report o questionari di valutazione o il calcolo di indicatori di qualità, si conclude con l'identificazione di punti i forza e punti di debolezza del processo e la definizione di azioni di miglioramento e della tempistica per la loro realizzazione (Eurostat, 2007).

[5] Le Autorità statistiche comprendono la Commissione (Eurostat), gli Istituti nazionali di statistica e le altre autorità nazionali responsabili dello sviluppo, produzione e diffusione delle statistiche europee.

2. I processi produttivi statistici

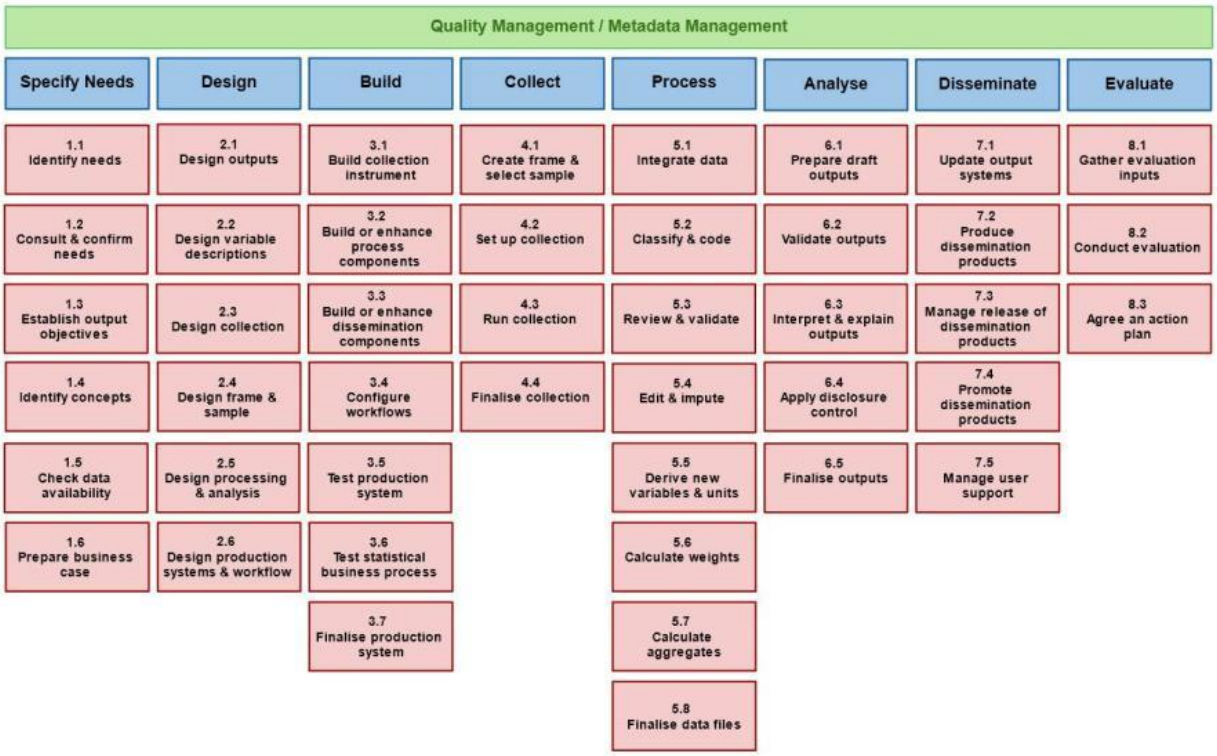
In questo manuale, per la descrizione dei processi produttivi da cui hanno origine le statistiche, viene adottato il modello **GSBPM** (*Generic Statistical Business Process Model*), sviluppato dall'Unece (2013). L'introduzione di tale schema risponde a un'esigenza di classificazione e armonizzazione delle diverse fasi dei processi messi in atto dagli Istituti nazionali di statistica; esso rappresenta inoltre un modello su cui è possibile basare la valutazione e il miglioramento della qualità di tali processi.

In particolare, si tratta di uno schema che può essere applicato a qualunque processo produttivo, dall'indagine tradizionale, all'acquisizione di dati amministrativi, alle elaborazioni statistiche, a prescindere dal settore tematico di riferimento, purché vi sia come risultato un output in termini di dati e metadati statistici.

Tale universalità è diretta conseguenza della flessibilità del modello, che non è costituito da una sequenza lineare di azioni bensì da una matrice di fasi e sotto-processi, di diversa ampiezza e importanza all'interno dei processi reali. Ciò permette di adattare la struttura del modello a processi di diversa dimensione e natura. Infatti, alcune fasi potrebbero essere applicate ad un processo e non essere applicate ad un altro; i sotto-processi non devono necessariamente essere seguiti secondo un ordine predeterminato o gerarchico, ossia alcuni si possono saltare, altri ripetersi più volte, dando vita a cicli iterativi.

Nella sua versione più recente il GSBPM è formato da otto fasi, ciascuna con un diverso numero di sotto-processi al proprio interno (Figura 1). Le fasi coprono i principali passaggi dello sviluppo di un processo statistico: dall'identificazione delle esigenze informative, alla diffusione e alla valutazione dei risultati, passando per la progettazione, la raccolta, il trattamento dei dati e vari altri *step* intermedi. Inoltre, sono definiti dei processi sovrastanti (*overarching*), tra cui la gestione dei dati e metadati, che include l'archiviazione, e la gestione della qualità.

Figura 1. Generic Statistical Business Process Model, ver. 5.0



La generalità del GSBPM lo rende particolarmente idoneo a rappresentare i processi di qualsiasi ente produttore di statistiche. Tuttavia, il modello rappresenta i diversi sotto-processi "alla pari", mentre ai fini della valutazione della qualità può essere più rilevante soffermarsi su alcuni piuttosto che su altri. Alla luce di ciò nel manuale (e nel relativo questionario) alcuni sotto-processi del GSBPM sono stati accorpati. Si è anche cercato di approfondire maggiormente gli aspetti più rilevanti per la produzione del Sistan. Tenendo in considerazione le finalità indicate, nella Tabella 1 che segue si riportano le fasi e i sotto-processi del GSBPM cosi come sono stati accorpati nelle sezioni della Parte II del manuale

Tabella 1. Fasi e/o sotto-processi del GSBPM e Sezioni della Parte II delle Linee Guida

Fasi e sotto-processi GSBPM	Sezioni del manuale
Specify needs (1.1. – 1.2. – 1.3. – 1.4. – 1.5.), Design outputs (2.1.), Design variable description (2.2.), Gather evaluation inputs (8.1.)	A. Identificazione delle esigenze degli utenti, definizione dei concetti scelta delle fonti e valutazione della soddisfazione
Design frame & sample (2.4.), Build or enhance process components (3.2.), Create frame & select sampe (4.1.), Calculate weights (5.6.), Calculate aggregates (5.7.), Gather evaluation inputs (8.1.)	B. Scelta del disegno, lista di riferimento, campionamento e stima
Design collection (2.3.), Build collection instrument (3.1.), Build or enhance process components (3.2.),Test production system (3.5.), Set up collection (4.2.), Run collection (4.3.), Gather evaluation inputs (8.1.)	C. Acquisizione dei dati
Design collection (2.3.), Build collection instrument (3.1.), Build or enhance process components (3.2.), Test production system (3.5.), Finalise collection (4.4.), Gather evaluation inputs (8.1.)	D. Conversione in formato elettronico (registrazione)
Design processing and analysis (2.5.), Test production system (3.5.), Integrate data (5.1.), Gather evaluation inputs (8.1.)	E. Integrazione
Design processing and analysis (2.5.), Test production system (3.5.), Classify & code (5.2.), Gather evaluation inputs (8.1.)	F. Codifica e classificazioni
Design processing and analysis (2.5.), Test production system (3.5.), Review & validate (5.3.), Edit & impute (5.4.), Gather evaluation inputs (8.1.)	G. Identificazione e trattamento degli errori
Design processing and analysis (2.5.), Test production system (3.5.), Derive new variables and units (5.5.), Gather evaluation inputs (8.1.)	H. Derivazione delle unità
Design processing and analysis (2.5.), Test production system (3.5.), Derive new variables and units (5.5.), Gather evaluation inputs (8.1.)	I. Derivazione delle variabili
6.1. Prepare draft outputs	J. Destagionalizzazione
From Run collection (4.3.) to Finalise outputs (6.5.)	K. Politica delle revisioni
Design processing and analysis (2.5.), Test production system (3.5.), Validate outputs (6.2.), Gather evaluation inputs (8.1.)	L. Validazione dei risultati
Design processing and analysis (2.5.), Build or enhance dissemination components (3.3.), Apply disclosure control (6.4.), Disseminate (7.1. – 7.2. – 7.3. – 7.4. –7.5.), Gather evaluation inputs (8.1.)	M. Diffusione dei dati e tutela della riservatezza, archiviazione, documentazione

Come già accennato, in questo manuale il GSBPM viene adottato per rappresentare sia i processi diretti sia quelli che utilizzano **dati di fonte amministrativa**. Nel primo caso gli strumenti per l'acquisizione dell'informazione sono progettati allo scopo primario di produrre statistiche sul fenomeno di interesse, mentre nel secondo caso la finalità originaria dei dati è di tipo gestionale e solo in via subordinata i dati sono usati per produrre informazioni statistiche su un collettivo di interesse.

È utile sapere che intorno al GSBPM sono sorte altre iniziative. Recentemente, la *Modernising Committe on Standards* dell'Unece (*United Nations Economic Commission for Europe*) ha sviluppato indicatori di qualità e mappato gli indicatori esistenti per ciascuno dei sotto-processi del GSBPM (Unece, 2016).

3. Il modello di riferimento per la qualità

Il modello di riferimento che si propone nel manuale integra le seguenti prospettive sulla qualità:

- la qualità delle statistiche prodotte, definita attraverso le dimensioni della qualità Eurostat;
- le fonti di errore che si generano durante il processo produttivo statistico e che hanno impatto sulla qualità finale dei risultati, che possono essere:
    - monitorate attraverso il calcolo di indicatori standard di tipo "indiretto", utili segnali per la tempestiva identificazione di eventuali problematiche,
    - valutate mediante le misurazioni statistiche della qualità, come la variabilità e la distorsione delle stime;
- la qualità del processo la cui realizzazione, attraverso l'identificazione di principi e suggerimenti, consente di contenere la variabilità non voluta e migliorare l'efficienza e quindi, di riflesso, si traduce in una maggiore qualità dei risultati.

Come già accennato, la qualità delle statistiche è oggi considerata non solo in relazione all'accuratezza[6], ossia la vicinanza tra la stima e il vero valore ignoto del parametro della popolazione, ma anche rispetto ad altre caratteristiche, che insieme a questa costituiscono le **dimensioni Eurostat della qualità** (Eurostat, 2003). Pe le definizioni delle dimensioni della qualità si faccia riferimento all'Appendice A. Oggi si guarda alla qualità anche in termini di capacità di produrre informazione statistica che: *i)* soddisfi i bisogni conoscitivi degli utenti (pertinenza), *ii)* sia diffusa in tempo utile e secondo un calendario prestabilito (tempestività e puntualità) e in modo accessibile con tutte le informazioni che ne permettano il suo corretto utilizzo (accessibilità e chiarezza), *iii)* sia coerente e confrontabile nel tempo e nello spazio. Queste dimensioni possono essere in conflitto tra di loro. Per esempio produrre stime molto accurate potrebbe richiedere tempi molto lunghi e andare a discapito della puntualità e tempestività. Peraltro, solo alcune di queste dimensioni possono essere misurate con degli indicatori, mentre per le altre si possono formulare solo dei giudizi qualitativi. Anche per le componenti misurabili in termini quantitativi, può essere poi difficile o molto costoso produrre delle misure. Infine, i costi non sono una dimensione della qualità ma sicuramente ne possono rappresentare un vincolo.

La qualità delle statistiche prodotte dipende dai processi produttivi sottostanti. Ciascuna delle azioni che compongono le fasi del processo è soggetta ad errori che derivano dalle caratteristiche della fase stessa (per esempio, rilevazione diretta, acquisizione di dati di fonte amministrativa, registrazione su supporto magnetico, classificazione, imputazione, ecc.). Gli errori di diversa natura che si creano durante il processo influenzano l'accuratezza delle stime prodotte in termini di **distorsione** e **variabilità** rispetto al parametro di interesse.

In letteratura, gli errori si dividono in prima istanza tra:

- sistematici, che tendono a seguire una legge deterministica e ad influenzare le stime inducendo una distorsione di segno sempre positivo o sempre negativo;
- accidentali, che sono frutto di cause episodiche di varia natura, che tendono a disporsi in modo simmetrico intorno al parametro di interesse.

Le cause di questo secondo tipo non provocano distorsioni dello stimatore, ma ne influenzano la variabilità.

[6] All'accuratezza è stato anche affiancato il concetto di affidabilità che riguarda le statistiche sottoposte a politica di revisione, per le quali sono prodotte più stime. L'affidabilità viene misurata in relazione alla vicinanza tra le stime prodotte per la stessa statistica in tempi diversi.

Per la quantificazione di tale errore, in letteratura si usa fare riferimento all'Errore Quadratico Medio (MSE, da *Mean Squared Error*), una misura che riflette le due componenti: distorsione e variabilità. Se il metodo di campionamento e la procedura di stima portano a uno stimatore non distorto, allora l'errore quadratico medio è semplicemente la varianza dello stimatore (Biemer e Lyber, 2003).

È appena il caso di osservare che le caratteristiche di distorsione e variabilità si applicano alla distribuzione di probabilità indotta dagli errori sull'universo di tutte le possibili stime che si possono, ipoteticamente, produrre attraverso l'indagine. Nella pratica l'indagine produce una sola realizzazione 'estraendola' dall'insieme delle stime possibili secondo il meccanismo di casualità indotto dalla legge di estrazione del campione e da quelle di generazione dei diversi errori. Quindi, anche il concetto di variabilità deve essere inteso come il rischio che l'unica stima da noi posseduta possa risultare anche molto distante dal vero valore del parametro sul quale si vuole fare inferenza.

Oltre alla suddivisione degli errori in sistematici e accidentali, nella valutazione della qualità è particolarmente utile rifarsi ad un'altra classificazione in funzione delle fasi o dei sotto-processi in cui si generano, come riportato nella Tabella 2, che fa riferimento alle fasi/sotto-processi considerati nella Parte II di queste Linee Guida. Qui di seguito si descrivono le tipologie di errore.

Errori di specificazione
Producono effetti sulla pertinenza del dato e nascono nella fase di progettazione quando le definizioni e i concetti operativi adottati per le popolazioni e le variabili di interesse non coincidono con quelle teoriche. Solitamente sono più importanti per le indagini da fonte amministrativa, per le quali non è possibile pianificare il sistema di rilevazione, e si riverberano sull'accuratezza dei dati soprattutto in termini di distorsione delle stime prodotte. Da alcuni autori, questa tipologia di errore viene denominata "validità del costrutto" (Groves et al, 2004).

Errori di copertura
Sono causati dalle imperfezioni presenti nelle liste usate per l'estrazione del campione e il contatto delle unità appartenenti alla popolazione obiettivo. Un primo effetto è quello di alterare le probabilità teoriche di inclusione nel campione. Essi si distinguono in errori di sotto-copertura, quando alcune delle unità appartenenti alla popolazione obiettivo rimangono escluse per qualche motivo dalla lista e di sovra-copertura, quando viceversa alcune unità non appartenenti alla popolazione obiettivo sono erroneamente incluse nella lista o sono presenti in forma duplicata. Nei dati di fonte amministrativa, usualmente non vi è un effetto sul disegno campionario, in quanto i dati sono utilizzati in modo esaustivo. Tuttavia, la popolazione contenuta negli archivi amministrativi utilizzati potrebbe non coincidere con quella obiettivo, portando a sovracopertura (facilmente rimuovibile eliminando le unità erroneamente incluse) o in forma più grave a sotto-copertura, in genere di tipologie di sottopopolazioni, aspetto che richiede la loro ricerca in altre fonti e l'integrazione di dati. Gli errori di copertura possono anche derivare dalle trasformazioni e aggregazioni applicate in fase di "ricostruzione" o "derivazione" dell'unità di interesse, quando le fonti amministrative non le contengano direttamente. Tutti questi errori possono provocare distorsione nelle stime se esiste un'associazione tra il meccanismo di generazione dell'errore e la caratteristiche di interesse. Ad esempio, se gli stranieri sono più soggetti a sotto-copertura e sono più giovani della media degli italiani, la percentuale di italiani nella popolazione sarà distorta per eccesso e l'età media della popolazione sarà sovrastimata.

Errori di campionamento
Derivano dal fatto che viene osservata soltanto una selezione di unità tra tutte quelle appartenenti alla popolazione obiettivo, il campione, un insieme sul quale normalmente il valore misurato non necessariamente è identico a quello del parametro della popolazione. Se ipoteticamente selezionassimo più volte il campione otterremmo valori ogni volta diversi, anche se prossimi, al valore del parametro di

interesse nell'intera popolazione. La misura della variabilità della stima riferita al parametro di interesse, calcolata rispetto a tutti i possibili campioni che possono essere estratti dalla popolazione prende il nome di errore campionario. In alcuni casi il campionamento può influenzare l'accuratezza della stima anche in termini di distorsione. Ciò accade se il valore medio della stima ottenuta considerando tutti i possibili campioni, risulta essere differente da quello del parametro di popolazione. La teoria del campionamento probabilistico è ormai un solido strumento statistico per tenere sotto controllo la variabilità e la distorsione campionarie mediante l'uso di disegni di campionamento basati sulla scelta casuale delle unità.

Errori di mancata risposta totale e di non osservazione

Si parla di errore di mancata risposta **totale** quando tutte le informazioni sull'unità statistica sono mancanti, o presenti ma in una misura considerata insufficiente. Possono essere causate da vari motivi: errori nelle informazioni per il contatto presenti nella liste di estrazione (errori di lista), rifiuti, impossibilità a partecipare alla rilevazione per assenza temporanea o altre cause, come la malattia. Un esempio di errori di lista sono i problemi negli indirizzi di abitazione che impediscono il contatto delle unità di interesse, compresa la verifica dell'appartenenza alla popolazione obiettivo. A volte l'informazione che si riesce a reperire sulle unità statistiche è così debole che non è possibile stabilire se le unità siano eleggibili, ossia appartengano alla popolazione obiettivo[7]. Nei dati di fonte amministrativa, si parla di "non osservazione", qualora vi sia un segnale di esistenza della unità, in genere l'identificativo della unità, ma l'insieme delle informazioni ad essa relative nell'archivio non siano presenti o siano insufficienti per considerarla acquisita. Questo tipo di errore è piuttosto raro e può essere difficile distinguerlo dall'errore di copertura.

Errori di mancata risposta parziale

Gli errori di mancata risposta **parziale** sono non osservazioni relative solo ad alcune variabili di interesse. Essi possono essere causati da una cattiva formulazione dei quesiti del questionario o da rifiuto a rispondere a quesiti che possono essere percepiti come sensibili. Mancate risposte totali e parziali si ripercuotono sull'accuratezza delle statistiche prodotte, con il rischio di generare una distorsione delle stime, tanto maggiore quanto più il fenomeno che genera l'errore di mancata risposta è associato statisticamente alla dimensione della caratteristica di interesse. Se, ad esempio, i rispondenti tendono a non rispondere sul proprio reddito tanto più quanto questo è elevato, allora il reddito medio del collettivo di interesse risulterà sottostimato.

Errori di misura

Sono costituiti da tutte le differenze tra il valore vero della caratteristica posseduta da un'unità statistica e quello osservato in fase di rilevazione o a lei attribuito in fase di trattamento dei dati. Le cause degli errori di misura in fase di rilevazione dei dati sono attribuibili: alla cattiva formulazione dei quesiti del questionario; all'erroneo atteggiamento dei rilevatori nel porre i quesiti; all'attitudine del rispondente che può volontariamente o meno rispondere il falso; all'erronea scelta della tecnica di rilevazione che può indurre il rispondente a non rispondere conformemente alla realtà. Per esempio, è noto che per i quesiti sensibili sia preferibile evitare una tecnica di intervista faccia-a-faccia, che potrebbe mettere in imbarazzo il rispondente e indurlo a mentire, mentre sono da preferire tecniche di autosomministrazione del questionario o tecniche telefoniche. Nelle altre fasi del processo produttivo statistico (registrazione, codifica, …) l'errore di misura generalmente dipende da due fattori: la presenza di personale (codificatori, addetti alla revisione, …) che svolge l'attività e può non interpretare correttamente il suo ruolo, e l'erronea programmazione degli strumenti utilizzati nel trattamento, per esempio un dizionario per la codifica con errate specificazioni, un piano di analisi errato, una cattiva specificazione delle regole in un piano di controllo e correzione. In letteratura gli errori di misura che si verificano in fase di raccolta dei dati e sono direttamente attribuibili al rispondente sono anche denominati errori di risposta, quelli che intervengono nelle fasi di processo

[7] Per un approfondimento sulla classificazione delle mancate risposte totali si veda l'Appendice C o si consulti Hidiroglou et. al (1993).

successive alla raccolta dei dati possono essere indicati con nomi alternativi, quali errori di elaborazione o errori di trattamento. Anche gli errori di misura possono incidere sull'accuratezza dei dati prodotti causando distorsioni quando l'errore è sistematico e quindi non si compensa in media, o un aumento della variabilità delle stime, se gli errori sono accidentali e quindi in media si compensano. Nei dati di fonte amministrativa, la qualità delle variabili rilevanti ai fini amministrativi è in genere buona, mentre potrebbe non essere tale per variabili che non sono direttamente di interesse per la funzione amministrativa o qualora sia necessario un processo di "derivazione" di una variabile non osservata. Inoltre, anche il ricorso a procedure di integrazione potrebbe portare ad errori di misura, causati dagli errori che sono collegati a tali tecniche (in particolare, falsi abbinamenti).

Errori di assunzione del modello
Si possono verificare nei processi di tipo rilevazione a causa della non corretta specificazione di ipotesi (implicite o esplicite) richieste dall'uso di metodi come la calibrazione e gli stimatori di regressione generalizzata. Il ricorso a modelli è anche alla base di alcune procedure quali per es. la destagionalizzazione. Nel caso dell'uso di dati di fonte amministrativa, il ricorso all'uso di modelli può essere frequente, per es. nella derivazione di variabili e unità. Anche quando si applicano procedure di controllo e correzione, si fa riferimento ad un modello che, se non correttamente specificato può introdurre l'errore piuttosto che rimuoverlo. Se applicati a procedure di generazione e/o derivazione di dati a livello micro, questo errore porta a un errore di trattamento, diversamente dal caso in cui i modelli si applichino a dati di tipo macro.

La Tabella 2 riassume le tipologie di errore descritte e le fasi/sotto-processi in cui esse si generano con più probabilità.

Tabella 2. Tipologie di errore, fasi e/o sotto-processi dove si generano

Tipologia di errore	Fase/sotto-processo o attività dove si genera
Campionario	
Errore di campionamento	B. Scelta del disegno, lista di riferimento, campionamento e stima
Non campionario	
Errore di Specificazione	A. Identificazione delle esigenze degli utenti, definizione dei concetti, scelta delle fonti e valutazione della soddisfazione C. Acquisizione dei dati (in particolare nella progettazione e sviluppo del questionario o dello strumento di rilevazione) I. Derivazione delle variabili
Errore di copertura	B. Scelta del disegno, lista di riferimento, campionamento e stima (in particolare nella creazione e/o aggiornamento delle lista di riferimento) C. Acquisizione dei dati (in particolare del dato di fonte amministrativa) E. Integrazione H. Derivazione unità
Errore di mancata risposta totale / non osservazione	B. Scelta del disegno, lista di riferimento, campionamento e stima (in particolare nell'identificazione/aggiornamento delle informazioni per il contatto delle unità) C. Acquisizione dei dati (sia nella raccolta diretta che da fonte amministrativa)
Errore di mancata risposta parziale	C. Acquisizione dei dati (progettazione e sviluppo del questionario o dello strumento di rilevazione, raccolta dei dati diretta e da fonte amministrativa)
Errore di misura (errori di risposta, di trattamento)	C. Acquisizione dei dati (progettazione e sviluppo del questionario o dello strumento di rilevazione, raccolta dei dati diretta e da fonte amministrativa) E. Integrazione I. Derivazione delle variabili D. F. G. Tutte le altre fasi di trattamento dei dati (conversione in formato elettronico, codifica, identificazione e trattamento degli errori)
Errore di assunzione del modello	Tutte le trasformazioni di dati che utilizzano un modello statistico (per es. modelli per la destagionalizzazione)

Tutte le tipologie di errore presentate hanno impatto prevalentemente sulla dimensione dell'accuratezza ad eccezione dell'errore di specificazione che può compromettere la pertinenza delle statistiche diffuse o anche la comparabilità e coerenza. La qualità del processo produttivo statistico può avere impatto anche su altre dimensioni dell'errore, quali per esempio la puntualità e la tempestività.

Gli errori che si generano in una fase, possono avere impatto anche in fasi successive. Per esempio alcuni errori di lista, quali errori nella classificazione delle unità rispetto a delle variabili utilizzate nel disegno di campionamento, possono inficiare l'efficienza del campione.

Da notare che l'errore quadratico medio già citato può essere scomposto nelle varie fonti di errore, campionario e non campionario, ciascuna delle quali contribuisce con le due componenti, variabilità e distorsione. Tuttavia, il calcolo dell'errore quadratico medio è in generale complesso e costoso e, se calcolato, di solito risulta essere un'approssimazione del vero MSE e considera solo un solo un sottoinsieme delle sue componenti.

Parte II

A. Identificazione delle esigenze degli utenti, definizione dei concetti, scelta delle fonti e valutazione della soddisfazione

Descrizione

La fase iniziale del processo produttivo statistico comprende tutte le attività messe in atto per identificare gli utenti e le esigenze di informazione statistica da questi espresse che andranno tradotte nei relativi obiettivi conoscitivi del processo produttivo statistico. Questa fase, che si concretizza nell'istituzione di un dialogo utente-produttore, è di fondamentale importanza per assicurare poi la pertinenza dell'informazione prodotta. Laddove vi sia un regolamento che norma precisamente le variabili e il relativo livello di classificazione da trasmettere e/o diffondere, è comunque opportuno verificare le necessità informative di altri utenti per realizzare eventuali sinergie nella produzione statistica.

Affinché le esigenze informative possano essere espresse in termini statistici è necessario che siano individuati i concetti e le dimensioni sottese, che questi siano correttamente resi operativi in termini di caratteristiche rilevabili. Questa fase include anche una preliminare riflessione riguardo l'identificazione delle unità statistiche, della popolazione obiettivo e delle classificazioni da utilizzare, il dettaglio informativo e territoriale richiesto, la tempestività con cui le statistiche dovrebbero essere messe a disposizione, le modalità di diffusione dei risultati.

Il processo che permette di passare dalla definizione degli obiettivi conoscitivi all'individuazione delle dimensioni del fenomeno, fino alla scelta delle variabili, consiste in un procedimento di tipo teorico-concettuale. Questo processo ha la finalità di determinare le caratteristiche tecnico-statistiche e operative dell'indagine attraverso un progressivo raffinamento delle relazioni tra le componenti del fenomeno osservato e le caratteristiche effettivamente misurabili sulle unità statistiche che si vogliono conoscere. Alla fine di tale processo risultano essere esplicitati: *i)* la/le popolazioni cui saranno riferite le statistiche; *ii)* le dimensioni spaziali e temporali di quest'ultime; *iii)* le variabili che misurano le caratteristiche di interesse riferite alla popolazione identificata; *iv)* le classificazioni teoriche relative a queste variabili. Per esempio tutto questo si può tradurre nel dover rilevare come popolazione, la popolazione attiva; per il tempo, la data cui si riferisce la popolazione, ossia al 31/03/2018; per lo spazio, il riferimento territoriale cui si riferisce la popolazione, ossia l'Italia; per le variabili da rilevare, lo stato occupazionale (attraverso una batteria di quesiti); per le classificazioni da adottare, la condizione occupazionale dichiarata, che porta a classificare la popolazione in: occupata, disoccupata, in cerca di occupazione. Questi aspetti, se non correttamente considerati, possono provocare gravi ricadute su alcune componenti della qualità come la pertinenza e l'accuratezza.

Una volta identificate le esigenze informative, dovrà essere effettuata una ricognizione delle fonti disponibili per valutare se sia necessario l'avvio di una nuova rilevazione, il ridisegno di una esistente, o se invece possano essere utilizzati dati già prodotti, anche per finalità amministrative (è anche possibile che emerga la necessità di una combinazione di queste attività). Inoltre, sono da considerare anche eventuali vincoli di budget, il fastidio statistico determinato dalla raccolta del nuovo dato, nonché valutazioni legate alla privacy.

Un aspetto relativo alla relazione con gli utenti, ma che viene affrontato alla fine del processo produttivo statistico, riguarda la valutazione del lavoro svolto, sia in termini di accesso all'informazione prodotta sia in

termini di soddisfazione degli utenti. A tal fine possono essere individuati tanto indicatori di monitoraggio quanto approntate vere e proprie rilevazioni sulla soddisfazione degli utenti. Esistono diverse tipologie di rilevazioni. Esse possono essere specifiche per un prodotto statistico e quindi mirate a comprendere la soddisfazione degli utenti rispetto alla pertinenza, alla qualità, all'accessibilità del dato, alla chiarezza delle informazioni a supporto dell'interpretazione del dato. Altre volte, le indagini sulla soddisfazione sono più generali e mirate a misurare la soddisfazione degli utenti rilevando anche il loro profilo come utilizzatori (frequenza e tipo di dati), il motivo dell'utilizzo e le loro caratteristiche demografiche, sociali e professionali dell'utilizzatore.

La principale fonte di errore che si genera in questa fase, riguarda l'errore di specificazione, qualora gli obiettivi conoscitivi non vengano adeguatamente tradotti nei concetti operativi del processo oppure se i dati amministrativi da utilizzare per le finalità statistiche non riflettano correttamente le definizioni statistiche.

Principio A.1. Identificazione degli utenti e delle esigenze informative e loro traduzione nei concetti statistici

Gli utenti dell'informazione statistica e le esigenze informative da soddisfare devono essere ben identificati e documentati. Tali esigenze devono essere tradotte in obiettivi conoscitivi concreti e, a seguire, devono essere chiaramente definiti i fenomeni di interesse, le variabili da rilevare, la popolazione target e le unità di analisi.

Suggerimenti

- I principali utenti devono essere chiaramente identificati e coinvolti nella definizione degli obiettivi e nell'eventuale progettazione o riprogettazione del processo.
- Gli utenti si possono classificare e raggruppare in base alle loro caratteristiche (dove lavorano, uso che fanno dell'informazione statistica, importanza che le statistiche rivestono nel proprio lavoro etc.). Nell'accezione di utente devono essere compresi anche i committenti e i preposti a organi di governo, centrali o locali, che utilizzano le statistiche per finalità decisionali.
- È utile tenere una documentazione aggiornata sul profilo dei principali utenti. A tal fine possono essere analizzati i dati degli utenti registrati presso i servizi dell'ente (sito web, servizi on line etc.) e le richieste ricevute attraverso canali interni (reclami, richiesta di informazioni etc.), presso i *contact center*, gli uffici di relazione con il pubblico o altri uffici.
- È opportuno in fase di registrazione degli utenti, chiedere l'autorizzazione al trattamento dei dati a scopo di indagine o per un successivo ricontatto, in modo da essere legalmente autorizzati a farlo.
- Le esigenze informative espresse dagli utenti possono essere acquisite attraverso l'istituzione di tavoli di confronto utente-produttore preferibilmente di natura continuativa e stabile, o attraverso strumenti di consultazione *ad hoc* come indagini esplorative, *focus group*, interviste a esperti del settore, etc. Ulteriori canali di interazione, che possono aiutare a identificare esigenze informative non soddisfatte, sono conferenze tematiche, incontri con esperti e *stakeholders* etc.
- Nel caso di molteplici utenti portatori di interessi diversi e divergenti che non è possibile soddisfare interamente, sarebbe auspicabile assegnare priorità alle richieste in modo da poter soddisfare quelle maggiormente rilevanti.
- Una volta consultati gli utenti principali e gli eventuali committenti (es. amministrazioni centrali, governo, istituzioni europee, organizzazioni internazionali etc.) dovranno essere chiariti, specificati e documentati in forma scritta gli obiettivi conoscitivi del processo di produzione dei dati, i fenomeni di interesse, le variabili da rilevare, la popolazione di riferimento, l'unità statistica su cui saranno rilevati i dati.

- Se i nuovi obiettivi conoscitivi sono definiti da una base normativa (es. regolamento o direttiva europea, legge nazionale, regolamenti, circolari e altre procedure), laddove non vi sia una chiara specificazione delle modalità e delle fonti di acquisizione a cui far riferimento per colmare la lacuna informativa, dovranno essere approntate azioni simili a quelle riferite in precedenza.

Principio A.2. Scelta delle fonti e minimizzazione del carico statistico sui rispondenti
Deve essere verificata l'esistenza eventuale di dati già disponibili - o la cui rilevazione è prevista per altre finalità - che permettano di soddisfare le esigenze informative degli utenti, al fine di minimizzare il carico statistico sui rispondenti e ridurre i costi.

Suggerimenti

- Prima di procedere al disegno di una nuova indagine è necessario effettuare un'analisi delle fonti di dati già disponibili per il fenomeno da indagare. Si raccomanda, quando possibile, di fare ricorso alle fonti amministrative per evitare eccessive e ridondanti richieste di informazioni ai rispondenti. Sarebbe opportuna, inoltre, la massima condivisione dei dati tra gli Enti produttori di statistiche al fine di limitare le occasioni di rilevazione attraverso il coordinamento già in fase di pianificazione, ad esempio nella predisposizione del Programma statistico nazionale. Pertanto, gli Enti produttori sono invitati a ricorrere allo sfruttamento a fini statistici di tutti i dati a loro disposizione e all'integrazione delle fonti di dati al fine di ridurre l'onere statistico e per garantire la completezza dell'informazione.
- La fase di esplorazione delle fonti potrebbe condurre alla scoperta di fonti di dati non ancora pienamente sfruttate ma utili a soddisfare tutti o parte degli obiettivi conoscitivi individuati. In questo caso è bene condurre un'istruttoria della nuova fonte per valutare preliminarmente l'effettiva rispondenza alle esigenze informative e l'opportunità dell'acquisizione. Anche in caso di utilizzo di fonti amministrative (interne, di soggetti e organi intermedi o fonti esterne) ogni decisione deve essere preceduta da una valutazione delle caratteristiche riguardanti la pertinenza e la qualità dei dati di input contenuti nella fonte.
- Nel caso di disponibilità di più fonti, la scelta deve essere condotta in base a un'analisi comparata che valuti: l'aderenza concettuale del dato amministrativo nel rappresentare quello statistico, la copertura degli archivi amministrativi rispetto alle popolazioni statistiche di interesse, la completezza delle informazioni presenti nelle variabili amministrative di interesse, il flusso di alimentazione e la sua periodicità. Le ipotesi e le motivazioni che hanno condotto alla scelta di utilizzare i dati di fonte amministrativa e il tipo di utilizzo all'interno del processo produttivo statistico dovranno essere opportunamente documentati.

Principio A.3. Soddisfazione della domanda conoscitiva
La rispondenza tra le esigenze informative degli utenti e l'informazione statistica prodotta deve essere garantita. La soddisfazione degli utenti deve essere periodicamente verificata.

Suggerimenti

Al fine di assicurare l'effettiva rispondenza dell'informazione statistica che si intende produrre con le esigenze informative emerse è buona norma che il processo di definizione e produzione dei dati sia affiancato da:

- tavoli di confronto che assicurino un dialogo costante tra produttori e utenti al fine di verificare ex-ante e in itinere oltre che i contenuti anche i requisiti di qualità e le caratteristiche generali del processo (unità, tecnica di raccolta, trattamento dei dati);
- negoziazioni per eventuali divergenze tra le richieste di utenti diversi, in condizione di risorse limitate;
- verifica dell'adeguatezza degli strumenti per la rilevazione delle caratteristiche di interesse;
- produzione e fornitura su richiesta di documentazione utile a dar conto dell'avanzamento e orientamento dei lavori;
- accordi che fissino le caratteristiche, i tempi e i modi di accesso/diffusione dell'informazione.

Ciò è ancor più opportuno nel caso di fenomeni emergenti o in fase di rapido cambiamento.

Per quanto riguarda la valutazione dell'effettiva soddisfazione delle esigenze informative, dopo la diffusione dei risultati, si suggerisce di effettuare:

- con regolarità, il monitoraggio del numero di pubblicazioni richieste, download di dati, accessi a sistemi informativi etc., per misurare l'interesse di una parte dell'utenza rispetto alla statistica prodotta;
- all'occorrenza, discussioni dei risultati del processo in workshop o seminari, insieme ad esperti di settore, per comprenderne a fondo il significato e la valenza;
- periodicamente, indagini sugli utenti per rilevarne la soddisfazione rispetto al prodotto e al servizio (modalità e condizioni di accesso al prodotto) tenendo conto del loro profilo.

Indicatori di qualità e performance

È possibile ricavare facilmente dai sistemi informativi di diffusione dei dati o pagine web un indicatore che riflette la pertinenza delle statistiche diffuse:

A.1. Frequenza di accesso ai dati pubblicati.

Da notare che alcune statistiche non cliccate possono essere altrettanto pertinenti (per es. attemperano ad un regolamento)

Indicatori sul carico statistico si possono calcolare per i processi di tipo multifonte, che affiancano l'uso di dati amministrativi o dati già disponibili da altre fonti a quelli di indagine. Essi possono essere calcolati relativamente alle unità e alle variabili (quando si osservano intere sottopopolazioni e intere variabili dalle fonti amministrative):

A.2: Rapporto tra le unità da indagine e quelle da fonte amministrativa

A.3. Rapporto tra le variabili da indagine e quelle da fonte amministrativa

Sulla soddisfazione degli utenti, gli indicatori possono derivare dall'analisi dei quesiti inclusi in indagini sulla soddisfazione.

Mappatura con i sotto-processi del GSBPM

1.1, 1.2., 1.3., 1.4., 1.5., 2.1., 2.2., 8.1.

Riferimenti bibliografici

Blanc, M., Radermacher, W. and Körner, T. 2001. “Quality and users.” International Conference on Quality in Official Statistics 2001. Session 15.1. Stockholm, Sweden.

Brackstone, G.J. 1993. “Data relevance: keeping pace with user needs.” Journal of Official Statistics. Vol. 9, no. 1, p. 49-56.

Code of Practice for Official Statistics, Edition 1.0, January 2009, UK Statistics Authority - London Statistics Canada. 2000 “Policy on Informing Users of Data Quality and Methodology.” Statistics Canada Policy Manual. Section 2.3. Last updated March 4, 2009. "http://icn-rci.statcan.ca/10/10c/10c_010_e.htm"

Croatian Bureau of Statistics (CBS) – User Satisfaction Survey- Zagreb, May 2015. http://www.dzs.hr/Eng/international/Quality_Report/Quality_Report_Documents/Quality_Report_Satisfaction_Survey.pdf

Department for Communities and Local Government - Engagement strategy to meet the needs of statistics users – London, January 2015 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/393100/UserEngagementStrategy__2015__-_Cover.pdf

Eurostat “Report on the EUROSTAT 2015 User satisfaction survey” "http://ec.europa.eu/eurostat/documents/64157/4375449/General+report-USS-2015/2ebe0f43-ad8d-4689-b63b-e772ea947dac"

EUROSTAT - Leadership Group (LEG) on Quality - Implementation Group “State-of-the-art regarding planning and carrying out Customer/User Satisfaction Surveys in NSIs” (LEG on Quality Recommendation No. 7)

Istat - Risultati della “Rilevazione sul grado di soddisfazione relativo ai prodotti e servizi offerti sul sito web www.istat.it”. Anno 2014 http://www.istat.it/it/files/2015/10/Report_questionario_2014_complessivo_per_web.pdf?title=Misurare+il+grado+di+soddisfazione+degli+utenti+-+01%2Fott%2F2015+-+Report+questionario+2014.pdf

ONS - Customer Satisfaction Survey 2015/16 "https://www.ons.gov.uk/aboutus/whatwedo/statistics/consultationsandsurveys/allconsultationsandsurveys/annualcustomersatisfactionsurvey"

B. Scelta del disegno, lista di riferimento, campionamento e stima

Descrizione

Questa fase tratta tutti gli aspetti relativi alla scelta del tipo di disegno (se campionario o esaustivo), il campionamento, l'estrazione del campione, il processo di stima e l'aggiustamento dei pesi campionari, il calcolo degli errori campionari.

Il disegno dell'indagine può essere esaustivo oppure prevedere l'osservazione di un sottoinsieme di elementi della popolazione, scelti secondo un meccanismo probabilistico oppure secondo un criterio non casuale. Generalmente, quando si utilizzano dati di fonte amministrativa, si ha a disposizione l'intera popolazione di interesse e non se ne estrae un suo campione, applicando così un'ottica di disegno esaustivo.

La teoria del campionamento è molto vasta e complessa e per un approfondimento teorico si può fare riferimento ad uno dei numerosi testi esistenti in letteratura. Qui si suppone che vi sia già una conoscenza di base della tematica e si vuole fornire una panoramica utile per la successiva definizione dei suggerimenti.

In una indagine campionaria, il disegno di campionamento consiste nel definire una distribuzione di probabilità che assegna ad ogni sottoinsieme (campione) della popolazione obiettivo una probabilità di essere osservato e uno schema operativo di selezione del campione. L'identificazione di questi due elementi dipende dal tipo di popolazione obiettivo e dei parametri che si vogliono stimare (a livello di intera popolazione e/o per suoi domini), dalla lista di riferimento e più in generale delle risorse che si hanno a disposizione.

Nei campioni probabilistici, ogni elemento della lista di riferimento (unità) ha una probabilità nota e non-nulla di essere inclusa nel campione. Il principio sottostante il processo di stima da un campione probabilistico è che le unità incluse nel campione siano rappresentative anche delle unità della popolazione non incluse nel campione. Ciò avviene attribuendo a ciascuna unità inclusa nel campione un peso (peso diretto, ossia inverso della probabilità di inclusione) che può essere visto come il numero di elementi della popolazione rappresentati da tale unità. Nei campioni non probabilistici il principio di stima non si avvale di una formalizzazione probabilistica legata all'inclusione casuale dell'unità nel campione. Il meccanismo di selezione del campione può quindi essere ragionato o soggettivo. La rappresentatività di un campione non probabilistico dipende da assunti teorici che sono formulati dal ricercatore. Per esempio, se ad un intervistatore viene chiesto di selezionare il campione, la sua scelta ricadrà su soggetti più disponibili e facilmente accessibili introducendo così una distorsione da selezione (*selection bias*), qualora la disponibilità a partecipare all'indagine sia correlata con un certa tipologia di risposte. Tuttavia, se si assume che la disponibilità a partecipare dipenda esclusivamente dall'età (o classi di età), l'intervistatore può selezionare non casualmente un campione, con l'attenzione di rispettare nel campione quote di interviste per età (quote ad esempio pari a quelle note per la popolazione obiettivo) eliminando o riducendo il *selection bias*. In tale campione detto per quote, le unità di una certa classe sono selezionate fino a quando non si raggiungono delle quote fissate.

In alcuni casi un campione non probabilistico è l'unica strada perseguibile. Il tipico esempio è quando non si ha a disposizione una lista esaustiva di unità della popolazione obiettivo. Alle unità non presenti nella lista è assegnata implicitamente una probabilità di inclusione nulla.

Alcuni tra i principali tipi di campionamento non probabilistico oltre a quello per quote sono: volontario, di esperti, *cut-off*, a valanga o palla di neve. Un esempio di campione di volontari si ha quando da un pop-up che si apre accedendo ad un sito viene chiesto di partecipare volontariamente alla rilevazione. A volte può

essere utile intervistare un campione di esperti (*focus group*). Questo tipo di campione è spesso di dimensioni molto contenute ed è spesso utilizzato per il test del questionario. Nel campionamento *cut-off* il criterio di scelta è deciso sulla base di alcune caratteristiche, per esempio quando viene deciso di osservare solo le unità al di sopra di una certa soglia scelta in funzione di una variabile nota e rilevante (per es. imprese il cui ammontare del fatturato raggiunge una certa percentuale del totale). Alle unità sotto la soglia è assegnata una probabilità di inclusione nulla e il campione è pertanto non probabilistico. Infine, per popolazioni di tipo elusivo, cioè che tendono a nascondersi spesso viene utilizzato un criterio a palla di neve, ossia attraverso le reti relazionali (sociali, culturali, politiche) di un gruppo di persone inizialmente contattate. Del resto, utilizzando un campionamento probabilistico tradizionale si possono avere due principali criticità: le unità delle popolazioni elusive non sono presenti nella lista di riferimento, le unità elusive se selezionate nel campione probabilistico sono difficili da contattare.

Esistono vari tipi di campionamento probabilistico ciascuno con diversi livelli di complessità e con i propri vantaggi e svantaggi: casuale semplice, sistematico, con probabilità proporzionali alla dimensione (PPS), a grappolo, stratificato, a più stadi. Nella pianificazione di un disegno campionario, in primo luogo si sceglie lo schema di campionamento (per es. a più stadi) e quindi gli eventuali criteri di stratificazione (definizione del numero degli strati, scelta delle variabili di stratificazione). Si determina l'ampiezza del campione che dipende dalla precisione delle stime richiesta. Quindi si procede a stabilire il metodo probabilistico di selezione delle unità campionarie: con probabilità uguali o variabili, proporzionali ad una misura di ampiezza supposta correlata con le variabili oggetto (tipicamente nei disegni a più stadi sulle famiglie, l'ampiezza demografica di un comune per le unità di primo stadio). Infine, si pianificano le numerosità campionarie per i diversi stadi di selezione e l'allocazione del campione tra gli strati. Brevemente, l'allocazione può avvenire a partire da una dimensione del campione fissata distribuendola tra gli stati oppure sulla base dell'errore di campionamento ammesso per le principali stime in relazione ai domini di riferimento, poi sommando le relative numerosità.

Sulla base del disegno pianificato, si procede all'estrazione del campione dalla lista di riferimento e alle successive fase di raccolta e di trattamento dei dati.

La lista di riferimento o archivio di estrazione deve rappresentare fedelmente la popolazione obiettivo. Tuttavia, nei casi reali si ha a disposizione una lista che rappresenta con un certo grado di approssimazione la popolazione obiettivo. Esempi di imperfezioni della lista di riferimento sono: la sotto-copertura (unità della popolazione obiettivo non presenti nella lista); la sovra-copertura (unità della lista che non appartengono alla popolazione obiettivo); le duplicazioni (unità che compaiono più volte nella lista e la cui molteplicità è riconosciuta solo nella rilevazione sul campo); errori nelle informazioni per il contatto (ad esempio l'indirizzo o il numero di telefono errato) che non permettono la somministrazione del questionario una volta inclusa l'unità nel campione; errori nelle informazioni che permettono la classificazione delle unità in strati funzionali al disegno di campionamento. Una delle maggiori cause delle imperfezioni della lista è il suo mancato aggiornamento. La selezione del campione spesso avviene su liste precedenti al periodo di riferimento dell'indagine campionaria e della concreta realizzazione sul campo. Per ridurre gli effetti di una lista di selezione imperfetta, si può ricorrere a tecniche di campionamento non standard come il campionamento indiretto, o in fase di stima modificando i pesi diretti.

La stima dei parametri della popolazione può essere effettuata ricorrendo diversi approcci di stima: *i)* metodi diretti, che usano i valori della variabile di interesse osservati sulle sole unità del campione appartenenti al dominio di interesse; *ii)* metodi indiretti: che utilizzano i valori della variabile di interesse osservati sulle unità del campione appartenenti ad un dominio più ampio contenente il dominio di interesse e/o ad altre occasioni di indagine; *iii)* metodi di stima basati su un modello di superpopolazione.

Nel processo di stima diretto, il parametro di interesse viene calcolato come funzione dei valori relativi alla variabile oggetto di indagine e dei pesi. Le procedure di aggiustamento dei pesi sono il modo più comune per correggere i problemi di mancata risposta totale e copertura (imperfezione delle lista). Si basano sull'ipotesi che le unità rispondenti rappresentino sia loro stesse che quelle non osservate. I pesi del disegno relativi alle unità non osservate sono quindi "redistribuiti" tra i rispondenti; una possibile procedura consiste nel moltiplicare i pesi dei rispondenti per un fattore di aggiustamento (inverso della probabilità di risposta) calcolato a partire dal tasso di risposta in gruppi ritenuti omogenei rispetto alla propensione a partecipare all'indagine. Allo stesso modo tali pesi possono essere corretti per problemi di copertura. I pesi dei rispondenti sono corretti per un fattore di aggiustamento pari all' inverso della probabilità di essere sotto-coperto dalla lista di riferimento. I pesi sono ulteriormente aggiustati per tener conto dei vincoli di uguaglianza tra alcuni parametri noti della popolazione e le corrispondenti stime campionarie, attraverso un meccanismo di calibrazione o di ponderazione vincolata, o più semplicemente attraverso la post-stratificazione, ottenendo così dei pesi finali utilizzati nella stima. Se le variabili utilizzate nella calibrazione sono anche correlate con il fenomeno di interesse si ottiene anche una maggiore precisione delle stime, oltre alla coerenza con altre fonti.

Nel caso di dati di fonte amministrativa la stima si ottiene direttamente dai dati se si può assumere che la popolazione obiettivo sia coperta dalle fonti amministrative e che la variabile statistica di interesse coincida nella definizione con quella amministrativa, a meno differenze casuali e non sistematiche.

Approcci indiretti, che utilizzano modelli statistici, possono essere applicati nel caso rilevazioni campionarie quando si abbiano esigenze di stima per domini di piccole dimensioni (domini in cui la dimensione della popolazione è ridotta e/o il fenomeno di interesse è raro). Nelle indagini che utilizzano campioni non probabilistici, e nel caso di dati di fonte amministrativa se l'ipotesi di assenza di errori di copertura e di specificazione non può essere considerata soddisfatta, si utilizza il terzo approccio alla stima basato sul modello. Dal campione si stima un modello che si ritiene generi la variabile di interesse, si predice il valore della variabile per tutte le unità non incluse nel campione secondo il modello stimato, si produce la stima per somma tra valori osservati nel campione e valori predetti dal modello.

Il processo di stima mediante campione genera un risultato che è affetto da variabilità indotta dall'osservazione parziale della popolazione obiettivo. Nei disegni di campionamento probabilistico, infatti, il processo di selezione può generare realizzazioni diverse del campione, introducendo una variabilità nella statistica campionaria, denominata varianza campionaria, che è una misura inversa della precisione delle stime. Nei disegni di campionamento non probabilistici la varianza delle stime dipende dalla varianza del modello statistico di superpopolazione utilizzato per predire i valori della variabile di interesse delle unità non osservate nel campione.

Oltre alla variabilità, le stime campionarie possono essere affette da distorsione. La distorsione ha una accezione diversa tra il processo di stima che utilizza campioni probabilistici e il processo di stima che utilizza campioni non probabilistici. Nel caso di campioni probabilistici la distorsione deriva da una non osservazione sistematica di alcuni elementi della popolazione (probabilità di inclusione nulla o probabilità di risposta nulla), che sono portatori di caratteristiche diverse rispetto al campione selezionato. Nei campioni non probabilistici, dove la selezione segue criteri soggettivi e non casuali, si parla di possibile *selection bias* e di scelta di un modello di lavoro per la predizione dei valori delle variabili di interesse sulle unità non incluse nel campione che è diverso dal vero modello che genera realmente la variabile di interesse (*model bias*).

Per la valutazione degli errori campionari delle stime prodotte si deve far ricorso a metodi di calcolo della varianza approssimati basati su metodi analitici o tecniche di ricampionamento.

Principio B.1. Campionamento e processo di stima

La scelta di una rilevazione esaustiva deve essere oggettivamente motivata e il suo ricorso giustificato.
Il disegno e la dimensione del campione devono essere tali da garantire il livello di accuratezza prefissato per le variabili chiave in corrispondenza dei principali domini di studio.
Le stime devono essere prodotte utilizzando metodologie consolidate. Le assunzioni alla base dell'uso di informazioni ausiliarie e di approcci da modello devono essere esplicitate e ne deve essere valutata l'effettiva validità. Le stime prodotte devono essere accompagnate da stime dell'errore.

Suggerimenti

Disegno

- Il ricorso ad una rilevazione esaustiva rispetto ad una rilevazione campionaria deve essere deciso sulla base di una molteplicità di elementi, che tengano conto degli obiettivi di stima, dei costi, del carico sui rispondenti, dell'accuratezza dei risultati.
- Una rilevazione esaustiva può essere condotta: laddove vi sia necessità di produrre stime delle quantità di interesse su piccoli domini di studio; quando la popolazione sia di dimensioni relativamente contenute tali da non compromettere il lavoro sul campo.
- Considerare che una rilevazione esaustiva porta un elevato carico statistico sui rispondenti.
- Considerare che una rilevazione esaustiva può comportare delle difficoltà di realizzazione e conseguentemente un elevato livello di errore non campionario.
- Una rilevazione campionaria ben pianificata e realizzata, a fronte dell'introduzione dell'errore campionario, può portare a stime più accurate perché meno affette da errore non campionario

Lista di riferimento

- Identificare chiaramente sia la popolazione obiettivo che la lista di riferimento da utilizzare per identificare e contattare le unità della popolazione.
- In presenza di più liste di riferimento, la scelta deve essere giustificata.
- Nel caso di una lista che non riflette adeguatamente la popolazione, utilizzare tutta l'informazione disponibile per integrarla con altre fonti.
- L'aggiornamento della lista dovrebbe essere il più possibile vicino al periodo di riferimento dei dati dell'indagine che la utilizza.
- È opportuno valutare la qualità della lista da utilizzare in termini di copertura e di errori nelle informazioni in essa contenute.

Campionamento non probabilistico

- Il ricorso a un disegno di campionamento non probabilistico deve essere giustificato sia dal punto di vista teorico che pratico.
- La non disponibilità di un archivio o lista di estrazione può obbligare la scelta di un disegno di campionamento non probabilistico.
- Trarre conclusioni sulla popolazione oggetto di studio a partire da un campione non probabilistico può essere fuorviante in quanto vi è un elevato rischio che le stime possano essere affette da distorsione (*selection bias e model bias*). In tali circostanze è, quindi, opportuno specificare le assunzioni sottostanti l'inferenza e cercare e documentare le evidenze a sostengo della validità delle assunzioni del modello.

Campionamento probabilistico

- Il disegno di campionamento deve essere adeguato rispetto agli obiettivi dell'indagine; deve considerare la tecnica di rilevazione e i costi connessi; deve tener conto delle informazioni contenute nella lista di selezione e deve garantire che ciascuna unità della lista di campionamento abbia una probabilità non nulla di essere inclusa nel campione (nel caso di disegni a più stadi, questo deve avvenire per ciascuno stadio).
- Informazione ausiliaria disponibile per tutte le unità del campione deve essere sfruttata nel disegno campionario
- È opportuno che il disegno di campionamento preveda una stratificazione delle unità in modo da creare strati omogenei di unità rispetto alle informazioni che si vogliono raccogliere e, se possibile, tale che i principali domini di studio possano essere ottenuti dalla unione di strati elementari. La stratificazione è molto importante per fenomeni distribuiti in modo asimmetrico nella popolazione (*skewed*). Questi disegni sono associati a maggiore precisione delle stime. Disegni che concentrano il campione sul territorio permettono di ridurre i costi in caso di interviste con faccia a faccia, ma possono comportare una perdita di precisione rispetto ad un disegno di pari unità ma non 'concentrato'.
- L'ampiezza ottimale del campione deve essere determinata con metodi statistici in modo da garantire una adeguata precisione delle stime per le principali variabili d'indagine a livello di intera popolazione e per i principali domini di studio. Laddove ci si attende una consistente riduzione della numerosità campionaria per via di un alto numero di unità non eleggibili o di mancate risposte totali, può essere utile selezionare più unità campione di quante ne servano (sovracampionamento).
- Il disegno di campionamento deve permettere la stima dell'errore campionario (varianza campionaria).

Selezione del campione

- La lista di riferimento per l'estrazione del campione deve essere identificata con chiarezza, valutandone l'adeguatezza rispetto agli obiettivi dell'indagine.
- La lista deve essere il più aggiornata possibile. Valutazioni sull'effetto del non aggiornamento della lista andrebbero condotte, in particolare relativamente alle informazioni che permettono il contatto delle unità e la loro stratificazione in gruppi utili per il disegno di campionamento e alla presenza nella lista di unità non appartenenti alla popolazione obiettivo.
- La selezione fisica del campione a partire dalla lista dovrebbe essere condotta mediante software generalizzato. L'utilizzo di software sviluppato ad hoc deve essere limitato a situazioni particolari e lo stesso deve essere ampiamente testato prima del suo utilizzo, per evitare che errori di programmazione possano inficiare la casualità del campione.

Processo di stima

- La procedura per derivare le stime di interesse (stime di livelli, rapporti, tabelle di contingenza, ecc.) deve essere chiara e ben definita.
- Nelle indagini con campioni probabilistici, i pesi che derivano direttamente dal disegno di campionamento (pesi diretti), devono essere corretti per compensare l'impatto di errori di natura non campionaria (mancate risposte totali, sotto-copertura), e per sfruttare le informazioni ausiliarie disponibili al fine di ricavare stime più precise delle quantità di interesse (es. calibrazione).
- La correzione dei pesi diretti per compensare problemi di natura non campionaria (mancata risposta totale, sotto-copertura) deve essere condotta utilizzando metodologie ben consolidate, condivise a livello nazionale o internazionale e deve essere documentata.

- Le informazioni ausiliarie utilizzate nel processo di stima (per aumentare la precisione delle stime, per garantire la coerenza con altre fonti) devono essere correlate con le variabili di indagine e provenire da fonti accurate. In presenza di più variabili ausiliarie è opportuno spiegare come si è proceduto alla scelta delle variabili effettivamente utilizzate.
- Le stime devono essere accompagnate da una misura dell'errore. Tali misure devono tener conto, se possibile, dei principali errori (campionari e non campionari) riscontrati nell'intero processo.
- Prima di produrre le stime è opportuno definire dei criteri per la pubblicazione delle stesse: ovvero stabilire il livello di errore oltre il quale la stima non viene pubblicata.
- Nell'effettuare l'elaborazione è preferibile utilizzare software generalizzato. Nel caso si faccia ricorso a software sviluppato ad hoc, l'intero programma deve essere attentamente testato prima di procedere alla elaborazione delle stime finali.
- Tutti i risultati del processi di stima si devono poter replicare (in modo esatto o con approssimazioni trascurabili), nel senso che ripetendo tutte le procedure di elaborazione si devono ottenere gli stessi risultati.
- Per indagini con campione probabilistico, dovrebbe essere prodotta una stima della varianza campionaria per le stime più importanti, a livello di intera popolazione e dei principali domini di studio. Tale stima deve tener conto delle caratteristiche del disegno (stratificazione, selezione su più stadi, …) e delle correzioni apportate ai pesi. Quando la stima della varianza campionaria è stata desunta applicando solo dei metodi approssimati, la scelta deve essere documentata.
- Nel caso di dati di fonte amministrativa, se si può assumere che la popolazione statistica obiettivo sia coperta dalle fonti amministrative e che la variabile statistica di interesse coincida nella definizione con quella amministrativa, si può derivare la stima direttamente dai dati.
- Nel caso di dati di fonte amministrativa, affetti da problemi di copertura, in taluni casi il ricorso ad approcci basati sulla calibrazione o approcci basati su modelli statistici di predizione può attenuare l'impatto di tali errori sulle stime.
- L'utilizzo di modelli nel processo di stima, sia esso applicato a campioni non probabilistici che a quelli probabilistici, deve essere giustificato e le assunzioni alla base degli stessi devono essere rese esplicite, plausibili e supportate da evidenza, per esempio testate su dati di altre indagini campionarie disponibili relative alla stessa popolazione.
- Laddove possibile si dovrebbe valutare quale possa essere l'impatto sulle stime, in termini di varianza e, se possibile, distorsione, di errori di natura non campionaria.

Documentazione

- Tutti gli aspetti relativi al disegno di campionamento devono essere opportunamente documentati. Per esempio se il disegno è a più stadi, per ogni stadio le unità di estrazione, le variabili di stratificazione, la probabilità di inclusione e lo schema di estrazione, il metodo di estrazione. Inoltre deve essere documentato l'eventuale processo di aggiustamento dei pesi. Per il processo di stima devono essere esplicitate i totali noti cui ci si vincola (nella calibrazione) e i domini di stima.
- È buona norma produrre una nota metodologica che descriva tutti gli aspetti relativi al disegno di campionamento e al processo di stima.

Indicatori di qualità e performance

Indicatori di qualità della fase di campionamento sono:

B.1. la frazione di campionamento ossia il rapporto tra la numerosità del campione e quella della lista cui si riferisce;

B.2. l'efficienza complessiva del disegno di campionamento utilizzato, calcolata attraverso il rapporto tra la varianza del campione utilizzato e quella di un ipotetico campione casuale semplice di pari numerosità;

Misure classiche dell'errore di campionamento, in campioni probabilistici, sono:

B.3. *Standard error* dello stimatore, ossia la radice quadrata della stima della sua varianza campionaria (misura che dipende dall'unità di misura della stima).

B.4. Coefficiente di variazione: ossia il rapporto tra lo *standard error* della stima campionaria e la media delle stime su tutti i possibili campioni, stimata come rapporto tra lo *standard error* e la stima stessa (di solito espressa in percentuale e quindi più facilmente interpretabile).

B.5. Intervallo di confidenza, ossia intervallo attorno alla stima che comprende il vero valore del parametro della popolazione con un dato livello di probabilità.

Mappatura con i sotto-processi di GSBPM

2.4, 3.2, 4.1, 5.6, 5.7, 8.1

Riferimenti bibliografici

Cochran, W. (1977). *Sampling Techniques*. Wiley, New York.

Deville, J. C., and Särndal, C. E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, 376-382.

Lavallée, P (2007) *Indirect sampling.* Springer, New York.

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"

Särndal C.E., Lundström S. (2005) *Estimation in Surveys with Nonresponse*. Wiley, New York.

Särndal C.E., Swensson B., Wretman J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Statistics Canada (2010) *Survey Methods and Practices.* Statistics Canada, Catalogue no. 12-587-X, Ottawa. http://www.statcan.gc.ca/pub/12-587-x/12-587-x2003001-eng.htm

Valliant R.,Dorfman A. H., Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley.

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

C. Acquisizione dei dati

Descrizione

La fase può consistere: nella raccolta diretta dei dati presso individui, imprese, istituzioni; nell'acquisizione di dati di fonte amministrativa o di altre fonti (es. big data); nell'acquisizione di dati provenienti da rilievi e/o misurazioni sul territorio (per es. centraline che rilevano inquinamento); nel supporto tecnico-metodologico all'acquisizione dei dati affidata ad uffici territoriali dell'ente, altri enti pubblici o privati, società esterne.

L'esigenza di ridurre i costi e il disturbo statistico per i rispondenti, orienta verso un sempre maggiore utilizzo di dati amministrativi per finalità statistiche. Questi ultimi sono dati raccolti e gestiti da un ente o da una pluralità di enti, solitamente pubblici, per esigenze legate ad adempimenti legislativi o amministrativi (es. concessione di benefici, somministrazione di sanzioni, ecc.) che non necessariamente coincidono con le finalità statistiche. Includono, tra gli altri, dati di natura demografica o fiscale.

Le tecniche di raccolta dirette vanno dalle tecniche per auto-somministrazione, alla rilevazione telefonica e faccia-a-faccia. Un importante insieme di tecniche è quello delle rilevazioni assistite da computer: telefonica (*Computer Assisted Telephone Interview*, CATI), personale (*Computer Assisted Personal Interview*, CAPI), web (*Computer Assisted Web Interview*, CAWI).

Per le tecniche assistite da computer contestualmente alla raccolta dei dati avviene la loro conversione in formato elettronico o registrazione (si confronti Sezione D).

I dati acquisiti possono provenire anche da approcci di tipo misto, che combinano dati raccolti sia tramite rilevazione diretta sia da fonte amministrativa.

Nella fase di raccolta dei dati, un'attenzione particolare va posta ai dati personali e a quelli sensibili e giudiziari (si veda Appendice D).

Per quello che riguarda l'obbligo di risposta[8], gli enti e gli organismi pubblici hanno l'obbligo di fornire i dati e le notizie che vengono loro richiesti per l'esecuzione dei lavori compresi nel Programma statistico nazionale, ad eccezione di quelli sensibili e giudiziari. Per i soggetti privati, l'obbligo di risposta sussiste, invece, limitatamente ai lavori del Programma statistico nazionale inseriti in un apposito elenco approvato ai sensi dell'art. 13 comma 3-ter del decreto legislativo n. 322 del1989.

Un'innovazione recente è rappresentata dall'utilizzo per finalità statistiche dei cosiddetti big data, ovvero i flussi d'informazione che passano per le reti create dalla tecnologia. La mole di questi dati è talmente grande da non poter essere gestita con strumenti convenzionali per estrapolare, gestire e processare le informazioni entro un tempo ragionevole. Le dimensioni di crescita dei big data sono definite dal cosiddetto modello "4V" (volume, velocità, varietà e variabilità).

Nella fase di acquisizione dei dati, sia essa diretta che da fonte amministrativa, si possono generare degli errori non campionari di varia natura, che possono causare aumento di variabilità e distorsione delle stime. Nel caso di utilizzo di dati di fonte amministrativa, i principali errori sono attribuibili a problemi di copertura degli archivi acquisiti e a possibili problemi nella fase di trasmissione dei dati. Nel caso di rilevazioni dirette su individui, imprese e istituzioni, le tipologie di errore maggiormente rilevanti sono: mancate risposte totali, mancate risposte parziali ed errori di misura. L'entità degli errori varia in funzione della tecnica scelta, del questionario, degli argomenti trattati, della capacità di ottenere la collaborazione da parte del rispondente,

[8] Ai sensi dell'art. 7, comma 1, del decreto legislativo n. 322 del 1989.

etc. Laddove la tecnica di rilevazione preveda l'impiego di intervistatori, questi possono a loro volta introdurre variabilità aggiuntiva e/o distorsione nelle stime (effetto intervistatore). Inoltre, gli aspetti gestionali relativi all'acquisizione possono influire sulla puntualità e tempestività dei dati.

Principio C.1. Acquisizione dei dati di fonte amministrativa
L'acquisizione di dati di fonte amministrativa provenienti da un altro ufficio dell'ente, da unità locali dislocate sul territorio, da uno o più enti/società esterni pubblici e privati deve essere regolamentata da accordi formali che fissino i requisiti e le modalità relative alla trasmissione, alla documentazione e ai livelli di qualità attesi. La qualità dei dati acquisiti deve essere periodicamente monitorata e valutata.

Suggerimenti

Nel caso di acquisizione di dati amministrativi da altro ufficio dell'ente, da unità locali dislocate sul territorio, da uno o più enti/società esterni pubblici e privati, è opportuno stabilire e mantenere buoni rapporti con i fornitori dei dati e collaborare per il miglioramento continuo della qualità dei dati fornendo regolarmente informazioni di ritorno sulla loro idoneità ad essere utilizzati per finalità statistiche. In particolare è opportuno:

- formalizzare accordi che fissino i tempi di trasmissione dei dati, i livelli di qualità attesi, la documentazione di supporto alla trasmissione. Nei casi in cui i dati siano prodotti da altri uffici dell'ente o tramite società esterne, i termini di tali accordi dovranno essere chiariti e condivisi anche con questi ultimi in modo che non vi siano incongruenze;
- identificare una persona di riferimento per il trasferimento dei dati da ciascuna fonte;
- collaborare affinché il modello amministrativo converga, quando possibile, con il modello di rilevazione del dato statistico;
- prevedere, nel caso di trasmissione da una pluralità di fonti che seguono modalità e formati diversi, una fase di aggregazione secondo un formato unico e attivare controlli di qualità sul materiale ricevuto;
- acquisire, in presenza di unità locali dislocate sul territorio, informazioni sui controlli effettuati in fase di acquisizione dagli stessi organi, le segnalazioni di eventuali problematiche riscontrate e la distribuzione dei carichi di lavoro del personale;
- considerare la normativa relativa alla rilevazione di dati sensibili e giudiziari. Nel caso in cui la raccolta dei dati non sia effettuata presso l'interessato (statistiche da fonte amministrativa o indagini presso terzi), la facoltatività della risposta dovrà essere assicurata attraverso l'adozione di specifiche misure organizzative tese a garantire l'adesione volontaria dell'interessato al trattamento dei suoi dati sensibili e giudiziari per finalità statistiche;
- collaborare, laddove possibile, con gli enti titolari del dato amministrativo per la definizione dei contenuti da rilevare in modo da far convergere le esigenze amministrative con quelle statistiche;
- fornire un feedback ai fornitori dei dati amministrativi per migliorare la qualità dei dati e il processo di trasmissione;
- monitorare le variazioni normative e procedurali nel tempo e sul territorio che possono avere impatto sui dati amministrativi acquisiti;
- predisporre una adeguata modulistica, per es. attraverso dei modelli ausiliari, per la raccolta di informazioni di supporto al dato amministrativo.

Durante la fase di ricezione e a conclusione delle attività di acquisizione è opportuno monitorare e misurare la qualità dei dati ricevuti (qualità dell'input), ossia:

- monitorare la qualità delle trasmissioni attraverso appositi indicatori per poter intervenire tempestivamente presso la fonte e, laddove possibile, ottenere dei nuovi trasferimenti di dati corretti;
- valutare la qualità dei dati amministrativi finali acquisiti attraverso il calcolo di misure appropriate (in termini di acquisizione, documentazione e accuratezza).

La documentazione di supporto ai dati amministrativi è di fondamentale importanza per il loro utilizzo. La completezza e chiarezza della documentazione deve essere valutata e in caso di carenze, richiesta dagli organismi fornitori.

Indicatori di qualità e performance

Per il monitoraggio e la valutazione dei dati amministrativi esistono numerosi indicatori in letteratura. In genere sono organizzati in indicatori relativi alla *fonte e alla fornitura*, alla *documentazione* o metadati e ai *dati* (Daas et al, 2009). Di seguito, senza voler essere esaustivi, si riportano alcuni indicatori utili come esempio.

Indicatori sulla fonte e fornitura

Attengono all'identificazione del titolare del dato amministrativo, all'esistenza di accordi, alla tempestività e puntualità della trasmissione. In particolare, si suggerisce di verificare almeno questi elementi:

C.1. Tempi previsti per la fornitura dei dati rispetto a quelli effettivi
C.2. Costi di fornitura

Indicatori sulla documentazione o metadati

C.3. Valutazione sulla completezza e chiarezza della documentazione a supporto dei dati: unità e variabili

Indicatori sui dati

Comprendono sia i controlli tecnici sui dati ricevuti (leggibilità e aderenza del file e del tracciato record a quanto concordato), sia misure della qualità dei dati come per esempio indicatori di copertura rispetto alla popolazione obiettivo dell'archivio amministrativo utilizzato e indicatori di tempestività dei dati. In Appendice B sono descritti con un maggior dettaglio gli indicatori qui elencati.

C.4. Aderenza del file al formato concordato
C.5. Leggibilità del file
C.6. Aderenza dei dati al tracciato record concordato
C.7. Esistenza di una chiave univoca identificativa dell'unità
C.8. Sotto-copertura della popolazione nell'archivio fornito, ossia numero di unità che dovrebbero far parte dell'archivio e non vi sono incluse sul totale delle unità (incluse e non incluse). Questa può essere anche una indicazione in termini descrittivi se non è possibile calcolarla numericamente, per esempio quando dai metadati sulla definizione delle unità sia disponibile l'informazione relativa a sottopopolazioni non incluse nell'archivio. Si può altresì valutare quella sotto-copertura dovuta ai ritardi della notifica di eventi, anche sulla base di invii di dati successivi.
C.9. Sovra-copertura della popolazione nell'archivio fornito ossia numero di unità dell'archivio che non appartengono alla popolazione su numero totale di unità. Si può altresì valutare quella sovra-copertura dovuta ai ritardi di cancellazione di eventi, anche sulla base di invii di dati successivi.
C.10. Completezza delle principali variabili: per ogni variabile di interesse, % di valori presenti sul totale dei valori dovuti.

C.11. Tempestività dei dati acquisiti: tempo tra la fine del periodo di riferimento dei dati nella fonte e il momento di disponibilità degli stessi

In caso di acquisizione di più archivi da unità locali dislocate sul territorio:

C.12. Indicatori C.8, C.9, e C.10. calcolati per singola unità locale dislocata sul territorio.

Principio C.2. Acquisizione diretta dei dati

Per garantire la qualità e completezza delle informazioni rilevate, si dovrebbe scegliere la tecnica di raccolta più idonea per la tematica oggetto di rilevazione, disegnare il questionario in modo che sia chiaro e facile da somministrare o compilare, favorire e incoraggiare la partecipazione dei rispondenti e curare attentamente la selezione e formazione dei rilevatori. Inoltre, la fase di raccolta dei dati dovrebbe essere monitorata in corso d'opera e valutata a posteriori attraverso strumenti idonei e indicatori oggettivi.

Suggerimenti

Al fine di garantire la qualità di questa fase, è necessario pianificare attentamente e mettere in atto una serie di azioni preventive, di monitoraggio e valutative volte a limitare gli errori che si possono generare o a misurarne l'entità.

Tecnica di raccolta

- Tenere conto della complessità e della vastità degli aspetti da rilevare. Per un argomento che richiede una elevata articolazione dell'intervista, ossia la presenza di "salti" o "svincoli" nel questionario, preferire tecniche assistite da computer, oppure nel caso di questionari cartacei preferire la presenza di un intervistatore in luogo della tecnica per autosomministrazione.
- Valutare la durata attesa dell'intervista. Dovrebbe esser considerato che interviste troppo lunghe non possono essere svolte con tecniche web e telefoniche.
- Preferire, per rilevare gli argomenti sensibili, una tecnica per autosomministrazione o telefonica, se vantaggioso per il complesso degli aspetti, nelle quali l'intervistatore non è presente oppure è presente ma in una forma meno invasiva.
- Preferire, se possibile, le tecniche assistite da computer, in quanto consentono: guadagni di efficienza nel processo produttivo di indagine con conseguente miglioramento della tempestività; l'anticipazione dei controlli (coerenza, dominio e flusso) sulle risposte fornite in fase di rilevazione del dato con possibilità di accertamento delle stesse durante l'intervista.
- Programmare il periodo dell'anno ottimale per effettuare l'indagine: nel caso di indagini presso le istituzioni, il periodo di rilevazione dei dati andrebbe concordato con una rappresentanza delle stesse.
- Valutare la disponibilità dei dati richiesti. Se essi devono essere reperiti all'interno dell'organizzazione, come spesso avviene per indagini presso imprese o istituzioni, è opportuno utilizzare tecniche postali o web, oppure tecniche telefoniche ma precedute da un invio preventivo del questionario.
- Valutare l'opportunità di utilizzare contestualmente più tecniche di tipo diverso (per es. CATI e CAPI) in modo da favorire il rispondente e migliorare i tassi di risposta, tenendo conto del possibile effetto tecnica, ossia impatto che diverse tecniche possono avere sull'errore di misura.

Questionario

Il questionario non è solo uno strumento di raccolta delle informazioni ma è anche un vero e proprio mezzo di comunicazione con il rispondente. Esso è una delle principali fonti di errore di misura e di mancata risposta parziale. È quindi opportuno che vi sia una strategia globale per la progettazione e il test del questionario, che consideri gli aspetti riportati qui di seguito.

- È opportuno strutturare il questionario in modo che raccolga efficacemente le informazioni di interesse senza comportare un eccessivo carico statistico sui rispondenti. La fluidità del questionario dovrebbe essere assicurata attraverso una logica organizzazione delle diverse sezioni. Anche gli aspetti grafici dovrebbero essere utilizzati in modo coerente e consistente all'interno di tutto il questionario (per i quesiti, per le modalità di risposta, per le istruzioni, per i salti,…). La lunghezza del questionario dovrebbe essere contenuta e valutata anche in funzione della tecnica.
- Il linguaggio dovrebbe essere facilmente comprensibile al rispondente; espressioni ambigue dovrebbero essere evitate e termini complessi dovrebbero essere corredati dalle opportune definizioni.
- I quesiti dovrebbero essere concisi e concreti, neutrali, ed esplicitare il tempo e luogo cui fanno riferimento, avere modalità di risposta mutuamente esclusive, essere corredati da istruzioni ed esempi che permettono una maggiore facilità di compilazione.
- Il questionario dovrebbe essere disegnato in modo funzionale alle attività successive: codifica e registrazione dei dati.
- I codici assegnati alle modalità di risposta nei questionari devono essere armonizzati con eventuali classificazioni standard nazionali e internazionali (per es. quella dei comuni). È opportuno prevedere sempre la distinzione tra i valori nulli (zero) e la codifica per i valori mancanti.
- Il questionario elettronico dovrebbe incorporare dei controlli di qualità (cfr. Sezione D).
- Il questionario andrebbe testato attraverso una valutazione interna all'ente e possibilmente anche attraverso un test sul campo quindi in condizioni simili a quelle reali di indagine.

Rispondenti

- Per favorire la partecipazione dei rispondenti, si suggeriscono alcune azioni quali, per esempio: pubblicizzare la rilevazione; inviare una lettera istituzionale di preavviso; fornire ai rispondenti una descrizione sintetica degli obiettivi dell'indagine; garantire esplicitamente la tutela della riservatezza; attivare un numero verde o un indirizzo e-mail per i rispondenti. È buona pratica anche prevedere un sistema di sollecito alle unità non rispondenti al primo contatto.
- Per l'obbligo di risposta da parte di soggetti privati, in assenza di una normativa specifica di carattere comunitario o nazionale, è richiesto l'inserimento della rilevazione in apposito elenco del Programma statistico nazionale.
- Programmare attentamente oltre al periodo di rilevazione anche gli orari di contatto e di intervista nell'arco della giornata in funzione del tipo di unità (per es. individuo o impresa).
- Stabilire se si accetteranno risposte da unità proxy, ossia da individui diversi dalla persona per la quale si intende raccogliere le informazioni. In questo ultimo caso, dovrebbe essere acquisita l'informazione su chi sia il rispondente proxy e sulle risposte da lui fornite.
- Utilizzare una classificazione degli esiti del contatto esaustiva (ad es. se l'unità è eleggibile o non eleggibile, rispondente e non rispondente con il motivo della non risposta). Ciò consente di comprendere la fonte di errore non campionario e di intervenire dove necessario per aumentare i tassi finali di risposta.

Rilevatori

- Effettuare una selezione mirata degli intervistatori in relazione agli obiettivi dell'indagine e al contesto in cui si svolge l'intervista. Ad esempio, quando la tematica concerne aspetti relativi alla violenza sulle donne si tende a preferire intervistatori di genere femminile.
- Gli intervistatori devono ricevere una formazione completa su tutti gli aspetti inerenti gli obiettivi dell'indagine e i contenuti del questionario, la comunicazione, la fase di contatto, le tecniche di conversione dei rifiuti, la gestione dei percorsi del questionario, l'uso del questionario elettronico, etc. Essi devono essere dotati del manuale di istruzioni, di eventuali strumenti ausiliari di supporto e di tutto il materiale utile al loro lavoro.
- Nel caso di dati sensibili e giudiziari l'intervistatore deve evidenziare al rispondente la facoltà di non rispondere a singoli quesiti (d.l. 322/1989; d.lgs.196/2003).
- Evitare un eccessivo turnover dei rilevatori così come un eccessivo carico di lavoro perché possono compromettere la qualità dei dati raccolti.
- Per il monitoraggio in corso d'opera degli intervistatori devono essere predisposti e attuati strumenti di supporto e controllo che variano da incontri con i rilevatori (*debriefing*) per fare emergere eventuali problemi, alla supervisione sul campo, all'effettuazione di telefonate di controllo, all'analisi di indicatori di performance e di qualità.

Indicatori di qualità e performance

Il calcolo e l'interpretazione degli indicatori di qualità in questa fase è fondamentale per comprendere le fonti di errore e apportare correzioni per i dati del processo in corso e miglioramenti per le sue edizioni successive. Si tratta di tassi di mancata risposta totale (su tutto l'insieme delle unità contattate o solo su quelle eleggibili) con le componenti per i motivi (rifiuto, mancato contatto, altri motivi) e stratificati per le variabili che possono avere interesse (per es. geografiche). Indicatori di mancato contatto dovuto a errori nelle informazioni quali indirizzo o telefono sono sintomo di carenze nella lista di riferimento. Indicatori di rifiuto possono riflettere carenze da parte degli intervistatori o nella strategia di comunicazione dell'indagine e richiedere interventi in questo senso, indicatori di abbandono a un certo punto dell'intervista possono riflettere problemi dovuti alla lunghezza del questionario. Questi indicatori possono essere calcolati durante la fase di raccolta dei dati per evidenziare problematiche da risolvere in corso d'opera, oppure al termine della fase di raccolta dei dati, per la valutazione complessiva della performance della fase. L'indicatore di mancata risposta totale può fornire informazioni indirette sulla presenza di distorsione nelle stime finali. Per una tassonomia utile al calcolo degli indicatori si veda Hidiroglou et al. (1993) e l'Appendice C. Per indicatori di qualità armonizzati a livello Europeo si faccia riferimento a Eurostat (2014).

Indicatori di errori nella lista di riferimento

C.13. Tasso di sovra-copertura: numero di unità contattate che non dovevano far parte della popolazione di riferimento sul numero totale di unità contattate

C.14. Tasso di mancato contatto per errore di lista: numero di unità che non è stato possibile contattare a causa di errori nelle informazioni per il contatto (indirizzo, numero di telefono)

Indicatori di mancata risposta totale

C.15. Tasso di risposta[9]: Numero di unità rispondenti / numero di unità totali (se si ha a disposizione l'informazione sullo stato di eleggibilità, questo indicatore si può calcolare anche sul totale delle unità eleggibili)

[9] In tasso di mancata risposta totale è il complemento a 1.

C.16. Tasso di rifiuto: Numero di unità che si sono rifiutate di partecipare alla indagine/numero di unità totali (o sulle eleggibili)

Indicatori di monitoraggio sugli intervistatori

C.17. Numero medio giornaliero di interviste per intervistatore (carico di lavoro)
C.18. Tasso risposta totale per intervistatore (numero di interviste completate su quelle previste)

Indicatori di mancata risposta parziale

C.19. Tasso di mancata risposta parziale: percentuale di dati mancanti e dovuti sul totale dei dati dovuti (per le principali variabili di interesse)

Indicatori sul carico statistico

C.20. Tempo medio di compilazione del questionario (che può includere o meno il tempo di reperimento dell'informazione)
C.21. Numero medio di quesiti cui l'unità deve rispondere
C.22. Quesito che più frequentemente segna l'abbandono al questionario

Mappatura con i sotto-processi GSBPM

2.3., 3.1., 3.2, 3.5., 4.2., 4.3., 8.1.

Riferimenti bibliografici

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

Brancato et el. Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System

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

Eurostat (2014) ESS Guidelines for the implementation of the ESS quality and performance indicators http://ec.europa.eu/eurostat/documents/64157/4373903/02-ESS-Quality-and-performance-Indicators-2014.pdf/5c996003-b770-4a7c-9c2f-bf733e6b1f31

FCSM (2001) "Measuring and Reporting Sources of Error in Surveys". Federal Committee on Statistical Methodology, Statistical Policy Working Paper 31. http://www.fcsm.gov/01papers/SPWP31_final.pdf

Hidiroglou MA, Drew DJ, Gray GB (1993) "A Framework for Measuring and Reducing Nonresponse in Surveys". Survey Methodology, 19, 1, pp. 81-94Wallgren 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

D. Conversione in formato elettronico (registrazione)

Descrizione

Questo sotto-processo consiste nella registrazione su supporto informatizzato dei dati rilevati mediante tecniche non assistite da computer, mentre per le tecniche assistite da computer essa avviene contestualmente alla raccolta dei dati.

La conversione dei dati in formato elettronico può essere automatizzata o comportare l'impiego di personale che inserisce manualmente i dati raccolti. Nel caso di lettura ottica, non sempre è possibile convertire in formato elettronico la totalità delle informazioni in modo automatico e può essere richiesto un intervento manuale. Un'attività implicita nella registrazione dei dati è la codifica, ossia il processo mediante il quale viene assegnato un valore numerico a ciascuna risposta, sulla base di un sistema predefinito in fase di progettazione. Anche questa operazione è spesso automatizzata, tuttavia le decisioni più complesse possono richiedere l'intervento umano. Nel caso in cui le informazioni vengano acquisite in formato aperto e codificate a posteriori attraverso una fase di vera e propria codifica, i principi e i relativi suggerimenti sono sviluppati nella Sezione F.

Per prevenire gli errori nel processo di registrazione (o di acquisizione dei dati nel caso di utilizzo di questionari elettronici) si inseriscono dei controlli, che possono essere: di dominio su specifiche variabili (per esempio l'età all'interno di un certo *range*), di coerenza tra variabili (per es. tra età e condizione professionale), di flusso (per es. quando ci sono sottosezioni di un questionario alle quali si risponde solo in base a domande filtro). I controlli possono essere vincolanti per il proseguimento dell'intervista o della registrazione (controlli hard) o non vincolanti (controlli soft).

Principio D.1. Conversione in formato elettronico (registrazione)
La procedura di registrazione, sia essa manuale o mediante lettura ottica, deve garantire un elevato livello di qualità delle informazioni registrate. Misure oggettive della qualità della registrazione andrebbero prodotte e valutate.

Suggerimenti

Registrazione dei dati contestuale alla rilevazione dei dati

- Se la registrazione su supporto informatico avviene durante la raccolta dei dati (tecniche assistite da computer), il questionario elettronico dovrebbe essere progettato in modo da massimizzare l'accuratezza delle informazioni registrate limitando il più possibile i tempi e il carico di chi (intervistato o intervistatore) digita le informazioni.
- Il numero di controlli nel questionario elettronico dovrebbe essere bilanciato: non eccessivo per evitare di interrompere troppo spesso il flusso dell'intervista, ma sufficiente a garantire la qualità delle informazioni più importanti raccolte.
- I controlli del questionario elettronico devono essere personalizzati rispetto al tipo di informazioni rilevate. Nei quesiti relativi a informazioni oggettive (per es. anno di nascita) possono essere utilizzati controlli di tipo "hard", mentre per i quesiti relativi ad attitudini e conoscenze è preferibile usare i controlli di tipo "soft". I controlli di tipo "hard" possono essere utilizzati anche nel caso di errori di flusso, ossia in corrispondenza a domande filtro cui seguono domande che devono /non devono essere somministrate.

- Nei questionari web autosomministrati, andrebbero evitati in controlli di tipo hard.
- Se la registrazione su supporto informatico avviene durante la raccolta dei dati (tecniche assistite da computer), prevedere nella formazione dei rilevatori anche l'argomento relativo alla digitazione delle informazioni.

Registrazione dei dati successiva alla rilevazione dei dati

- Se la registrazione dei dati su supporto informatico è svolta successivamente alla raccolta dati mediante operatori è opportuno provvedere ad una adeguata formazione degli operatori e dotarli del materiale di supporto (per es. manuali con regole).
- Il software adottato per la registrazione dovrebbe prevedere una serie di controlli al fine di minimizzare l'errore di digitazione: controlli vincolanti sui codici identificativi e preferibilmente controlli non vincolanti (di dominio, di flusso e di coerenza) sugli altri dati. I controlli, tuttavia, non devono essere eccessivi, per evitare troppo frequenti interruzioni dell'attività di registrazione. Obbligare gli operatori a correggere solo i propri errori di digitazione impedendo di correggere qualsiasi incongruenza causata dalle risposte dell'intervistato.
- In caso di adozione della lettura ottica, il questionario dovrebbe essere disegnato in modo da facilitare il riconoscimento automatico dei caratteri.
- In caso di adozione della lettura ottica, si deve prevedere che una parte dei dati non sia acquisita automaticamente, perché alcuni caratteri possono non essere riconosciuti dal software, o perché alcuni questionari possono pervenire in cattive condizioni. Di conseguenza la lettura ottica deve essere affiancata dalla registrazione da parte di operatori.
- Nel caso di registrazione da parte di una società esterna all'ente, devono essere previsti il livello massimo di errore accettabile e le procedure di verifica della qualità dei lotti registrati. I dati registrati devono essere inviati utilizzando un protocollo di trasmissione sicura.
- La qualità della registrazione va considerata sia in relazione all'accuratezza (minimizzazione dell'incidenza degli errori di registrazione), sia in relazione al tempo impiegato per tale fase, che non deve essere tale da provocare elevati ritardi nel rilascio dei dati.
- Le valutazioni effettuate sull'accuratezza della registrazione e sul tempo richiesto possono essere utilizzate per migliorare il processo produttivo nelle successive edizioni dell'indagine.

Indicatori di qualità e performance

Il controllo di qualità della registrazione prevede la selezione di un campione casuale di record da lotti di materiale registrato e la ri-digitazione con verifica delle eventuali discrepanze sul materiale cartaceo. Se l'errore identificato supera una certa soglia, in genere viene ri-digitato tutto il lotto. Tale procedura è spesso implicita nei contratti per la digitazione dei dati in outsourcing. Nel caso di applicazione della procedura descritta, un indicatore calcolabile è:

D.1. Percentuale di errori di digitazione identificati dalla procedura di controllo

Nel caso di adozione di lettura ottica un indicatore di performance della procedura può essere calcolato come:

D.2. Percentuale di questionari acquisiti in lettura ottica sul numero di questionari previsti per tale modalità di registrazione.

Mappatura con i sotto-processi del GSBPM

2.3, 3.1, 3.2, 3.5, 4.4.

Riferimenti bibliografici

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

Statistics Canada (2010) *Survey Methods and Practices.* Statistics Canada, Catalogue no. 12-587-X, Ottawa. http://www.statcan.gc.ca/pub/12-587-x/12-587-x2003001-eng.htm

E. Integrazione

Descrizione

La procedura di integrazione dei dati consiste nell'utilizzo congiunto di informazioni provenienti da processi produttivi diversi relativi ad una stessa area informativa. L'integrazione assicura pertanto la possibilità di mettere in relazione le informazioni provenienti da fonti diverse, una volta definite le unità da osservare.

L'integrazione tra due fonti di dati può essere di tipo micro e di tipo macro, nel primo caso l'obiettivo è quello di rintracciare i record delle due fonti che si riferiscono alla stessa unità, nel secondo caso si possono ricostruire dei parametri relativi a variabili osservate nelle due fonti.

L'integrazione tra più fonti di dati può avere diverse finalità. Nel caso di archivi amministrativi, l'integrazione è condotta per colmare problemi di copertura di un archivio, per rendere disponibili nuove variabili non presenti nell'archivio di riferimento, o per imputare valori mancanti. Nel caso di integrazione di dati di indagine con dati amministrativi può essere condotta per colmare problemi di mancata risposta (totale o parziale), per rendere disponibili nuove variabili, o per condurre analisi di *record check* volte ad individuare e valutare l'impatto di eventuali errori di misura.

L'integrazione tra fonti amministrative oppure tra fonti amministrative e dati d'indagine può avvenire in diversi modi. Se le unità presentano un codice identificativo univoco e privo di errori, allora si può procedere con un abbinamento esatto (*merging*) basato su tale codice. Quando un unico codice identificativo non c'è, si utilizzano i metodi di *record linkage* (aggancio dei record), attraverso i quali è possibile abbinare le unità, se esistono delle variabili chiave che congiuntamente considerate contribuiscono a identificare l'unità (quali nome, cognome, data, di nascita, indirizzo, …).

Le procedure di *record linkage,* sia se l'integrazione è tra fonti amministrative, o tra fonti amministrative e dati di indagine, sono caratterizzate da una serie di fasi:

- la pre-elaborazione, che risulta importante, ma spesso prescinde dall'obiettivo dell'integrazione di fonti diverse e riguarda la necessità di rendere compatibili e omogenee le informazioni contenute; questa fase comprende la scelta delle variabili di abbinamento, il miglioramento della qualità dei dati nelle fonti da integrare, in alcuni casi la standardizzazione delle variabili ed eventuali operazioni di suddivisione e ordinamento dei record delle basi di dati da integrare;
- l'applicazione di un metodo di *record linkage*, che può essere deterministico quando si fa riferimento a regole formali per stabilire se coppie di record nelle due fonti distinte fanno riferimento alla stessa unità o probabilistico quando la regola di decisione è basata su modelli probabilistici;
- l'analisi statistica dei dati abbinati sulla base delle informazioni provenienti dall'applicazione del metodo di *record linkage*, al fine di valutare la robustezza dei risultati.

Relativamente all'applicazione di un metodo di *record linkage*, per ridurre la complessità statistica e computazionale, spesso si ricorre a variabili di *blocking*, variabili categoriali affidabili attraverso le quali vengono partizionati i data set da integrare, per poi procedere al confronto tra i record appartenenti alla stessa partizione. Infine, la decisione se una coppia di record sia o meno relativa alla stessa unità avviene attraverso una funzione di confronto sulle variabili di abbinamento. La funzione più usata è quella che verifica l'uguaglianza (attribuendo valore 1) o la diversità (attribuendo valore 0) del valore della variabile di abbinamento nei due record messi a confronto.

Spesso le procedure di integrazione consistono in una combinazione di diversi metodi di abbinamento.

La procedura di integrazione può essere soggetta ad errori che è necessario valutare. Questi possono essere dovuti alla qualità delle variabili chiave e quindi pregiudicare la possibilità di aggancio dei record di due basi di dati. Tra gli errori di questo tipo si riscontrano quelli di trascrizione, ad esempio un individuo può immettere una data di nascita sbagliata, e quelli di registrazione. Questi errori avvengono durante la fase di registrazione dei dati. Errori in questa fase sono in qualche modo controllabili da parte dell'ente che produce i dati, ma difficilmente si riesce ad eliminarli prima della procedura.

Gli errori nelle variabili chiave riducono l'efficacia dell'informazione congiunta delle variabili per l'aggancio delle unità di due basi di dati, generando due tipi di errore di *linkage*:

- i falsi abbinamenti, alcuni record possono essere abbinati anche se in realtà fanno riferimento a unità diverse;
- i falsi non abbinamenti, alcuni record delle due basi di dati fanno riferimento alla stessa unità ma nell'abbinamento non si è in grado di individuarli a causa degli errori nelle variabili chiave.

Gli errori nell'integrazione tra dati possono causare altri errori nelle fasi successive con un conseguente impatto sull'accuratezza dei dati stessi.

Principio E.1. Integrazione tra fonti di dati
L'integrazione tra fonti di dati 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. La validità dei risultati del processo di integrazione deve essere valutata, se possibile, calcolando opportuni indicatori.

Suggerimenti

Le procedure di integrazione possono consistere in una combinazione di diversi metodi. A tale proposito è necessario che l'intera procedura sia ben definita, stabilendo accuratamente l'ordine con cui applicare le diverse metodologie, nei diversi campi di applicazione e in relazione agli obiettivi prefissati.

- I metodi utilizzati nell'integrazione devono essere condivisi e consolidati a livello internazionale.
- Se per l'integrazione di diverse fonti di dati è utilizzata una procedura informatica ad hoc, la procedura deve essere preventivamente testata per evitare che errori di programmazione possano inficiare l'accuratezza dei risultati del processo di integrazione.
- Il processo di integrazione deve avvenire nel rispetto delle normative per la tutela della riservatezza.
- È importante migliorare la qualità dei dati nelle basi di dati da integrare**:** per quanto possibile, è utile fare in modo che le basi di dati a disposizione siano estremamente accurate per evitare poi errori negli abbinamenti.
- Nel caso si ritenga opportuno utilizzare i metodi di *record linkage*, è necessario scegliere le variabili chiave. La scelta delle variabili chiave è estremamente delicata. In linea di principio, tutte le variabili in comune fra le due basi di dati possono essere usate congiuntamente per identificare le unità, ma molte di queste non sono necessarie per l'integrazione. In genere si sceglie il numero minimo di variabili chiave che congiuntamente identificano le unità, fra le variabili in comune nelle due basi di dati che sono universali (ovvero tutte le unità devono rispondere a queste variabili) e permanenti (ovvero immodificabili nel tempo).
- È opportuno selezionare le variabili chiave fra le variabili più accurate, complete e non sensibili, ovvero che non violino il diritto alla riservatezza delle unità.
- È necessario anche procedere alla standardizzazione delle variabili**:** può risultare utile trasformare in modo opportuno le modalità delle variabili chiave in modo da rendere più semplice per i computer il

riconoscimento delle differenze. Questo avviene in particolare per variabili come "nome", "cognome" e "indirizzo". Per queste variabili spesso si preferisce eliminare i titoli (come *sig.*, *dr.*, per gli individui, *srl*, *spa* per le imprese, *via*, *piazza* per gli indirizzi). In alcuni casi le modalità di queste variabili vengono trasformate in modo da limitare gli effetti derivanti da errori di digitazione o possibili differenze nella pronuncia di nomi stranieri.

- Inoltre, per facilitare il controllo dei record da parte dei programmi software per il *record linkage*, può essere necessario ordinare (*sorting*) opportunamente i record nelle due basi di dati e dividerli in gruppi (*blocking*). Quest'ultima operazione può influenzare in modo notevole i risultati del *record linkage*.
- Tutte le operazioni di trattamento condotte sulle singole fonti di dati ai fini dell'integrazione, nonché le metodologie di integrazione stesse devono essere documentate.

A seguire dei passi indicati si procede con l'applicazione del metodo di *record linkage,* deterministico o probabilistico. 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.

Indicatori di qualità e performance

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*. Pertanto, laddove possibile, si deve produrre una stima dei seguenti due indicatori:

E.1. Tasso di falsi abbinamenti (record erroneamente abbinati che nella realtà rappresentano due distinte unità)
E.2. Tasso di falsi mancati abbinamenti (unità erroneamente non abbinate dalla procedura).

La valutazione di questi errori è particolarmente costosa e spesso di difficile applicabilità perché richiede o l'esistenza di fonti di confronto più accurate o un controllo manuale.

Mappatura con i sotto-processi GSBPM

2.5., 3.5., 5.1., 8.1.

Riferimenti bibliografici

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.

Essnet Data Integration Cros portal. https://ec.europa.eu/eurostat/cros/content/data-integration_en

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

Scanu M. (2003) Metodi Statistici per il record linkage, Metodi e Norme-n.16, Istat

F. Codifica e classificazioni

Descrizione

La codifica è la trasformazione di valori non numerici in dati numerici o, più in generale, in categorie predefinite più facili da trattare dal punto di vista statistico. Spesso questa operazione è piuttosto semplice in quanto le modalità di risposta di un questionario cartaceo, e i relativi codici numerici sono predefiniti (per es. Maschio=1, Femmina=2), cioè basati su una classificazione (o nomenclatura) data, e questi semplici valori vengono riportati nel file di dati finali. Altre volte le modalità di risposta sono espresse in termini di testo libero, cui assegnare i codici numerici, anche qui sulla base di un sistema o di uno schema classificatorio. Ne è un esempio la codifica dell'attività economica svolta da un individuo in una indagine sull'occupazione.

Nei casi più semplici, e quando nella raccolta dei dati si utilizzano tecniche che prevedono l'assistenza di computer, è anche possibile che la codifica venga effettuata dall'intervistatore in modo contestuale alla raccolta dei dati. Nel caso di acquisizione di dati da fonti amministrative la codifica deve permettere di ricondurre le classificazioni amministrative a quelle statistiche ed è un processo che si svolge dopo la fase di acquisizione dei dati.

L'attività di codifica viene definita:

- automatica, quando viene utilizzata una applicazione software in modalità *batch* che attribuisce automaticamente codici a variabili rilevate a testo libero;
- assistita, quando la codifica viene effettuata dal rispondente, dall'intervistatore o dal codificatore, a seconda della tecnica, con l'ausilio di una specifica applicazione software;
- manuale, quando viene effettuata da personale appositamente istruito senza l'ausilio di una applicazione software dedicata.

Nella maggior parte delle applicazioni, i metodi di codifica coesistono, in quanto una parte del materiale viene codificato in modo automatico e quello che i sistemi automatici non riescono a risolvere viene trattato da operatori, o in modo assistito da computer oppure in modo manuale. Nei primi due casi l'attività viene svolta sulla base di dizionari informatizzati, ossia elenchi di riferimento che vengono aggiornati periodicamente.

L'attività di codifica è strettamente legata all'adozione di classificazioni alle cui voci sono ricondotte le modalità di risposta. La scelta di una classificazione è cruciale, in quanto classificazioni differenti possono mettere in luce aspetti diversi del fenomeno che si intende indagare e avere quindi un impatto sulla pertinenza delle statistiche prodotte.

Nel caso di utilizzo di dati amministrativi, gli errori di classificazione, ossia gli errori che si commettono nell'allineare le classificazioni delle variabili incluse nell'archivio amministrativo con quelle relative all'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 (cfr. Sezione H).

Gli errori che si generano durante l'attività di codifica hanno impatto sull'accuratezza dei dati. Infatti, errate specificazioni nello schema predefinito per la codifica e nel software automatico possono portare ad errori di misura prevalentemente di natura sistematica, e quindi risultare in distorsione delle stime, mentre da parte degli operatori potrebbero essere introdotti errori sia di natura casuale che sistematica. Tuttavia, la qualità della codifica dipende anche dalla completezza e qualità della risposta fornita dal rispondente, caratteristica

questa più difficile da controllare, ma che richiede un'attenta progettazione e realizzazione del questionario e dei quesiti in esso contenuti.

Principio F.1. Codifica e classificazioni
La procedura di codifica deve utilizzare classificazioni standard o ben strutturate e garantire un elevato livello di qualità delle informazioni codificate, minimizzando costi e tempi. Le classificazioni adottate devono rispecchiare il fenomeno oggetto di studio ed essere tali da evitare ambiguità nella codifica. Quando disponibile, deve essere utilizzato software generalizzato. Misure oggettive della qualità della codifica andrebbero prodotte e valutate.

Suggerimenti

Schema

- Essere riconosciuto come standard a livello nazionale o internazionale (es. la classificazione delle malattie ICD9CM del Ministero della salute oppure l'*International Standard Classification of Occupations* (ISCO) dell'*International Labor Organisation*),
- Avere codici esaustivi rispetto alle modalità di risposta e mutuamente esclusivi. Possono essere utilizzate aggregazioni o disaggregazioni di codici purché rimanga possibile riportarsi alla classificazione nota.
- Essere aggiornato e adeguato al fenomeno oggetto di indagine.
- Prevedere la codifica di qualsiasi modalità di risposta, anche quelle non informative, come per esempio il "non so/non risponde", il "rifiuto a rispondere" o il "non applicabile" per quesiti derivanti da domande filtro, cui devono essere assegnati un valore convenzionale e uguale per tutte le classificazioni.

Procedura di codifica

- Essere applicata nel modo più coerente possibile su tutte le unità oggetto di studio;
- Prevedere, nel caso di risposte chiuse, una pre-codifica delle modalità di risposta durante la fase di raccolta dati, allo scopo di ridurre i costi e i tempi;
- Prevedere la possibilità di distinguere il valore ammissibile nullo dalle mancate risposte e dalle diverse modalità non informative;
- Prevedere, in caso di risposte aperte e quando effettuata dopo l'acquisizione delle risposte, l'ausilio di un apposito software per ridurre i costi e migliorare l'accuratezza rispetto alla codifica manuale;
- Assicurare l'accesso ai dati testuali (non codificati), nel caso in cui i dati amministrativi per variabili testuali utilizzino una classificazione che non è riconducibile a quella obiettivo, in quanto ciò può consentire la piena riconducibilità alla classificazione di interesse;
- Essere assegnata, nel caso di codifica manuale, a personale appositamente formato e, se possibile, esperto della specifica classificazione impiegata e della materia amministrativa nel caso di uso di tale tipo di dati; gli operatori dovrebbero essere dotati di tutti gli strumenti necessari (manuali, istruzioni,…);
- Prevedere, per i casi più complessi, l'analisi da parte di un gruppo di esperti della classificazione adottata. È possibile, in caso di difficoltà nell'assegnazione di particolari codici, operare uno smistamento preliminare su livelli più alti, per poi demandare l'individuazione del codice specifico a un operatore più esperto. Una procedura centralizzata, oltre a ridurre i costi, favorisce la diffusione dell'esperienza tra gli operatori;

- Prevedere il monitoraggio dell'operato dei codificatori e, in situazioni più critiche, la valutazione del loro operato anche attraverso metodi per il controllo statistico della qualità. È possibile, infatti, che operatori differenti prendano decisioni differenti su come codificare una stessa risposta per effetto della loro diversa esperienza, formazione o inclinazione personale, specie se il processo non è assistito da software. Ciò è particolarmente delicato nei casi in cui le risposte sono a modalità aperta;
- Prevedere, in caso di codifica automatica di risposte aperte, la creazione di un file di riferimento (dizionario) dove raccogliere le frasi riconosciute dal software. Il suo contenuto può essere migliorato attraverso appositi studi a campione, effettuati dai codificatori più esperti, sull'accuratezza della procedura di codifica automatica. In tal modo i dizionari possono essere arricchiti nel tempo, sfruttando l'esperienza dei processi statistici precedenti.
- Se disponibile, utilizzare software generalizzato.
- Tutte le procedure relative alla codifica devono essere esaustivamente documentate.

Indicatori di qualità e performance

Per valutare la qualità della codifica è necessario ripetere l'operazione di codifica di uno stesso lotto di valori almeno due volte e/o avere a disposizione i valori codificati corretti. Ciò, tuttavia, comporta naturalmente un investimento non indifferente in termini di tempo e risorse economiche. A costi limitati, è quindi solo possibile calcolare il seguente indicatore di performance dell'efficienza del sistema di codifica.

F.1. Percentuale di valori che sono codificati rispetto al totale dei valori sottoposti a codifica per tipo di codifica (automatica, assistita da computer, manuale).

Mappatura con i sotto-processi GSBPM

2.5, 3.5., 5.2., 8.1.

Riferimenti bibliografici

Groves R.M., Floyd J.F. Jr, Couper M.P., Lepkowski J.M., Singer E., Tourangeau R.(2004) Survey Methodology. Wiley series in methodology

Istat (2007). Metodi e software per la codifica automatica e assistita dei dati. Tecniche e strumenti, n. 4, 2007

Grant EL, Leavenworth RS. (1996) Statistical quality control. 7th edition, New York, McGraw-Hill.

G. Identificazione e trattamento degli errori

Descrizione

Questa fase, anche comunemente identificata come controllo e correzione, consiste nell'applicazione di una varietà di metodi che hanno l'obiettivo di migliorare la qualità dei dati. Infatti, i dati rilevati attraverso l'indagine statistico o da fonte amministrativa possono presentare: *i)* errori dovuti a una qualunque delle fasi di acquisizione e messa a punto delle informazioni (raccolta, codifica, registrazione), chiamati errori di misura; *ii)* dati mancanti per alcune variabili (mancate risposte parziali). Questi rientrano nella tipologia degli errori non campionari.

Al fine di identificare gli errori e le mancate risposte parziali presenti nei dati è possibile implementare dei controlli (o regole o *edit*), che comprendono:

- controlli di consistenza, che verificano se prefissate combinazioni di valori assunti da variabili rilevate in una stessa unità soddisfano certi requisiti (regole di incompatibilità);
- controlli di validità o di *range*, che verificano se i valori assunti da una data variabile sono interni all'intervallo di definizione della variabile stessa;
- controlli statistici, utilizzati al fine di isolare quelle unità statistiche che presentano, per alcune delle variabili in esse contenute, valori che si discostano in modo significativo dai valori che le stesse variabili assumono nel resto delle unità campionarie o rispetto ad una rilevazione precedente (per es. tecniche per l'analisi degli *outlier*). Questi valori sono con alta probabilità errati, ma l'asserzione della loro non correttezza necessita di ulteriori e approfondite verifiche.

A seguito della localizzazione di errori e mancate risposte parziali si può procedere a: controllo sul questionario cartaceo e/o ricontatto dell'unità (se possibile) e correzione manuale; cancellazione dell'intera informazione riguardante l'unità; imputazione.

L'imputazione è il processo di assegnazione di valori coerenti al posto di dati mancanti, inammissibili o incoerenti che hanno violato le regole di controllo.

Il principio alla base di un processo di imputazione è quello di utilizzare informazioni ausiliarie disponibili così da approssimare il più accuratamente possibile i valori mancanti o errati, attraverso un modello di assegnazione (cioè la formulazione di un insieme di ipotesi sulle variabili che richiedono imputazione) e produrre stime di qualità. Tale principio dovrebbe comportare una riduzione della distorsione e della varianza generati dal non aver osservato tutti i valori desiderati. A seconda delle variabili ausiliarie a disposizione è possibile scegliere diversi metodi di imputazione.

In generale i metodi di imputazione possono essere raggruppati in due categorie: deterministici e stocastici (o probabilistici). I primi sono metodi che, a seguito di applicazioni ripetute, producono gli stessi risultati. I secondi sono caratterizzati da una certa variabilità dei risultati. Tra i metodi deterministici vi sono: l'imputazione deduttiva, da serie storica, con il valore medio, da modello di regressione senza componente stocastica e l'imputazione con donatore di distanza minima. I metodi stocastici includono l'imputazione da donatore di tipo casuale e di distanza minima con selezione casuale del donatore da un insieme di unità candidate, da modello di regressione con componente casuale e altri metodi deterministici a cui vengono aggiunti residui casuali.

Sia la localizzazione degli errori che l'imputazione possono seguire diversi approcci: interattivo, se l'individuazione e la correzione delle incompatibilità è basata sull'interazione tra esperto e dati, per cui il processo di verifica e correzione dipende strettamente da decisioni umane prese caso per caso; automatico se

le procedure per l'individuazione e la correzione sono interamente automatizzate e affidate ad un software (ad hoc o generalizzato); misto derivante dalla combinazione delle due componenti quella automatica e quella interattiva.

Le procedure di identificazione e trattamento possono inoltre riguardare tutte le osservazioni in errore (solitamente con una procedura automatica) o solo le unità errate con impatto significativo sulle stime finali (editing selettivo), o, ancora, utilizzare una procedura interattiva per le unità influenti e automatica per le rimanenti (procedura di tipo misto).

Fra i principali problemi connessi all'attività di verifica dei dati vanno però considerati i costi ed i tempi necessari per mettere a punto gli strumenti idonei e per portare a termine le operazioni connesse al controllo e correzione dei dati. Questo spinge a implementare procedure di controllo già in fase di *acquisizione* presso le unità, in modo da rendere più agevole il reperimento di informazioni corrette laddove si verifichino situazioni non compatibili o anomale, a sviluppare tecnologie per l'integrazione del controllo e correzione dei dati con le fasi di intervista o di registrazione, così da eliminare o in ogni caso minimizzare la parte di errori attribuibile ad errori di compilazione o registrazione dei modelli (che rappresentano generalmente la parte più consistente del totale degli errori). Alcune tipologie di errori (di codifica, di percorso, di dominio, ecc.) vengono corretti contestualmente alla fase di intervista o di registrazione, producendo una migliore qualità finale dei dati ed un risparmio nei tempi e nei costi connessi alle fasi successive di controllo (interattivo o automatico) dei dati.

Le informazioni derivate dalla procedura di controllo e correzione come, ad esempio, la frequenza di attivazione delle regole di controllo o il tasso di imputazione per variabile, rappresentano dei campanelli d'allarme di possibili problemi nel processo produttivo (ad esempio attribuibili a difetti del questionario) e possono fornire un'idea sulle principali fonti di errore. Tali preziose informazioni devono essere analizzate e utilizzate per attivare un processo virtuoso volto al miglioramento delle successive edizioni dell'indagine.

Mancate risposte parziali ed errori di misura possono compromettere seriamente l'accuratezza delle stime di interesse, aumentando la variabilità e introducendo possibili distorsioni delle stime. I metodi di controllo e correzione hanno proprio l'obiettivo di pervenire ad un insieme di dati completo e corretto. Tuttavia, se non opportunamente applicate, le procedure stesse possono essere fonte di ulteriori errori nei dati.

Principio G.1. Identificazione e trattamento degli errori
Le procedure di identificazione e trattamento degli errori devono essere scelte in relazione alle diverse tipologie di errore e ai dati, alle caratteristiche del processo, nonché ai vincoli di tempo e risorse. Esse devono essere basate su metodologie statistiche consolidate e devono essere opportunamente valutate e documentate.

Suggerimenti

Strategia

- L'insieme dei dati provenienti dalla fase di raccolta diretta o acquisizione da fonti amministrative e convertiti in formato elettronico deve essere verificato rispetto alla completezza e alla coerenza delle informazioni. Tale verifica deve avere come obiettivo quello di massimizzare i livelli di qualità e quindi la scelta di correggere i dati dovrebbe essere presa soltanto se si giudica che gli errori individuabili siano tali da rendere troppo bassa la qualità dell'informazione rispetto ai livelli

prestabiliti e se si pensa che l'insieme delle informazioni ausiliarie che si possiedono permettono di correggere i dati e di migliorarne la qualità.

- L'approccio da seguire (interattivo ed automatico) deve essere valutato in base ad opportuni e rigorosi criteri che portino alla scelta di metodologie ottimali in termini sia di efficacia (qualità dei risultati) che di efficienza (tempi, costi, disponibilità di personale con elevata esperienza, carico sui rispondenti).
- Organizzare le procedure di controllo e correzione per priorità, concentrando le risorse sul trattamento degli errori più gravi e delle unità e variabili più importanti, anche attraverso la verifica sul modello di rilevazione compilato o il ricontatto del rispondente. In ogni caso, la revisione interattiva andrebbe limitata agli errori più rilevanti e che non possono essere risolti in modo automatico, così da permettere una riduzione dei costi, un miglioramento della tempestività, limitare il numero e l'onere (fastidio) di risposta per le unità ricontattate.
- La procedura di controllo e correzione da adottare deve essere determinata da più elementi, principalmente dal tipo di errore, ma anche dalle caratteristiche dei dati e in particolare dal tipo di variabili esaminate (qualitative o quantitative), dalla presenza di valori anomali all'interno della distribuzione, dalla disponibilità di serie storiche per il confronto dei dati con valori pregressi, dalla numerosità dei dati stessi (bassa, media, alta).
- Differenziare i metodi in base alle ipotesi sulla tipologia di errore: trattare gli errori sistematici con regole deterministiche, basate sulla conoscenza del meccanismo che può aver generato l'errore; trattare i presunti errori casuali con un metodo stocastico, che preserva meglio la struttura della frequenza dei dati e offre una variabilità più realistica dei dati imputati.
- Nei casi di archivi/fonti integrati scegliere la strategia più idonea, valutando i vantaggi connessi all'implementazione di procedure di controllo e correzione solo sull'archivio integrato rispetto a quelle applicate singolarmente sugli archivi prima dell'integrazione. Il livello di qualità attesa nei dati finali e le risorse effettivamente disponibili per ottenere tale livello (tempi, costi, complessità delle procedure da applicare) deve guidare nella scelta dello scenario da preferire.
- Testare la procedura di controllo e correzione prima della sua applicazione ai dati reali.
- Se disponibile, utilizzare software generalizzato che implementa metodologie note.
- Tutte le procedure di controllo e correzione devono essere esaustivamente documentate.

Individuazione degli errori

- Le regole di controllo devono essere il risultato di una collaborazione tra esperti della materia oggetto di rilevazione, personale dell'indagine, esperti di dati amministrativi ed esperti nelle metodologie di controllo e correzione.
- È generalmente indicato utilizzare controlli o regole di *edit* di tipo logico nel caso di variabili qualitative o di tipo statistico nel caso di variabili quantitative.
- Verificare che le regole siano coerenti e non ridondanti e tali da evitare una eccessiva correzione dei dati (*over-editing*) non giustificata da un apprezzabile miglioramento della qualità dei risultati.
- I dati mancanti (*missing value*) devono essere riconoscibili rispetto ai valori non dovuti e, nel caso di variabili quantitative, anche rispetto agli zeri strutturali.
- Per gli errori di natura sistematica, la definizione delle regole deterministiche atte alla loro identificazione dovrebbe scaturire dall'analisi degli indicatori relativi alle regole di controllo. Gli errori sistematici devono essere identificati e corretti prima degli errori casuali e dell'editing selettivo.

- L’individuazione di errori influenti deve seguire un approccio basato sull’editing selettivo, le cui priorità devono possibilmente riflettere una funzione punteggio che valuti il rischio di errore e l’influenza sulla stima.
- Per il riconoscimento di valori anomali devono essere utilizzati metodi robusti che vanno da semplici analisi univariate a metodi grafici complessi, in base alle relazioni esistenti tra le variabili nelle diverse sottopopolazioni. In ogni caso, la plausibilità di un valore anomalo deve essere attentamente valutata prima di sottoporlo al processo di correzione.
- Per gli errori casuali dovrebbe essere utilizzata una metodologia consolidata basata sul principio di minimo cambiamento (ad esempio il paradigma di Fellegi-Holt).

Imputazione

- L'imputazione di dati deve avvenire sulla base di metodologie e tecniche che diano prefissate garanzie di qualità e di efficienza, come il mantenimento delle distribuzioni originali dei dati, l'oggettività e la riproducibilità. In ogni caso, è sempre necessario tenere traccia di quali e quante imputazioni sono state effettuate.
- Qualsiasi metodo di imputazione equivale ad assumere, implicitamente o esplicitamente, un modello basato su informazioni ausiliarie. La selezione delle variabili ausiliarie deve essere effettuata tenendo conto della forza dell’associazione con le variabili da imputare e quanto esse contribuiscono a spiegare il meccanismo della mancata risposta. Il modello di imputazione, che incorpora le variabili ausiliarie, deve essere attentamente validato per ogni variabile soggetta a imputazione separatamente, e per gruppi di variabili.
- Nella scelta del donatore considerare che uno specifico donatore dovrebbe essere utilizzato per un numero limitato di riceventi, mentre per uno specifico ricevente bisognerebbe limitare il numero di donatori diversi.
- Preferire un metodo deduttivo se si ha la possibilità di sfruttare le informazioni presenti per poter dedurre il valore da sostituire al dato mancante da una o più variabili ausiliarie, o nel caso in cui le informazioni disponibili conducano ad un solo valore ammissibile o, ancora, quando la natura dell’errore sia ben nota. Questo metodo è indicato anche per i dati di fonte amministrativa.

Revisori

- Curare la formazione dei revisori e fornire adeguate istruzioni, in forma scritta, sulle regole da seguire per l’applicazione dei controlli e per il trattamento dei diversi casi di errore possibili. Le istruzioni dovrebbero essere sviluppate, testate, e poi revisionate periodicamente, e la loro applicazione dovrebbe essere monitorata, anche per evitare il fenomeno dell’editing “creativo”, ovvero della presenza di regole soggettive nella correzione.
- Predisporre un sistema di supporto e supervisione dei revisori e di valutazione. In presenza di un possibile effetto dei revisori sulle stime, si consiglia di effettuare una valutazione anche attraverso sperimentazioni (ad esempio analisi di qualità confrontando i dati grezzi con quelli finali validati per ottenere una misura dell’attività dei revisori sia in termini di entità, sia in termini di tipologia degli interventi di correzione effettuati).

Valutazione

- Per garantire la valutabilità di eventuali sotto-fasi del piano di controllo e correzione, è necessario conservare i valori originali e quelli imputati nei diversi stadi della procedura.

- Il processo di controllo e correzione deve essere attentamente monitorato attraverso indicatori (es. tassi di mancata risposta parziale per variabile, tassi di attivazione delle regole di compatibilità, tassi di imputazione, confronti tra distribuzioni prima e dopo la procedura, differenze fra le stime prodotte calcolate sui dati grezzi e su quelli puliti).
- Valutare l'opportunità di stimare la variabilità aggiuntiva attribuibile all'imputazione nel caso si applichino tecniche di controllo e correzione.

Indicatori di qualità e performance

Indicatori di qualità della fase di trattamento dei dati possono essere:

a livello micro informazioni su:

G.1. Numero di *edit* attivati per singole variabili;
G.2. Numero e tipo di variabili responsabili dell'attivazione degli edit;
G.3. Numero di inconsistenze, errori di dominio e di codifica;
G.4. Numero di *outlier* individuati il tipo di trattamento cui sono stati sottoposti.

a livello macro:

G.5. Tasso di mancata risposta parziale per variabile, come risulta al termine delle fasi di *editing* e di imputazione logico-deduttiva;
G.6. Tasso di imputazione delle singole variabili al termine della fase di imputazione;
G.7. Numero totale di *edit* attivati;
G.8. Tasso di imputazione totale per il *dataset* (riferito alle sole variabili soggette ad imputazione)
G.9. Tassi di imputazione distinti per tipo di imputazione: modifica da valore *non blank* ad altro valore non *blank*, imputazione netta da *blank* a valore *non blank*, cancellazione da valore *non blank* a *blank*;
G.10. Tasso di non imputazione totale riferito al *dataset*, come misura della qualità di base dei dati.

Mappatura con i sotto-processi GSBPM

2.5, 3.5., 5.3, 5.4., 8.1.

Riferimenti bibliografici

EUROSTAT (2014). Handbook on Methodology of Modern Business Statistics "Statistical Data Editing", http://ec.europa.eu/eurostat/cros/content/statistical-data-editing

Luzi O. e Grande E. (2003) "Metodologie per l'imputazione delle mancate risposte parziali: analisi critica e soluzioni disponibili in Istat" ISTAT, Servizio delle Metodologie di Base per la Produzione Statistica

G.Barcaroli, L.D'Aurizio, O.Luzi, A.Manzari, A.Pallara "Metodi e software per il controllo e la correzione dei dati" Documenti Istat, n. 1/1999

Fellegi I.P.,Holt D. A systematic approach to automatic Edit and imputation. Journal of American Statistical Association, Vol.71, N.353, pagg.17-35, 1976

Eurostat (2007) CIS 4. The 4th Community Innovation Survey, Quality Report for Country Italy EUROSTAT WEB page: http://forum.europa.eu.int/Public/irc/dsis/Home/main - Section S&T and Innovation Statistics/ CIS4/CIS4 Quality Reports.

H. Derivazione delle unità

Descrizione

Questo sotto-processo consiste nella creazione di nuove unità statistiche, laddove queste non siano direttamente osservate dal processo di rilevazione o acquisizione di dati di fonte amministrativa. Nella progettazione di un'indagine la popolazione, l'unità statistica e l'unità di rilevazione trovano generalmente corrispondenza nell'osservazione pianificata, e le unità statistiche sono spesso create in modo semplice.

Più complesso può essere il caso dell'uso di dati di fonte amministrativa, dove le unità statistiche non sempre esistono come tali negli archivi amministrativi, richiedendo spesso, oltre all'aggregazione e alla divisione di unità, delle procedure più complesse di creazione e derivazione.

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 possono essere create per derivazione dagli oggetti amministrativi 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.

In generale, la ricostruzione dell'unità statistica può essere:

semplice, quando l'unità coincide o è facilmente riconducibile a quella obiettivo per aggregazione o divisione. Per esempio: in una rilevazione, l'unità statistica "famiglia" è creata per aggregazione delle unità "individuo" (relazione 1: n); gli individui iscritti nelle liste dell'Anagrafe sono unità statistiche della popolazione degli individui (relazione 1:1); nell'archivio Emens, prodotto mensilmente dall'Inps, l'unità statistica lavoratore è ottenuta aggregando diversi profili contributivi (relazione 1: n);

assistita da esperto, quando la ricostruzione dell'unità statistica è più complessa e richiede l'ausilio di esperti di settore che, nel caso di dati di fonte amministrativa 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;

assistita da integrazione tra archivi. Per esempio, nell'uso di dati di fonte amministrativa, 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;

mista, quando richiede l'ausilio congiunto sia delle conoscenze di esperti che dell'integrazione tra archivi. 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).

Nel ricostruire l'unità statistica si può incorrere nel cosiddetto errore di derivazione, che ha impatto sull'accuratezza e in particolare sulla copertura (sia in termini di sotto-copertura che di sovra-copertura).

Principio H.1. 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.

Suggerimenti

- Nel caso di rilevazioni dirette, la creazione di nuove unità non dovrebbe comportare grosse criticità. È tuttavia importante chiedersi se nel processo di derivazione e creazione delle nuove unità si sono commessi errori che possono avere impatto sulla copertura della popolazione di interesse.
- Nell'utilizzo di dati amministrativi, è necessario in fase di progettazione: lo studio degli oggetti contenuti nell'archivio amministrativo di riferimento 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.
- È importante che nel processo di derivazione delle nuove unità si valuti attentamente l'applicabilità delle tecniche disponibili in letteratura.
- È necessario individuare in che modo poter ricostruire l'unità statistica a partire dall'unità amministrativa, se in modo semplice, o assistita da esperto, o assistita da integrazione con altri archivi, o mista.
- Una volta individuate e derivate le unità di interesse statistico è necessario valutare l'errore di derivazione.
- Il processo di derivazione delle unità dovrebbe essere riproducibile e documentato. Le ipotesi sottostanti tale processo dovrebbero essere esplicitate e documentate.

Indicatori di qualità e performance

Una misura dell'errore di derivazione può essere fornita dal numero di unità che non possono essere attribuite univocamente all'unita nuova o derivata (o alla popolazione di interesse).

Altre misure possono derivare dalla procedure utilizzate nel processo di derivazione. Per es. qualora la ricostruzione dell'unità sia assistita dall'integrazione, gli errori di derivazione potrebbero derivare da errori di *linkage*, per la cui valutazione è spesso necessario il ricorso al controllo manuale con operatori esperti.

Mappatura con i sotto-processi GSBPM

2.5., 3.5., 5.5., 8.1.

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

I. Derivazione delle variabili

Descrizione

Questo sotto-processo consiste nella costruzione di variabili che non sono esplicitamente raccolte nel processo di indagine o osservate nelle fonti amministrative utilizzate, ma che è necessario diffondere. La derivazione di nuove variabili si ottiene applicando una funzione di trasformazione semplice, ossia tramite formule aritmetiche (regole deterministiche), oppure attraverso l'applicazione di modelli statistici (con componente casuale) a una o più variabili che sono state raccolte o acquisite.

Nell'uso di dati di fonte amministrativa, la derivazione di nuove variabili può comportare la trasformazione di variabili amministrative in variabili statistiche e richiedere un'analisi delle relative definizioni. In questo ambito possono essere sfruttate le informazioni interne ad una singola fonte oppure disponibili da più fonti. In quest'ultimo caso l'attività di derivazione delle variabili implica un'attività preliminare di integrazione tra archivi, con gli eventuali errori che si possono generare nell'applicazione di procedure di *record linkage* (si veda anche Sezione E sull'integrazione). In generale, gli errori che si possono generare nel processo di derivazione delle variabili dipendono da un'errata specificazione delle regole deterministiche o del modello casuale.

Per valutare la validità delle variabili derivate si possono seguire differenti approcci a seconda della disponibilità o meno di variabili di controllo, e in particolare:

- un confronto puntuale tra i dati, il calcolo di misure di scala e di forma distributiva degli errori e quello di funzioni di distanza, quando vi sia la disponibilità di variabili di controllo diretto, ossia variabili statistiche con definizioni coincidenti o raccordabili che abbiano funzione di standard di riferimento (*gold standard*);
- l'applicazione di tecniche per la ricerca di valori anomali (*outlier*) e tecniche regressive anche multivariate per lo studio delle relazioni funzionali, quando vi sia 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;
- lo studio della coerenza tra variabili (intra-fonte o tra fonti), l'analisi fattoriale o i modelli a classi latenti, quando non vi sia la disponibilità né di variabili di controllo né di variabili funzionali.

Strettamente legato alla derivazione delle variabili è il tema delle classificazioni adottate, trattato nella Sezione F.

Nell'attività di derivazione delle variabili si possono generare errori di trattamento che hanno impatto sull'accuratezza finale dei risultati. Nell'uso di dati amministrativi qualora la variabile amministrativa non corrisponda correttamente a quella statistica (errore di specificazione) e si commettano errori nel processo di derivazione (errori di misura), ciò può compromettere la pertinenza delle statistiche prodotte.

Principio I.1. Derivazione delle variabili
Il procedimento di derivazione delle variabili deve seguire pratiche consolidate. Tutte le regole e 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 deve essere documentato.

Suggerimenti

- È opportuno esplicitare le regole o le ipotesi sottostanti il processo di derivazione delle variabili.
- L'intero processo di derivazione delle variabili e dovrebbe essere riproducibile e documentato.
- Nell'uso di dati di fonte amministrativa, è opportuno analizzare e armonizzare 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 variabili statistiche.
- È opportuno valutare la validità del processo di derivazione delle variabili sfruttando il più possibile l'informazione a disposizione (variabili di controllo di tipo: *gold standard*, funzionali; correlazioni tra variabili).
- Nel caso di utilizzo di dati di fonte amministrativa, i metodi di validazione delle variabili dovrebbero essere applicati anche in funzione della finalità di utilizzo. Se l'archivio è utilizzato direttamente per la produzione statistica è auspicabile che la validazione delle variabili 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.

Indicatori di qualità e performance

I principali indicatori che possono essere calcolati attengono alla misurazione dell'errore di processo e derivano dalle tecniche suggerite per la verifica della validità delle variabili. Quindi nel caso di disponibilità di variabili di controllo di tipo *gold standard*, si tratta di misure di distanza tra la variabile derivata e quella presa come standard di riferimento. Nel caso di disponibilità di variabili di controllo funzionali per esempio si possono analizzare indici di correlazione e ricercare valori anomali (*outlier)*.

Nel caso che il processo di derivazione delle variabili avvenga tramite integrazione di archivi amministrativi, le misure dell'errore di derivazione potrebbero derivare da errori di *linkage*, per la cui valutazione è spesso necessario il ricorso al controllo manuale con operatori esperti.

Mappatura con i sotto-processi GSBPM

2.5., 3.5., 5.5., 8.1.

Bibliografia

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-databusiness-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.

J. Destagionalizzazione

Descrizione[10]

La stagionalità, nella dinamica di una serie storica, è quella componente che si ripete ad intervalli regolari ogni anno, con variazioni di intensità più o meno analoga nello stesso periodo (mese, trimestre, etc.) di anni successivi e di intensità diversa nel corso di uno stesso anno. La sua presenza, potendo mascherare altri movimenti di interesse, tipicamente le fluttuazioni cicliche, viene spesso considerata di disturbo nell'analisi della congiuntura economica; essa, ad esempio, rende problematica l'interpretazione delle variazioni osservate su una serie storica tra due periodi consecutivi dell'anno (ossia la variazione congiunturale), essendo queste spesso influenzate in misura prevalente dalle oscillazioni stagionali piuttosto che da movimenti dovuti ad altre cause (come al ciclo economico). Questi ultimi possono essere, invece, correttamente evidenziati calcolando le variazioni congiunturali sui dati destagionalizzati, dai quali, cioè, è stata opportunamente rimossa la componente stagionale.

Per destagionalizzazione si intende quindi un metodo statistico atto a identificare e rimuovere le fluttuazioni di carattere stagionale di una serie storica, che impediscono di cogliere correttamente l'evoluzione dei fenomeni considerati.

L'impiego di dati destagionalizzati permette di comparare l'evoluzione di diverse serie storiche e trova ampia applicazione nell'utilizzo congiunto delle statistiche prodotte da diversi Paesi.

Un'altra pratica, strettamente connessa alla precedente, è quella di correggere i dati per la cosiddetta componente di calendario, determinata dalla diversa composizione del calendario nei singoli periodi dell'anno, che contribuisce anch'essa ad offuscare il segnale congiunturale di interesse. Il diverso numero di giorni lavorativi o di giorni specifici della settimana in essi contenuti, come anche il modo in cui si collocano, nei periodi messi a confronto, le festività nazionali civili e religiose, fisse e mobili, e gli anni bisestili, possono costituire una fonte di variazione di breve periodo per molte serie storiche. Tali effetti, non necessariamente analoghi tra paesi o settori, inficiano la comparabilità nel tempo dei fenomeni economici e pertanto sono spesso rimossi unitamente alla componente stagionale. Il ricorso a tale trasformazione dei dati consente, in particolare, di cogliere in maniera più adeguata sia le variazioni tendenziali (calcolate rispetto allo stesso periodo dell'anno precedente), sia le variazioni medie annue. In molti casi, accanto ai dati destagionalizzati e corretti, vengono prodotte anche serie storiche al netto dei soli effetti di calendario.

Generalmente, l'ipotesi sottostante alla costruzione di una procedura di destagionalizzazione è che ogni serie storica, osservata a cadenza infra-annuale, sia esprimibile come una combinazione delle seguenti componenti non osservabili:

- una componente di trend, che rappresenta la tendenza di medio-lungo periodo, talvolta denominata anche ciclo-trend;
- una componente stagionale, costituita da oscillazioni di periodo annuale;
- una componente irregolare, dovuta a movimenti erratici, cioè a fluttuazioni di breve periodo non sistematiche e non prevedibili.

[10] La descrizione è tratta dal sito dell'Istat http://www.istat.it/it/strumenti/metodi-e-strumenti-it/analisi che a sua volta riflette i risultati di un gruppo di lavoro Istat per la definizione degli standard in materia (Istat, AAVV, 2015) dove si possono reperire altri riferimenti bibliografici rilevanti.

Nell'ambito della produzione statistica ufficiale, gli approcci metodologici più diffusi alla destagionalizzazione sono essenzialmente i due, il cui impiego viene anche incoraggiato nelle linee guida europee sulla destagionalizzazione (Eurostat, 2015):

- i metodi di tipo *Arima model based* (AMB), basati sull'ipotesi che esista un modello statistico parametrico (Arima) in grado di descrivere adeguatamente la struttura probabilistica del processo stocastico generatore della serie storica osservata
- i metodi *filter based* (FLB) di tipo non parametrico o semiparametrico, in cui la stima delle componenti avviene senza ipotizzare l'esistenza di un modello statistico rappresentante la serie analizzata, ma mediante l'applicazione iterativa di una serie di filtri lineari costituiti da medie mobili centrate di diversa lunghezza.

La fase di applicazione di uno dei metodi indicati, per la eliminazione della componente stagionale, è preceduta da una fase di pretrattamento dei dati in cui: si sceglie lo schema di scomposizione che lega le diverse componenti delle serie storica (additiva, moltiplicativa, log-additiva, ecc.), si identificano e si eliminano valori anomali (*outliers*) ed effetti di calendario.

J.1. Destagionalizzazione

Le procedure di destagionalizzazione devono essere mirate ad eliminare la componente stagionale di una serie storica. I dati destagionalizzati devono essere privi di effetti residui della stagionalità. L'approccio utilizzato per destagionalizzare i dati deve essere giustificato e basato su metodologie consolidate e condivise. Le assunzioni sottostanti l'approccio utilizzato devono essere verificate periodicamente. Gli utenti devono essere chiaramente informati sull'esistenza di dati destagionalizzati e sulle metodologie applicate.

Suggerimenti

- Una serie storica va destagionalizzata solo se c'è evidenza che la serie stessa è chiaramente influenzata da fattori stagionali e quando la sottostante stagionalità può essere identificata in modo sufficientemente affidabile, cioè quando essa non è oscurata o nascosta da un alto livello di fluttuazioni irregolari.
- La destagionalizzazione dovrebbe essere preceduta da un trattamento preliminare dei dati volto a correggere l'influenza dovuta al diverso numero di giorni lavorativi, alle festività (fisse o mobili, civili o religiose) e, infine, a valori anomali (*outlier*). Tutte le procedure di pre-trattamento devono seguire metodologie consolidate e condivise e devono essere adeguatamente documentate.
- La stima della componente stagionale deve essere condotta utilizzando procedure consolidate e condivise. A cadenze temporali regolari è necessario rivedere le specifiche utilizzate per il pretrattamento e per la stima della componente stagionale, per tener conto sia di eventuali revisioni dei dati grezzi già diffusi, sia della diffusione di nuovi dati.
- La metodologia adottata deve essere adeguatamente documentata insieme al software utilizzato e alla relativa versione. Le specifiche della procedura utilizzata devono essere disponibili per poter essere diffuse su eventuale richiesta degli utenti.
- Per la validazione della destagionalizzazione è necessario utilizzare le diagnostiche standard (grafici, test statistici volti a valutare l'assenza di stagionalità residua, la stabilità della componente stagionale, i residui dei modelli, etc…).

Indicatori di qualità e performance

La qualità della procedura di destagionalizzazione può essere valutata affiancando quanto già fornito negli output dei software per la destagionalizzazione in termini di grafici, statistiche descrittive, criteri parametrici e non parametrici, con analisi diagnostiche grafiche e test statistici aggiuntivi. Le linee guida Europee (2015) suggeriscono inoltre di guardare anche alla plausibilità dei risultati e non solo alla significatività dei test statistici.

Le principali misure suggerite nelle linee guida Europee sono orientate a identificare: l'assenza di errori nella specificazione del modello; l'assenza di effetti residuali stagionali/di calendario o l'eccesso di aggiustamento per gli effetti stagionali/di calendario; l'adeguato trattamento di *outlier* e di cambiamenti nelle dinamiche stagionali della serie; la stabilità delle componenti di ciclo-trend e stagionale e l'assenza di pattern nella componente irregolare; la non influenza della componente irregolare sulle altre componenti della serie.

Mappatura con i sotto-processi GSBPM

6.1.

Bibliografia

Istat AAVV. (2015) Destagionalizzazione di serie storiche con metodologia Arima model based (AMB) implementata nel software JDemetra+. Istat.

Eurostat (2015), Ess Guidelines on Seasonal Adjustment. Manuals and Guidelines. ISBN: 978-92-79-45176-8. DOI: 10.2785/317290. URL: http://dx.doi.org/10.2785/317290

K. Politica di revisione

Descrizione[11]

Per revisione si intende una modifica di un dato statistico precedentemente diffuso e per politica di revisione si intende l'insieme delle regole che stabiliscono le modalità con le quali i dati sono sottoposti a revisione.

In alcuni processi la necessità di diffondere tempestivamente le stime di interesse comporta il rilascio di stime preliminari, o provvisorie, che sono successivamente revisionate man mano che nuove (o aggiornate) informazioni si rendono disponibili. Talvolta tale revisione può essere determinata anche dall'applicazione di differenti procedure di stima, da cambiamenti nelle metodologie o da eventi straordinari. Pertanto, le revisioni possono essere:

- ordinarie, di frequenza annuale o infrannuale, determinate dalla disponibilità di nuove informazioni e/o dall'aggiornamento delle procedure impiegate per l'aggiustamento dei dati;
- straordinarie, se la loro frequenza è superiore all'anno (solitamente 5 anni), dovute a cambiamenti metodologici dei dati di base, modifiche di classificazione e di definizione delle variabili;
- non programmate, casuali, legate ad errori di calcolo o nei dati di base.

Le informazioni statistiche diffuse con elevata tempestività (es. statistiche congiunturali) sono caratterizzate, per definizione, da un significativo grado di incertezza delle stime, a causa della ridotta disponibilità di fonti statistiche a breve distanza dal periodo di riferimento.

L'analisi delle revisioni mira a quantificare, sintetizzare e valutare il processo di revisione delle stime preliminari rispetto a quelle pubblicate in periodi successivi (ad esempio un mese, un trimestre o un anno dopo).

Per misurare e analizzare il processo delle revisioni delle stime relative a uno stesso indicatore si utilizza una particolare rappresentazione tabellare denominata "triangolo delle revisioni" (noto anche come *real-time database*). Tale struttura organizza per riga le serie storiche rilasciate a partire da una certa data, mentre dalle colonne è possibile ricostruire la storia delle stime diffuse per ciascun riferimento temporale (mese o trimestre) dal primo rilascio fino all'ultimo disponibile.

Principio K.1. Politica di Revisione
La politica di revisione deve riportare le modalità e i tempi di aggiornamento delle stime. La procedura di revisione deve essere definita con chiarezza e resa nota agli utenti dei dati. L'analisi delle revisioni dovrebbe essere utilizzata anche per il miglioramento della qualità.

Suggerimenti

- Nell'ambito della politica di revisione, tutte le informazioni relative al processo di revisione delle stime devono essere specificate in modo chiaro ed esplicito, evidenziando le fonti utilizzate e la loro tempestività, il numero di revisioni previste, i motivi e il relativo calendario.
- La politica di revisione deve essere comunicata in anticipo ai fruitori dei dati. La pubblicazione di stime soggette a revisione va corredata con l'indicazione dei tempi e delle modalità della revisione, affinché l'utente ne sia preventivamente informato.

[11] La descrizione è tratta dalla pagina ufficiale del sito dell'Istat http://www.istat.it/it/congiuntura/revisioni

- Ciascuna revisione deve essere documentata utilizzando la rappresentazione del "triangolo delle revisioni", che permette di ricostruire la storia delle stime diffuse e consente una valutazione dell'impatto della politica di revisione. La documentazione deve anche comprendere i risultati relativi al calcolo dei principali indicatori di revisione e rimandare a documenti con analisi più approfondite delle revisioni (qualora disponibili).
- Il "triangolo delle revisioni" deve essere aggiornato regolarmente in occasione della diffusione di nuovi dati. La scelta delle informazioni da diffondere nel triangolo (dati di livello e/o di variazione, destagionalizzati o meno) deve tener conto delle esigenze degli utenti esterni.
- Laddove l'analisi delle revisioni evidenziasse un andamento sistematico delle stime (tendenza della stima preliminare a sottostimare o sovrastimare la successiva), si dovrebbe cercare di individuarne le cause e, quindi, intervenire sul processo di produzione in modo da rimuoverle, ove possibile.
- Revisioni occasionali, non previste dalla politica di revisione, devono essere documentate e motivate. Gli utenti devono essere informati di tali revisioni e dei motivi per cui sono state effettuate.

Indicatori di qualità e performance

Gli indicatori di revisione sono delle misure sintetiche della distribuzione delle revisioni assolute e relative, dove per revisione assoluta si intende la differenza tra una stima al tempo *t* e una stima preliminare, mentre la revisione relativa considera al denominatore anche la stima al tempo *t*. La revisione assoluta è generalmente calcolata sui tassi di variazione mentre quella relativa sui livelli.

Vengono utilizzate misure relative all'ampiezza media delle revisioni, per esempio la Revisione Media Assoluta (RMA) e la Revisione Media Assoluta Relativa (RMAR). Vi sono poi misure relative alla direzione delle revisioni, per esempio la Revisione Media (RM), della quale vengono analizzate variabilità e significatività rispetto al valore nullo. Infine si considerano le misure della variabilità delle revisioni e misure relative all'impatto delle revisioni sul segno dei tassi di variazione. Per un approfondimento si consulti Istat (2017) elencato in bibliografia.

Mappatura con i sotto-processi GSBPM

La politica di revisione, in base anche ai motivi sottostanti, è caratterizzata dal fatto che più sotto-processi possono essere effettuati nuovamente, dall'acquisizione (4.3.) alla predisposizione degli output (6.5).

Riferimenti bibliografici

Istat (2017). I principali indicatori sintetici sulle revisioni http://www.istat.it/it/files/2017/03/indicatori-sintetici-sulle-revisioni.pdf

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

UK Office for National Statistics https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/revisions

L. Validazione dei risultati

Descrizione

Nella fase di validazione del prodotto statistico finale, si effettua un controllo dei risultati e della loro qualità prima che questi vengano diffusi, per verificare che siano conformi alle aspettative e non presentino delle anomalie derivanti da errori nel processo produttivo statistico. Tale confronto può avvenire sulla base delle conoscenze del fenomeno in oggetto, sulla base di serie storiche di dati dello stesso processo oppure sulla base di dati provenienti da altre fonti e relative allo stesso fenomeno o a fenomeni correlati. Qualora si riscontrasse un'anomalia nei risultati quindi in fase di validazione tutte le attività del processo produttivo statistico devono essere ricontrollate, e gli indicatori di qualità devono essere valutati, per cercare l'eventuale errore che l'ha generata. Spesso questa ricerca consente di correggere le procedure e può portare a migliorare le edizioni successive del processo produttivo statistico.

Principio L.1. Validazione dei risultati
I risultati prodotti dovrebbero essere valutati prima della loro pubblicazione, possibilmente insieme a esperti del settore per verificare se vi siano o meno delle anomalie. Inoltre, dovrebbero essere calcolati ed analizzati, in modo rigoroso, gli indicatori di qualità del processo.

Suggerimenti

- È opportuno che i risultati del processo produttivo statistico, prima di essere diffusi, vengano controllati da esperti della materia trattata, interni oppure esterni all'ente, soprattutto in caso di valori sospetti.
- I risultati del processo devono essere valutati mediante confronti con i risultati di precedenti edizioni, sia nel caso di indagine diretta, sia nei casi di utilizzo di dati di fonte amministrativa, dove siano stati utilizzati i dati della stessa fonte. Il confronto può essere effettuato anche con fonti esterne all'ente. In caso di discrepanze, le differenze 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 come accade ad esempio per alcuni rapporti demografici. Anche in questo caso eventuali differenze devono essere giustificate e documentate.
- I punti di criticità del processo, che potrebbero aver portato a valori anomali o errati dei risultati, possono essere individuati più agevolmente mediante il calcolo di indicatori di qualità, sia in riferimento alla qualità dei dati di input che alla qualità del processo stesso. Nel caso di indagini è opportuno verificare la copertura della popolazione obiettivo e i tassi di risposta. In generale si possono calcolare indicatori di coerenza tra statistiche.
- Nel caso siano possibili margini di miglioramento agendo sulla fonte amministrative impiegata, il risultato della validazione dovrebbe concretizzarsi in informazioni di ritorno per l'ente titolare del dato amministrativo.

Indicatori di qualità e performance

Misure che possono essere calcolate in fase di validazione sono indicatori di coerenza, per esempio con altre fonti, tra stime derivanti da processi con diversa periodicità oppure tra stime preliminari e definitive.

L’indicatore di coerenza tra fonti viene costruito dividendo la differenza tra le stime (quella derivante dal processo in esame con quella derivata da un’altra fonte di confronto) con la stima della fonte di confronto.

L.1. Coerenza (differenza relativa) tra stime

L.2. Comparabilità delle stime nel tempo: lunghezza della serie storica comparabile dei dati

Nei caso di statistiche su flussi in entrata e in uscita (per es. quando trasferimenti da una regione ad un’altra), si possono calcolare degli indicatori di confronto:

L.3. Indicatore di Asimmetria: discrepanze tra dati relativi a flussi per esempio per coppie di paesi/regioni/..

Mappatura con i sotto-processi GSBPM

2.5., 3.5., 6.2., 8.1.

M. Diffusione dei dati e tutela della riservatezza, archiviazione, documentazione

Descrizione

A conclusione del ciclo produttivo vi è la fase di diffusione dei risultati agli utenti attraverso vari canali, corredati da tutto quanto è di supporto nell'accesso e nell'utilizzo di dati statistici, con le necessarie garanzie di tutela della riservatezza. Questa attività viene affiancata da altre attività che riguardano l'archiviazione dei micro e macrodati prodotti e la produzione della documentazione necessarie per le finalità interne.

Le pratiche adottate nella diffusione e la documentazione a corredo dei dati pubblicati migliorano la qualità e soprattutto le componenti dell' "accessibilità" e "chiarezza". L'accessibilità è legata al tipo di supporto utilizzato (database on-line attraverso interfaccia grafica di ricerca, CD-Rom, volume cartaceo) e alla facilità di reperimento dell'informazione, oltre che alla possibilità per l'utente di scaricare i dati in formati riusabili (ad esempio file csv, txt, xls, RDF[12], SDMX[13]). 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. File di microdati sono oggi anche accessibili, da remoto, attraverso specifici laboratori. La chiarezza è, invece, legata alla disponibilità di metadati relativi ai contenuti informativi e alle principali caratteristiche del processo di produzione e di indicatori di qualità, al fine di consentire agli utenti la corretta interpretazione e l'uso consapevole dei dati.

La legge istitutiva del Sistema statistico nazionale[14] prevede che debba essere tutelata la riservatezza dei rispondenti, e, in particolare, che i dati oggetto di diffusione debbano essere adeguatamente trattati (cfr. Appendice D).

La tutela della riservatezza può essere garantita attraverso l'adozione di misure atte a proteggere i dati diffusi in modo che non siano possibili violazioni della riservatezza dei rispondenti, applicando metodi che consentono di quantificare il rischio di identificazione di una unità attraverso le sue caratteristiche. Gli approcci utilizzabili per tutelare la riservatezza dei dati diffusi si basano sulla restrizione all'accesso ai dati e sull'applicazione di metodi per proteggere i dati da violazioni.

Nella protezione dei dati aggregati in tabelle sono previste le seguenti tre fasi: *i)* valutare quali tabelle possano essere a rischio in base al loro contenuto, *ii)* stabilire i criteri in base ai quali una cella è a rischio di violazione della riservatezza (regola di rischio), *iii)* applicare le procedure per la tutela della riservatezza. Queste dipendono dal tipo di tabelle e dati. Tra le regole di rischio più utilizzate vi è quella di frequenza (o della soglia): una cella è a rischio se il numero di contributori ad essa afferente è inferiore al "valore soglia" prefissato. Il "Codice di deontologia e di buona condotta per i trattamenti di dati personali a scopi statistici e di ricerca scientifica effettuati nell'ambito del Sistema statistico nazionale"[15] prevede che il valore minimo che può assumere la soglia sia pari a tre. Per le tabelle di intensità oltre alla regola di frequenza è possibile ricorrere a regole di rischio basate sulla concentrazione del carattere osservato (ad esempio regole della dominanza e del rapporto). I metodi per la protezione nelle tabelle possono consistere in metodi di riduzione dell'informazione diffusa (attraverso l'accorpamento di modalità adiacenti, la definizione di combinazioni di

[12] RDF – Resource Description Framework è il formato utilizzato nei Linked Open Data.

[13] SDMX – Statistical Data and Metadata eXchange è il formato utilizzato per lo scambio dati tra organizzazioni statistiche.

[14] D.to L.vo 322/89 e successive integrazioni e modifiche di razionalizzazione del DPR. n. 166/2010.

[15] Allegato A3 al Codice in materia di protezione dei dati personali 12 Giugno 2014

modalità tale che la distribuzione del carattere non presenti alcuna cella sensibile, la soppressione dei valori nelle celle.), oppure possono essere di tipo perturbativo pre o post tabellare.

Per il rilascio di dati elementari, ossia di record contenenti informazioni sulle singole unità statistiche, i metodi di misurazione del rischio sono più complessi e spesso basati su modelli probabilistici. Per la protezione di dati elementari si utilizzano principalmente: la ricodifica delle variabili (riducendone il dettaglio diffuso, per es. età in classi quinquennali piuttosto che annuali); la soppressione di specifiche informazioni che possono rendere identificabile una unità; la perturbazione dei dati.

Le misure a garanzia della riservatezza dei dati, sono fondamentali per accrescere la fiducia dei soggetti intervistati verso l'ente produttore di statistica e indirettamente permettere di aumentare la partecipazione alle rilevazioni e di migliorare la qualità dei dati.

L'archiviazione dei dati e dei metadati è un'attività che riguarda non solo i dati finali, siano essi microdati o macrodati, ma tutti i dati di input e di output delle principali fasi del processo produttivo statistico. Ciò è fondamentale per garantire la tracciabilità dei dati e la riproducibilità dei risultati.

Infine, la documentazione del processo e della qualità non è solo utile agli utenti dei dati ma è anche indispensabile per le finalità interne all'ente. In particolare serve a garantire la riproducibilità del processo, ad assicurare la continuità della produzione anche al variare del personale impiegato, a consentire la valutazione della qualità e a interpretare correttamente la qualità ossia comprendere la qualità dei risultati in relazione alle procedure e alle metodologie applicate.

Principio M.1. Diffusione dei dati e tutela della riservatezza, archiviazione e documentazione
La diffusione dei dati deve avvenire in modo trasparente per gli utenti. I macrodati e i microdati diffusi devono essere preventivamente trattati per garantire una adeguata tutela della riservatezza e supportati dalle informazioni che ne accrescono l'utilizzabilità. I microdati validati e i macrodati devono essere archiviati corredati da metadati e indicatori di qualità. Tutte le fasi del processo devono essere adeguatamente documentate.

Suggerimenti

Diffusione e tutela della riservatezza

- È opportuno garantire il contemporaneo accesso ai dati da parte di tutti i potenziali utenti, ivi compresi quelli appartenenti all'amministrazione produttrice dei dati, disponendo affinché qualsiasi accesso privilegiato prima della diffusione sia limitato, controllato e reso noto.
- È necessario predisporre misure per assicurare la massima facilità di accesso ai dati e ai metadati;
- È necessario definire e rendere noto il tempo che intercorre tra la data di riferimento dei risultati prodotti (l'ultimo giorno se si tratta di un periodo di riferimento) il momento della loro diffusione, noto come indicatore di tempestività.
- È opportuno definire ex-ante e rendere pubblico il calendario delle diffusioni dei dati che tenga il più possibile conto delle esigenze degli utenti.
- È necessario diffondere le statistiche in forma chiara e comprensibile, sempre accompagnate dai metadati, presentandole in modo da offrire un'interpretazione il più possibile imparziale e in modo da rendere possibili confronti significativi nel tempo e nello spazio. Eventuali limitazioni dei dati,

quali ad esempio l'esistenza di interruzioni nelle serie storiche, l'eventuale carattere provvisorio dei dati rilasciati, il livello territoriale per cui i dati sono significativi, dovrebbero essere rese note.

- È auspicabile corredare la diffusione dei dati con informazioni sulle metodologiche adottate e con indicatori di qualità.
- Nel caso siano diffuse stime preliminari, indicare chiaramente tale specifica, così come dovrà essere indicata successivamente la politica di revisione adottata. Allo stesso modo revisioni occasionali non previste dalla politica di revisione dovranno essere annunciate, motivate e documentate.
- Qualora dovessero essere individuati errori dopo la diffusione, provvedere tempestivamente alla loro correzione e alla pubblicazione delle rettifiche, spiegandone le ragioni.
- La diffusione dell'informazione statistica, laddove possibile, dovrebbe avvalersi di avanzate tecnologie dell'informazione e della comunicazione (sito web, modalità di accesso trasparenti, formato dei dati e licenze che consentano il riuso delle informazioni).
- È auspicabile avviare attività di industrializzazione nei processi di diffusione attraverso l'adozione di standard internazionali per renderli *metadata-driven*, facilitare l'interoperabilità semantica e favorire l'*open (statistical) data*.
- Per quanto possibile, va prevista la distribuzione gratuita dell'informazione statistica su sito web, accessibile secondo gli standard internazionali.
- Sui dati diffusi è necessario verificare il rischio di violazione della riservatezza e in caso adottare le tecniche più appropriate per la tutela della riservatezza.
- Per la tutela della riservatezza nella diffusione dei dati è suggerito, ove possibile, l'uso di software generalizzati.

Archiviazione

- È opportuno archiviare file di dati lungo tutto il processo produttivo insieme con i descrittori del file e i metadati utili per l'interpretazione dei dati. Almeno i seguenti file dovrebbero essere archiviati: i file di input di fonte amministrativa; i dati derivanti dal processo di rilevazione presso le unità di rilevazione; i file grezzi e puliti antecedenti e successivi alle procedure di identificazione e correzione dell'errore; i microdati validati finali successivi alla procedura di validazione dei risultati.
- I microdati validati, che hanno passato tutte le fasi del processo produttivo e i controlli di qualità, devono essere archiviati insieme ai metadati necessari per la loro interpretazione (tracciati record, variabili e classificazioni associate). Anche i macrodati, sia direttamente pubblicati sia utilizzati per la costruzione di tavole o grafici anch'essi pubblicati, devono essere opportunamente archiviati e chiaramente referenziati per edizione del processo, periodo di riferimento dei dati, canale di diffusione e data di pubblicazione.
- Allo stesso modo, gli indicatori di qualità devono essere archiviati e resi disponibili a fini valutativi, naturalmente essi saranno differenti a seconda del tipo di dato e delle fasi del processo produttivo. Si avranno pertanto indicatori specifici nel caso di data set provenienti da fonti amministrative o ottenuti tramite rilevazione sul campo. Gli indicatori sono stati suggeriti per ogni fase/sotto-processo del processo produttivo in questo manuale. Il *dataset* risultante dalla fase di controllo e correzione avrà, ad esempio, indicatori relativi ai tassi di imputazione o al confronto delle distribuzioni prima e dopo l'applicazione della procedura.
- I dati personali devono essere archiviati separatamente da quelli identificativi, in base alle situazioni previste dall'art. 11 del "Codice di deontologia e di buona condotta per i trattamenti di dati personali a scopi statistici e di ricerca scientifica effettuati nell'ambito del Sistema statistico nazionale".
- È necessario adottare tutte le misure possibili utili a garantire il segreto statistico a tutela delle persone fisiche e giuridiche, quali per esempio: la protezione dei questionari durante la raccolta dei dati, il loro trasferimento e la loro conservazione; il giuramento da parte di tutto il personale che ha

accesso ai dati sulla tutela della confidenzialità; l'applicazione di restrizioni all'accesso ai dati sia fisico ai luoghi che virtuale ai server dove sono archiviati; la verifica del rischio di violazione.

- A tutto il personale coinvolto nel trattamento dei dati in cui sia possibile incorrere in violazioni della riservatezza dovranno essere fornite adeguate istruzioni sulla tutela del segreto statistico e sul rispetto delle norme in materia di protezione dei dati personali.

Documentazione

- È opportuno che siano documentati tutti i metadati di tipo strutturale, e cioè tutti gli elementi che descrivono i dati: unità, popolazioni, operatori statistici, variabili.
- È opportuno che siano documentati tutti i metadati referenziali, ossia tutti gli elementi che descrivono le attività del processo produttivo e la qualità: fasi e sotto-processi effettuati, attività messe in campo per prevenire, monitorare o valutare gli errori, indicatori di qualità.
- È opportuno che la documentazione organizzata secondo i punti precedenti sia il più possibile completa anche in vista della predisposizione delle relazioni sulla qualità, laddove previsto, secondo l'articolo 12 del regolamento UE 759/2015 sulla loro trasmissione a Eurostat.

Indicatori di qualità e performance

Possono essere calcolati in questa fase indicatori di pertinenza relativi agli accessi, così come suggeriti tra gli indicatori nella Sezione A.

Anche se dipendono da tutto il processo produttivo statistico è quando si diffondono i dati che si calcolano gli indicatori di puntualità e tempestività del rilascio. La prima riflette eventuali discrepanze tra la data effettiva di diffusione e quanto previsto (da regolamento, nel calendario di diffusione, …). La seconda misura il tempo intercorrente tra il periodo di riferimento delle stime e quando queste sono disponibile agli utenti.

M.1. Puntualità: differenza tra data programmata di diffusione/trasmissione dei dati e data effettiva.

M.2. Tempestività dei risultati: differenza tra data di diffusione/trasmissione dei risultati e la data cui i risultati o le stime si riferiscono.

Sulle stime prodotte possono anche essere calcolati indicatori di comparabilità nel tempo e di coerenza (si veda Sezione L), quali:

M.3. Lunghezza della serie storica disponibile e confrontabile

M.4. Coerenza con statistiche disponibili sullo stesso fenomeno.

Mappatura con i sotto-processi GSBPM

2.5., 3.3., 6.4., 7.1., 7.2, 7.3., 7.4., 7.5., 8.1.

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 (2011). Linee guida per il miglioramento della qualità della diffusione delle statistiche ufficiali da parte dei soggetti del Sistema statistico nazionale (approvate dal Comstat nella seduta del 16 Dicembre 2011) http://www.sistan.it/fileadmin/Repository/Home/QUALITA_E_SVILUPPO/CODICE/MONITORAGGIO/Linee_guida.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

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

Regolamenti UE 223/2009 e aggiornamento 759/2015.

Statistics Canada (2009). Survey methods and practices

Statistical Disclosure Control website with resources from CASC (2000-2003) CENEX (2006) and the ESSnet (2008-2009) projects. http://neon.vb.cbs.nl/casc/index.htm

Glossario

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, file, 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.

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.

Attendibilità

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

Classificazione

Un insieme di osservazioni discrete, esaustive e mutuamente esclusive che può essere assegnata a una o più variabili per essere misurate nella raccolta o nella presentazione dei dati (OECD Glossary).

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.

Copertura

Cfr Errore di copertura.

Controlli di dominio (o di range)

Processo di verifica se un certo valore dei dati ricade all'interno di un intervallo precedentemente specificato (OECD Glossary, 2007).

Confidenzialità

Proprietà dei dati che indica il grado con cui la loro violazione non autorizzata possa essere pregiudizievole

o dannosa per l'interesse della fonte o di altre parti rilevanti (SDMX Metadata Common Vocabulary, 2009).

Controlli di flusso

Controlli per verificare il flusso ossia l'ordine di digitazione dei campi (variabili) e le condizioni sotto le quali sono ammissibili (Manuale Blaise per lo sviluppo di questionari elettronici).

Controlli di coerenza

Controlli sulla conformità dei dati a condizioni logiche o restrizione sui valori di un dato o gruppo di dati che devono essere soddisfatte affinché i dati possano essere considerati corretti (Glossario Memobust).

Controlli hard

Controlli che si applicano ad errori che devono essere risolti prima che il questionario possa essere considerato "pulito" (Manuale Blaise per lo sviluppo di questionari elettronici).

Controlli soft

Controlli che si applicano ad errori che non è indispensabile risolvere to prima che il questionario possa essere considerato "pulito" (Manuale Blaise per lo sviluppo di questionari elettronici)**.**

Controllo e correzione

Un attività che ha l'obiettivo di identificare, comprendere e correggere valori mancanti o errati nei dati (Glossario Memobust).

Dato giudiziario

Si intende un dato personale idoneo a rivelare provvedimenti di cui all'art. 3 comma 1, lettera a), o) e da r) a u) del DPR 14/2002 n.313 in materia di casellario giudiziario e di anagrafe delle sanzioni amministrative dipendenti da reato e dei relativi carichi pendenti, o la qualità di imputato o di indagato ai sensi dell'art. 60 e 61 del codice di procedura penale.

Dati grezzi

Dati ottenuti dalla fase di rilevazione presso le unità, con gli eventuali controlli in corso di intervista, e sul quale siano state applicate le eventuali procedure di codifica e revisione manuale. Vengono forniti come input al software relativo alla procedura di controllo e correzione.

Dato sensibile

Si intende un dato personale idoneo a rivelare l'origine razziale o etnica le condizioni religiose filosofiche o di altro genere, le opinioni politiche, l'adesione a partiti, sindacati, associazioni o organizzazioni a carattere religioso, filosofico, politico, sindacale, nonché i dati personali idonei a rivelare lo stato di salute e la vita sessuale.

Distorsione (*bias*)

Un effetto che priva un risultato statistico della rappresentatività distorcendolo sistematicamente, in opposizione al errore casuale che può distorcere in una occasione ma si annulla in media (OECD Glossary, 2007).

Imputazione

È il processo di assegnazione di valori coerenti al posto di dati mancanti, inammissibili o incoerenti che hanno violato le regole di controllo.

Effetto (o errore dell') intervistatore

Effetto sulle risposte dei rispondenti derivante dai diversi modi in cui gli intervistatori gestiscono la rilevazione. (Biemer P.P. et al, 1991).

Eleggibilità di un'unità

Una unità è eleggibile se appartiene alla popolazione oggetto di indagine.

Errori accidentali o casuali

Errori la cui origine è da attribuirsi a fattori aleatori non direttamente individuabili. *(http://www.istat.it/it/strumenti/metodi-e-strumenti-it).*

Errore di assunzione del modello

Gli errori da assunzione del modello si verificano con l'uso di metodi, come la calibrazione, stimatori di regressione generalizzata, benchmarking, destagionalizzazione o altri modelli non inclusi nelle componenti dell'accuratezza precedenti, allo scopo di calcolare statistiche o indici (OECD Glossay, 2007).

Errore di copertura

Gli errori di copertura di una lista sono le discrepanze in termini di omissioni (sotto-copertura), errate inclusioni e duplicazioni (sovra-copertura) tra gli elementi della lista e quelli della popolazione obiettivo che intende rappresentare.

Errore di lista

Errori nelle variabili della lista che pregiudicano il contatto dell'unità o la sua corretta classificazione in strati funzionali al disegno di campionamento.

Errore di Mancata Risposta Parziale

L'errore di mancata risposta parziale si verifica quando unità del campione o della popolazione su cui rilevare le informazioni non rispondono ad alcuni quesiti del questionario (tradotta e riadattata da Biemer e Lyberg, 2003).

Errore di Mancata Risposta Totale

L'errore di mancata risposta totale si verifica quando unità del campione o della popolazione su cui rilevare le informazioni non rispondono totalmente al questionario, oppure le risposte vengono considerate insufficienti e intera unità considerate come non rispondenti (tradotta e riadattata da Biemer e Lyberg, 2003).

Errore di misura

Per errore di misura si intende una discrepanza tra valore "vero" e valore "osservato" di una variabile in

un'unità.

Errore di Risposta

Gli errori di risposta sono quelli che hanno origine nel processo di rilevazione dei dati, a causa del rispondente, intervistatore, questionario e tecnica di raccolta dati (tradotto e riadattato da OECD Glossary).

Errori sistematici

Errori la cui origine è da attribuirsi a difetti strutturali o organizzativi del processo di produzione dell'informazione statistica, alla struttura del modello, o al sistema di registrazione adottati, e si manifestano nella maggior parte dei casi come deviazioni "in una stessa direzione" dal valore vero di una o più variabili rilevate. *(http://www.istat.it/it/strumenti/metodi-e-strumenti-it).*

Errore di Trattamento (*processing*)

Gli errori di trattamento sono quelli che hanno origine nelle fasi successive al processo di rilevazione dei dati, a causa di difetti nella loro implementazione (tradotto e riadattato da OECD Glossary).

Errore Quadratico Medio (o *Mean Squared Error*)

È uguale alla somma del quadrato della distorsione e la varianza dello stimatore. (OECD Glossay, 2007).

***Linked Open Data* (dati collegati di tipo aperto)**

Dati di tipo aperto pubblicati in un formato atto ad essere collegati tra loro, si basa sulle tecnologie del web semantico. I dati sono descritti semanticamente tramite metadati e ontologie; seguono il paradigma *Resource Description Framework* (RDF) per cui le risorse sono univocamente individuate da un *Uniform Resource Identifier* (URI) sul Web. I dati sono detti "linked" per la possibilità di referenziarsi ("collegarsi") tra loro. Nel referenziarsi, si usano relazioni ("link") che hanno un preciso significato e spiegano il tipo di legame che intercorre tra le due entità coinvolte nel collegamento. I Linked (Open) Data consentono quindi l'integrazione e l'interoperabilità tra dati.

Lista di riferimento (*frame*)

Una lista, materiale o dispositivo che delimita, identifica e permette l'accesso agli elementi di una popolazione obiettivo. Gli elementi (unità) sono linkabili con quelli della popolazione obiettivo (finita e identificabile), vi sono delle informazioni che permettono di localizzare l'unità, vi sono delle informazioni che permettono di classificare le unità in modo utile alle procedure di campionamento (semplificata e tradotta da Lessler and Kalsbeek, (1992).

Matching

Abbinamento di microdati da fonti diverse basato su caratteristiche presenti in quelle fonti (CODED).

Metodi deduttivi (imputazione)

Metodi di imputazione che si basano sulla possibilità di sfruttare le informazioni presenti in modo da poter dedurre il valore da sostituire al dato mancante da una o più variabili ausiliarie.

Metodi deterministici (imputazione)

Metodi nei quali imputazioni ripetute per unità aventi le stesse caratteristiche considerate producono sempre gli stessi valori imputati (es. *Imputazione deterministica con media*, o *con moda* per variabili

qualitative, *Imputazione con regressione, lineare* in genere per variabili quantitative, *log-lineare* o *logistica* per variabili qualitative, *Imputazione dal più vicino donatore*).

Metodi stocastici (imputazione)

Metodi nei quali imputazioni ripetute per unità aventi le stesse caratteristiche considerate possono produrre differenti valori imputati; si caratterizzano per la presenza di una componente aleatoria, corrispondente ad uno schema probabilistico associato al particolare metodo d'imputazione prescelto (es. *Imputazione con donatore casuale all'interno delle classi*, *Imputazione con regressione casuale*, versione stocastica dell'imputazione con regressione, in cui i valori imputati sono sempre stimati con l'equazione di regressione nella quale si aggiunge la componente residuale, *Imputazione multipla, Hot deck sequenziale, Hot deck gerarchico, Reti neurali*).

Microdati validati

È il file di dati individuali generato successivamente alla fase di validazione dei dati, sia essa interna all'indagine sia che utilizzi fonti esterne, e quindi è il file dei microdati che consente la riproducibilità dei dati diffusi dall'Istat. Alcune procedure di indagine potrebbero non consentire la distinzione tra questa tipologia di file e quella dei puliti, che finiscono per coincidere(Glossario a h*ttp://siqual.istat.it/SIQual*).

Ontologia di dominio

Rappresentazione formale, condivisa ed esplicita di una concettualizzazione di una peculiare porzione di realtà (dominio di interesse).

***Open Data* (dati di tipo aperto)**

Un contenuto o un dato si definisce aperto se chiunque è in grado di utilizzarlo, ri-utilizzarlo e ridistribuirlo, soggetto, al massimo, alla richiesta di attribuzione e condivisione allo stesso modo" [definizione della Open Knowledge Foundation, 2004]. Il Legislatore italiano con la Legge 17 dicembre 2012, n. 221 ha formalizzato una definizione di dati aperti (formalmente "dati di tipo aperto") inserendola all'interno dell'art. 68 del Codice dell'Amministrazione Digitale. Secondo tale definizione, sono dati di tipo aperto, i dati che presentano le seguenti tre caratteristiche: *a)* sono disponibili secondo i termini di una licenza che ne permetta l'utilizzo da parte di chiunque, anche per finalità commerciali, in formato disaggregato; *b)* sono accessibili attraverso le tecnologie dell'informazione e della comunicazione, ivi comprese le reti telematiche pubbliche e private, in formati aperti ai sensi della lettera a), sono adatti all'utilizzo automatico da parte di programmi per elaboratori e sono provvisti dei relativi metadati; *c)* sono resi disponibili gratuitamente attraverso le tecnologie dell'informazione e della comunicazione, ivi comprese le reti telematiche pubbliche e private, oppure sono resi disponibili ai costi marginali sostenuti per la loro riproduzione e divulgazione.

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.

Politica di revisione

L'insieme delle regole che stabiliscono le modalità con le quali i dati sono sottoposti a revisione.

Popolazione obiettivo

Insieme delle unità statistiche di studio. (Survey methodology, 2004, Groves RM et al.).

Privacy

Diritto di una persona o organizzazione di mantenere segreti i propri affari e relazioni personali che coinvolge l'obbligo da parte di chi mantiene l'informazione del soggetto a fare altrettanto (Unece, 2009).

Processo produttivo statistico

È una sequenza di procedure interdipendenti e legate tra loro che, ad ogni stadio, consumano risorse (tempo persona, energia, machine, denaro) per convertire input (dati, materiali, parti, etc.) in output. Fattori essenziali di un processo generico sono input, output a un insieme di passi o attività che trasformano l'input in output. (tradotto da http://www.businessdictionary.com/definition/process.html).

Record linkage

Il processo di attribuzione di informazioni contenute in due o più fonti diverse di microdati quando riferibili ad una medesima unità.

Revisione

Per revisione si intende una modifica di un dato statistico precedentemente diffuso. A seconda dei motivi le revisioni si classificano in: ordinarie, straordinarie, altre revisioni non programmate(Istat, sito congiuntura, area revisioni).

Sotto-copertura

Cfr Errore di copertura.

Sovra-copertura

Cfr Errore di copertura.

Stima preliminare

Termine usato per descrivere la prima versione di una serie di dati rilasciati o per descrivere versioni precedenti a quelle finali. In ogni caso si tratta di dati soggetti a revisione (OECD Glossary).

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.

Titolare

Il titolare del lavoro è il responsabile dell'intero processo statistico, dalla fase di progettazione a quella di diffusione. Generalmente coincide con il responsabile del trattamento dei dati personali (Programma Statistico Nazionale).

Unità di primo stadio

In un disegno di campionamento a più stadi il campione di unità elementari si ottiene come risultato di più stadi di campionamento: gli elementi della popolazione sono prima raggruppati in sottopopolazioni disgiunte, chiamate unità di primo stadio e viene estratto un campione probabilistico di tali unità. Successivamente, all'interno di ogni unità di primo stadio selezionata nel campione vengono estratte con campionamento probabilistico le unità di secondo stadio e così via fino ad arrivare alle unità elementari. (Tradotto e adattato da Sarndal, et al., 2003).

Unità di rilevazione

Unità che viene contattata per ottenere le informazioni relative alle unità di analisi. L'unità di rilevazione può coincidere con una unità di analisi oppure essere una unità funzionale all'acquisizione delle informazioni su altre unità. L'unità di rilevazione è definita per tutte le indagini dirette, e per quelle indagini amministrative in cui i dati sono raccolti presso una pluralità di enti: ad esempio, gli Istituti di cura sono le unità di rilevazione della "Indagine rapida sui dimessi dagli istituti di cura" (Glossario a h*ttp://siqual.istat.it/SIQual*).

Unità di secondo stadio

Le unità di secondo stadio sono i raggruppamenti di unità della popolazione obiettivo che vengono estratti all'interno delle unità di primo stadio, durante il secondo stadio di un disegno di campionamento a più stadi.

Unità statistica o di analisi

Entità oggetto di osservazione del processo. Le unità di analisi possono essere sia collettivi direttamente osservabili, ad esempio famiglie, componenti delle famiglie, imprese, lavoratori dipendenti, sia collezioni di eventi, ad esempio vacanze, ricoveri ospedalieri. Una unità di analisi può coincidere con una unità di rilevazione quando fornisce informazioni anche su stessa (Glossario a h*ttp://siqual.istat.it/SIQual*).

Validazione (dell'*output*)

Intesa in termini di validazione degli output o dei risultati è il processo di monitoraggio dei risultati di una produzione di dati e la verifica della qualità dei risultati stessi (Eurostat CODED).

Variabile

Una variabile è una caratteristica di una unità osservata che può assumere una misurazione numerica o una categoria da una classificazione (per es. reddito, età , peso, occupazione, …) (OECD GLossary).

Web Semantico

Insieme di modelli e standard Web in cui le risorse vengono descritte e correlate fra loro in modo formale attraverso l'uso opportuno di metadati. In questo modo si abilitano agenti software automatici a comprendere il significato dei dati e delle informazioni.

Riferimenti bibliografici generali

Biemer P.P. Lyberg L.E. (2003) Introduction to Survey Quality Wiley & Sons. Hoboken, New Jersey

Biemer P.P., Groves R.M., Lyberg L.E., Mathiowetz N.A., Sudman S. (1991) Measurement errors in survey, John Wiley & Sons, 1991

Eurostat (2003), Definition of Quality in Statistics. Eurostat Working Group on Assessment of Quality in Statistics, Luxembourg, October 2-3.

Eurostat (2015) Quality Assurance Framework in the European Statistical System http://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646

Eurostat (2014) ESS Guidelines for the implementation of the ESS quality and performance indicators http://ec.europa.eu/eurostat/documents/64157/4373903/02-ESS-Quality-and-performance-Indicators-2014.pdf/5c996003-b770-4a7c-9c2f-bf733e6b1f31

Eurostat (2012) ESS Quality Glossary http://ec.europa.eu/eurostat/ramon/coded_files/ESS_Quality_Glossary.pdf

Eurostat (2011) European Statistics Code of Practice revised edition 2011 (disponibile nelle lingue dell'Unione). http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-32-11-955

Eurostat (2007) Handbook on Data Quality Assessment Methods and Tools (DatQAM), https:// https://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-HANDBOOK%20ON%20DATA%20QUALITY%20ASSESSMENT%20METHODS%20AND%20TOOLS%20%20I.pdf

Eurostat COncepts and DEfinitions Database (CODED) http://tinyurl.com/ESSQualityGlossaryinCODED

FCSM (2001) "Measuring and Reporting Sources of Error in Surveys". Federal Committee on Statistical Methodology, Statistical Policy Working Paper 31

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

Hidiroglou MA, Drew DJ, Gray GB (1993) "A Framework for Measuring and Reducing Nonresponse in Surveys". Survey Methodology, 19, 1, pp. 81-94

Istat (2012) Linee guida per la qualità dei processi statistici , Versione 1.1, Dicembre 2012 – Istat –Roma http://www.istat.it/it/files/2010/09/Linee-Guida-Qualit%C3%A0-_v.1.1_IT.pdf

Istat (2016) Linee guida per la qualità dei processi statistici che utilizzano dati amministrativi , Versione 1.1, Agosto 2016 – Istat –Roma http://www.istat.it/it/files/2010/09/Linee-Guida-fonte-amministrativa-v1.1.pdf

Leadership Expert Group on Quality (2001) Summary Report from the Leadership Group (LEG) on Quality (Luglio, 2001) "https://www.istat.it/it/files/2011/11/LEG-on-quality.pdf"

Lessler, J., and Kalsbeek, W. (1992) *Nonsampling Errors in Surveys*. Wiley, New York.

OECD Glossary of Statistical terms https://stats.oecd.org/glossary/index.htm

Särndal C. E., Swensson, B., Wretman, L. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Statistics Canada (2010) *Survey Methods and Practices.* Statistics Canada, Catalogue no. 12-587-X, Ottawa. http://www.statcan.gc.ca/pub/12-587-x/12-587-x2003001-eng.htm

UNECE (2013) Generic Statistical Business Process Model (GSBPM) (Ver 5.0., December 2013) http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0

UNECE (2016) Version 1.0 of the Quality Indicators for the Generic Statistical Business Process Model (GSBPM) - For Statistics derived from Surveys, May 2016

UNECE (2009), Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes, Geneva, 2009

Appendice A. Definizioni Eurostat sulla qualità delle statistiche

Pertinenza
La pertinenza è definita come il grado in 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.

Accuratezza
L'accuratezza dei risultati viene definita dal punto di vista statistico come il grado di vicinanza tra le stime e i corrispondenti valori veri.

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 definita come il periodo di tempo che intercorre tra la data di rilascio dei dati e la data di rilascio programmata, quest'ultima può essere annunciata dal calendario ufficiale di diffusione, stabilita da un Regolamento oppure frutto di un accordo preventivo tra partner.

Coerenza e comparabilità
La coerenza tra due o più statistiche si riferisce a quanto i processi statistici che le hanno prodotte hanno utilizzato i medesimi concetti – classificazioni, definizioni e popolazioni obiettivo – e metodi armonizzati. Statistiche coerenti possono essere correttamente combinate e usate congiuntamente. Esempi di uso congiunto si hanno quando le statistiche fanno riferimento alla stessa popolazione, periodo di riferimento e regione, ma comprendono differenti gruppi di variabili (es. dati sull'occupazione e dati sulla produzione) o quando comprendono le stesse variabili (es. dati sull'occupazione) ma per diversi periodi, regioni o altri domini. Si definisce coerenza: i) tra domini: le statistiche sono riconciliabili con quelle ottenute attraverso altre fonti o domini statistici; ii) tra statistiche annuali e infra-annuali: statistiche di diversa periodicità sono riconciliabili; iii) con la Contabilità Nazionale: le statistiche sono riconciliabili con i conti nazionali; iv) interna: le statistiche sono consistenti all'interno di un certo *dataset*.

La comparabilità è una misura dell'impatto delle differenze nei concetti statistici adottati e negli strumenti/procedure di misurazione quando si confrontano le statistiche tra aree geografiche e nel tempo. È considerata come un caso particolare della coerenza e si riferisce al caso in cui le statistiche fanno riferimento alle stesse variabili e vengono combinate per fare confronti nel tempo, tra regioni o tra altri tipi di domini.

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, file, 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.

Appendice B. Alcuni Indicatori per la valutazione della qualità delle fonti amministrative

La qualità di un processo che utilizza in modo prevalente fonti amministrative dipende, fra i vari aspetti, anche dalla qualità dell'input stesso. Di seguito verranno riportati alcuni esempi di indicatori atti alla valutazione della qualità dei dati delle fonti amministrative, tratti dal progetto "Blue-ETS" (Daas P. e Ossen S, 2011), che risultano da una collaborazione tra diversi Istituti nazionali di statistica europei per il miglioramento della qualità delle statistiche economiche, ma applicabili a qualsiasi ambito in cui vi sia un impiego prevalente di fonti amministrative.

In particolare gli indicatori presentati riguardano: i controlli tecnici, l'accuratezza, la completezza, l'integrabilità e gli aspetti temporali. Per ragioni di spazio non è qui possibile presentare tutti gli indicatori elaborati nell'ambito del progetto Blue-ETS, pertanto, per ulteriori approfondimenti è possibile consultare il sito dedicato al progetto stesso (https://www.blue-ets.istat.it/) e in particolare il *Deliverable* 4.2 all'interno della sezione "*Deliverables and results*".

Nel seguito si parlerà di unità, tuttavia è necessario ricordare che, nei data set amministrativi, si fa riferimento ad oggetti che possono coincidere con unità oppure con eventi.

Controlli tecnici

Leggibilità. Sono indicatori dell'accessibilità dei file di dati.

- Percentuale dei file con un'estensione sconosciuta, corrotti o che per qualsiasi motivo sia impossibile aprire.
- Per ciascun file, percentuale del file non leggibile, in termini di GB/MB o numero di record.

Accuratezza

Autenticità delle unità. Questo aspetto si riferisce alla legittimità dei record nella fonte. L'eventuale illegittimità può derivare sia da una mancata corrispondenza del record con un'unità reale, sia a un'errata assegnazione della chiave identificativa e di conseguenza una corrispondenza del record con un'unità reale diversa da quella legittima. Per quest'ultima fonte di errore è necessario avere a disposizione una lista di riferimento completa.

- Percentuale di unità con identificativi non validi sintatticamente.
- Percentuale di unità con informazioni contraddittorie rispetto alla lista di riferimento.

Errore di misura. Questo errore si riferisce alla discrepanza tra i valori acquisiti delle variabili e i valori veri che avrebbero dovuto essere misurati. A volte è possibile che valori errati siano evidenziati dai titolari stessi della fonte, ma più spesso per valori sospetti si deve provare a risalire alla fonte di errore con l'aiuto del fornitore. Per individuare possibili valori sospetti è utile verificare la presenza di valori incoerenti nel file e fare un confronto con altre fonti contenenti dati simili, ove possibile.

In particolare, per valutare l'entità dei dati sospetti o incoerenti fra loro, è possibile ricorrere ai seguenti indicatori:

- percentuale di record per i quali la combinazione dei valori dei dati fornisce risultati privi di logica o incoerenti.
- Percentuale di record per i quali la combinazione dei valori dei dati fornisce risultati sospetti o improbabili, ma non necessariamente errati.

Completezza

Sottocopertura. È un indicatore che si riferisce all'assenza di unità che avrebbero dovuto essere incluse nella fonte. Per valutarla è necessario avere qualche conoscenza della popolazione obiettivo nella sua interezza, per esempio una lista di riferimento. Se ciò non fosse disponibile è opportuno procedere con studi per la sua costruzione.

- Percentuale di unità nella lista di riferimento assenti nel file di dati.

Sovracopertura. Si riferisce alla presenza all'interno della fonte di unità che non avrebbero dovuto esservi incluse.

- Percentuale di unità presenti nel file di dati non incluse nella popolazione di riferimento

Ridondanza. Sono indicatori che trattano la duplicazione non necessaria di unità all'interno del file di dati.

- Percentuale di unità all'interno del file con lo stesso identificativo.
- Percentuale di unità all'interno del file con gli stessi valori per un sottoinsieme scelto di variabili (che risultano essere duplicati).
- Percentuale di unità all'interno del file con gli stessi valori per tutte le variabili (che risultano essere duplicati).

Valori mancanti. Con questo indicatore si vuole analizzare la presenza di dati mancanti dal punto di vista delle variabili nella fonte, sia in riferimento a un'unica variabile che a una combinazione delle stesse. Per una valutazione migliore di questo indicatore è necessario che nel tempo sia stata sviluppata una comprensione approfondita dei dati.

- Percentuale di unità con valori mancanti per una particolare variabile.
- Percentuale di unità con esclusivamente valori mancanti per un sottoinsieme scelto di variabili.

È anche possibile utilizzare metodi grafici per individuare valori mancanti nelle variabili.

Integrabilità

Confrontabilità delle unità. L'indicatore è utile per un confronto fra le unità della fonte oggetto di valutazione e le unità presenti in altre fonti utilizzate dall'ente produttore, ai fini della loro integrabilità. L'indicatore va calcolato prima di effettuare le operazioni di integrazione.

- Percentuale di unità all'interno della fonte con definizioni analoghe a quelle già utilizzate dall'ente.
- Percentuale di unità all'interno della fonte che, dopo un'operazione di armonizzazione, corrisponderebbero a definizioni utilizzate dall'ente.

Confrontabilità delle variabili. Analogamente a quanto fatto per le unità, con questo indicatore si propone una valutazione della vicinanza tra le variabili inserite nella fonte oggetto di analisi e quelle di altre fonti utilizzate dall'ente.

- Percentuale di unità nelle due fonti che presentano lo stesso valore per una variabile oggetto di studio.

Aspetti temporali

Tempestività. Questo indicatore si riferisce all'intervallo di tempo tra la fine del periodo di riferimento dei dati e il momento di ricezione della fonte di dati. In alternativa, se l'ente in carico del processo è anche il titolare della fonte, si può sostituire la ricezione della fonte con la data in cui essa è resa disponibile agli utenti (se ciò accade).

- Differenza in giorni tra la data di ricezione della fonte di dati e la fine del periodo di riferimento dei dati.
- Differenza in giorni tra la data in cui la fonte è accessibile agli utenti e la fine del periodo di riferimento dei dati.

Puntualità. È un indicatore usato per valutare la distanza tra la data effettiva di ricezione della fonte e la data in cui essa sarebbe dovuta essere fornita da contratto.

- Differenza in giorni tra la data effettiva di ricezione della fonte e la data in cui essa sarebbe dovuta essere fornita da contratto.

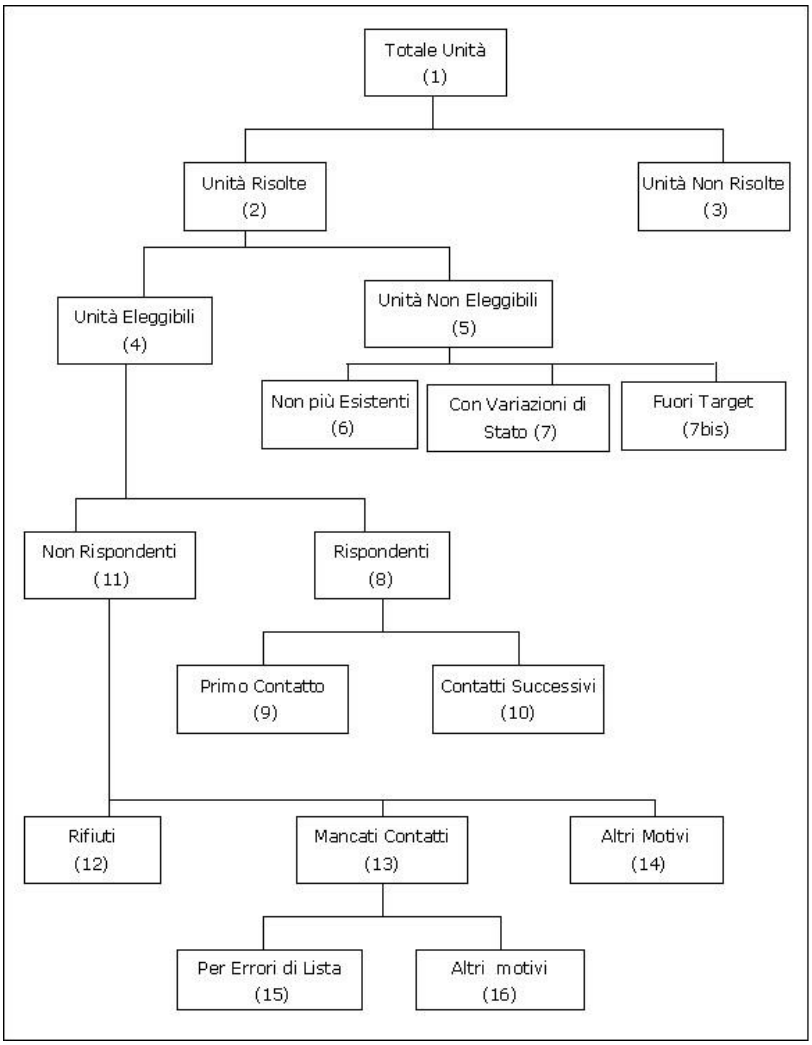
Dinamica delle unità. Questo indicatore serve per valutare l'utilità della fonte in relazione alla sua capacità di catturare i cambiamenti delle unità nel tempo. I cambiamenti principali sono dovuti all'ingresso di nuove unità nella popolazione ("nascite") e alla loro uscita ("morti"). Trattandosi di movimenti la cui registrazione è di vitale importanza per l'accuratezza della fonte scelta, la valutazione della capacità della fonte di seguirli nel tempo andrebbe valutata anche con un ritorno sul campo che si concentri sulle unità di nuovo ingresso o recente uscita.

Indicando con *t-1* e *t* due momenti consecutivi nel tempo nell'unità temporale scelta, possono essere calcolati i seguenti valori:

- percentuale delle nascite al tempo *t* = (nascite *t* / totale unità *t*) x 100
- percentuale delle morti al tempo *t* = (morti *t* / totale unità *t*) x 100
- percentuale delle morti al tempo *t-1* = (morti *t* / totale unità *t-1*) x 100

Appendice C. Schema di classificazione delle unità per il calcolo di indicatori di Copertura e Mancata risposta Totale

L'Istat adotta la seguente classificazione delle unità, che è generalizzabile ad qualsiasi processo di rilevazione:



Di seguito, le definizioni della casistica rappresentata

Totale Unità (1): numero complessivo delle unità oggetto di indagine. Per le indagini campionarie coincide con il numero di unità campionate

Unità Risolte (2): un'unità è risolta se è stato possibile accertare se era eleggibile

Unità Eleggibili (4): un'unità è eleggibile se appartiene alla popolazione oggetto di indagine

Unità Non Eleggibili (5): l'unità non appartiene alla popolazione oggetto di indagine pur essendo presente nell'archivio o lista di estrazione

Unità Non più Esistenti (6): l'unità non esiste pur essendo presente nell'archivio o nelle liste di estrazione per mancato aggiornamento, ritardo nell'aggiornamento o per un errore di inclusione

Unità con Variazioni di Stato (7): l'unità ha modificato il suo stato in modo tale da non essere più eleggibile per l'indagine (come trasferimento di residenza all'estero, cambiamento di attività economica o di numero di addetti)

Unità Fuori Target (7bis): l'unità non appartenente alla popolazione obiettivo; si tratta di una errata inclusione nella lista di riferimento

Rispondenti (8): unità di rilevazione per la quale è stato possibile rilevare l'informazione

Rispondenti al Primo Contatto (9): unità di rilevazione per la quale è stato possibile rilevare l'informazione al primo contatto

Rispondenti ai Contatti Successivi (10): unità di rilevazione per la quale è stato possibile rilevare l'informazione solo dopo più di un contatto

Non Rispondenti (11): unità di rilevazione per la quale non è stato possibile rilevare l'informazione

Non Rispondenti per Rifiuto (12): l'unità è eleggibile ma si rifiuta di partecipare all'indagine

Non Rispondenti per Mancato Contatto (13): unità eleggibile che non è stato possibile contattare per errori nella lista o per altri motivi

Non Rispondenti per Altri Motivi (14): unità eleggibile, contattata ma che non è stata in grado di fornire le informazioni richieste (malato, anziano, assenza del titolare dell'impresa)

Unità Non Contattate per Altri Motivi (16): l'unità è eleggibile ma non si è riusciti a contattarla per irreperibilità o altro (famiglia in vacanza, nessuna risposta al telefono)

Unità Non Contattate per Errori di Lista (15): l'unità è eleggibile ma non ha partecipato all'indagine perché non si è riusciti a contattarla per imprecisioni o informazioni insufficienti nella lista (come indirizzo errato)

Appendice D. Normativa sui dati personali e tutela della riservatezza

La legge istitutiva del Sistema statistico nazionale (d.lgs. n. 322 del 1989 e successive integrazioni, e modifiche di razionalizzazione del DPR. n. 166/2010), prevede che debbano essere adottate tutte le misure a garanzia della riservatezza dei rispondenti. Ulteriori principi in materia di tutela della riservatezza dei dati sono stabiliti dal "Codice di deontologia e di buona condotta per i trattamenti di dati personali a scopi statistici e di ricerca scientifica effettuati nell'ambito del Sistema statistico nazionale" (d.lgs. n. 196/2003).

Recentemente, è stato pubblicato nella Gazzetta Ufficiale Europea il Regolamento Europeo in materia di protezione dei dati personali[16]**.** Il nuovo Regolamento mira a garantire una disciplina sulla protezione dei dati personali uniforme ed omogenea in tutta la UE, al fine di assicurare un livello coerente ed elevato di protezione e rimuovere gli ostacoli alla circolazione dei dati personali all'interno dell'Unione Europea. Esso è immediatamente applicabile senza necessità di recepimento con atti nazionali (http://www.garanteprivacy.it/regolamentoue). La sua entrata in vigore, a decorrere dal 25 maggio 2018, impone ulteriori elementi di attenzione su vari aspetti tra cui le misure da prendere in caso di *data breach*, l'istituzione del data *protection officer* (responsabile della protezione dei dati).

Raccolta dati

Nella fase di raccolta dei dati, un'attenzione particolare va posta ai dati personali[17] e a quelli sensibili e giudiziari[18], che devono essere trattati nel rispetto della vigente normativa in materia di protezione dei dati personali (allegato A.3 del già citato Codice di deontologia). Tali dati possono essere utilizzati dal titolare del lavoro anche per ulteriori scopi statistici, in conformità all'art. 6-bis del decreto legislativo n. 322 del 1989. Attualmente la normativa non consente l'acquisizione di dati sensibili e giudiziari presso i rispondenti o presso soggetti pubblici e privati che li detengono, senza il consenso esplicito dell'interessato cui i dati si riferiscono, a meno che questo non sia previsto da espressa disposizione normativa.

Trattamento

Il trattamento dei dati sensibili e giudiziari è possibile solo nell'ambito di un lavoro statistico previsto nel Programma statistico nazionale. Diversamente per i dati personali non sensibili e giudiziari, raccolti presso un soggetto diverso dall'interessato (imprese, istituzioni), il trattamento dei dati è possibile purché sia resa l'informativa attraverso il Programma statistico nazionale o con idonee modalità (internet, stampa ecc.) da comunicare preventivamente al Garante.

Conservazione

Inoltre, il "Codice di deontologia e di buona condotta" regola gli aspetti relativi alla conservazione dei dati personali. Questi possono essere conservati anche dopo la conclusione del lavoro statistico per cui sono stati raccolti qualora siano necessari per ulteriori trattamenti statistici del titolare. Diversamente, i dati identificativi possono essere conservati, invece, solo se ricorre una delle ipotesi previste all'art. 11 dello stesso codice, quindi fino a quando risultino necessari per: a) indagini continue e longitudinali, b) indagini di controllo, di qualità e copertura, c) definizione di disegni campionari e selezione di unità di rilevazione, d) costituzione di archivi delle unità statistiche e di sistemi informativi, e) altri casi in cui risulti essenziale e adeguatamente documentato per le finalità perseguite.

[16] Regolamento UE n. 2016/679, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali nonché alla libera circolazione di tali dati e abroga la direttiva 95/46/CE (Regolamento generale sulla protezione dei dati).

[17] Ai fini della qualificazione della natura personale dei dati si precisa che i "dati personali" possono riguardare sia gli individui e le famiglie, che costituiscono le unità di analisi tipiche delle statistiche sociali e demografiche, sia le persone fisiche che svolgono, ad esempio, attività d'impresa e che, in quanto tali, rientrano invece, insieme ad altre categorie di soggetti, nel campo di osservazione proprio delle statistiche economiche.

[18] Per le definizioni di dato personale, dato sensibile e giudiziario si veda il glossario.

Qualora siano conservati, i dati identificativi devono essere custoditi separatamente da ogni altro dato salvo che ciò, in base a un atto motivato per iscritto, risulti impossibile in ragione delle particolari caratteristiche del trattamento o comporti l'impiego di mezzi manifestamente sproporzionati. Infine, i dati idonei a rivelare lo stato di salute e la vita sessuale devono essere conservati separatamente da altri dati personali trattati per finalità che non richiedono il loro utilizzo (d.lgs. n.196 del 2003, art. 22, comma 7).

Diffusione

La già citata normativa regolamenta la materia anche in relazione alla diffusione. In particolare, il Codice di deontologia definisce il concetto di identificabilità di un'unità statistica, mediante l'uso di mezzi ragionevoli, ovvero la possibilità di stabilire un'associazione significativamente probabile tra la combinazione delle modalità delle variabili relative all'unità statistica e i dati identificativi della medesima. Nella definizione di informazioni "riservate" rientrano anche i dati personali, inclusi quelli sensibili e i dati giudiziari (così come definiti all'art. 4 d.lgs. n. 196).

Tuttavia, il Codice di deontologia del Sistan (l'art. 4, comma 2), contempla la possibilità di diffondere variabili in forma disaggregata qualora ciò "risulti necessario per soddisfare particolari esigenze conoscitive anche di carattere internazionale e comunitario", purché tali variabili non siano idonee a rivelare informazioni sensibili e giudiziarie. Ciò era previsto anche nell'art.13 comma 3-bis del d. lgs. 322 del 1989 ai sensi del quale, in deroga ai limiti posti dalla disciplina in materia di segreto statistico, è prevista la possibilità di diffondere variabili in forma disaggregata indipendentemente dalla natura personale dei dati (quindi anche per le imprese).