rivista di statistica ufficiale

n. 3
2015

Temi trattati

Istat
Istituto Nazionale di Statistica

rivista di statistica ufficiale

n. 3
2015

Temi trattati

Istat

Per contattare la redazione o per inviare lavori scrivere a:
Segreteria del Comitato di redazione della Rivista di Statistica Ufficiale
All'attenzione di Gilda Sonetti
Istat – Via Cesare Balbo, 16 – 00184 Roma
e-mail: rivista@istat.it

rivista di statistica ufficiale

n. 3/2015

Periodico quadrimestrale
ISSN 1828-1982

Registrato presso il Tribunale di Roma
n. 339 del 19 luglio 2007

Istituto nazionale di statistica
Via Cesare Balbo, 16 – Roma

Analisi e misurazione dei processi di supporto[1]

Dario Russo,[2] Piero Demetrio Falorsi[3]

Sommario

Non esiste una ricetta universale per la progettazione ottimale dei processi di supporto. È necessario agire caso per caso con riferimento alle caratteristiche specifiche di un'organizzazione, utilizzando diverse leve di intervento: un'attenta progettazione organizzativa, un sistema efficace di regole e controlli, una politica di approvvigionamento equilibrato e un uso intelligente della tecnologia. Questo articolo descrive le fasi principali di un progetto per migliorare l'efficienza dei processi di supporto che parte da una specifica analisi del contesto in cui si muove l'organizzazione. Per esemplificare l'approccio, si considera come caso di studio un progetto pilota, svolto in un istituto finanziario in cui ci si focalizza sull'analisi e la riconfigurazione del processo di back office per mezzo di un approccio di simulazione.

Parole chiave: Analisi Lean, Modelli guidati da eventi, Tassonomia di modelli organizzativi, Simulazione di processi.

Abstract

There is not a universal framework for the optimal design of support processes: it is necessary to tailor the strategy on a case by case basis and with reference to the specific characteristics of the company. It is useful to adopt a strategy based on different leverages: a careful organizational design, an effective system of rules and checks, a balanced sourcing policy, an intelligent use of technology.
This article describes the main steps of a project to improve the efficiency of the support processes recalling certain models of investigation and with reference to direct experiences carried out in the banking sector. The analysis and redesign process moves from the evaluation of the activities of the institutions and it is focused on the efficiency of the support processes.

Keywords: Lean Analysis, Event-Driven Models, Organizational Processes Taxonomy, Process Simulation.

[1] Le opinioni espresse in quest'articolo ricadono nell'esclusiva responsabilità dell'autore e non riflettono necessariamente le posizioni delle istituzioni di appartenenza.

[2] Banca d'Italia *e-mail:* dario.russo@bancaditalia.it.
[3] Istat, *e-mail:* falorsi@istat.it.

1. Introduzione

Il presente lavoro ha lo scopo di proporre un approccio integrato di analisi e progettazione organizzativa finalizzato all'aumento di efficienza dei *processi di supporto*. Le considerazioni qui esposte hanno una valenza generale e possono trovare applicazione in qualsiasi realtà aziendale.

Tuttavia, in questo lavoro, l'attenzione è stata focalizzata su quegli enti (nel seguito indicati sinteticamente come *Enti*) il cui scopo principale è quello di fornire al pubblico servizi di tipo *non market*. *Enti* di questo tipo sono, ad esempio, quelli appartenenti alla Pubblica Amministrazione (PA), le organizzazioni *no-profit*, le agenzie di controllo e regolatrici del mercato, ecc.

I processi di supporto non costituiscono l'obiettivo principale dell'azione degli *Enti*, ma rappresentano un *input* necessario e imprescindibile al raggiungimento degli obiettivi istituzionali. Esempi di processi di supporto sono quelli per la gestione del personale, i servizi informatici non esposti al pubblico[4].

In un quadro di tagli crescenti ai *budget* pubblici, il tema del recupero di efficienza, diventa centrale. I processi di supporto, che non sono direttamente collegati alle finalità istituzionali, sono i candidati naturali per il taglio dei costi.

Tuttavia, azioni non ben meditate possono incidere in modo significativo anche sui servizi istituzionali che gli *Enti* erogano al pubblico; per non incorrere in disfunzioni e mitigare i rischi, diventa quindi necessario dotarsi di approcci organizzativi e metodologici che consentano di migliorare l'efficienza, senza incidere sulla qualità dei servizi forniti.

I modelli e gli approcci di analisi organizzativa orientati al recupero dell'efficienza nascono in un contesto industriale maturo[5] (ad esempio, nell'industria dell'auto, ma anche nel settore informatico) e solo recentemente questi approcci sono stati proposti anche per il settore *non market*[6] (in particolare per gli enti pubblici) che, rispetto a una realtà industriale matura presenta una serie di particolarità (e/o complessità) come ad esempio:

- ✓ vi è una difficoltà oggettiva (anche di tipo concettuale) a definire il valore di mercato dei servizi prodotti, sia quelli esterni che quelli di supporto;
- ✓ non viene sempre adottata con rigore una contabilità analitica e risulta spesso molto complesso valorizzare i costi delle differenti tipologie di servizio;
- ✓ il modello organizzativo dell'*Ente* è sovente sviluppato per rispondere a normative e regolamentazioni e non è di tipo *output oriented*;
- ✓ i modelli organizzativi sono rigidi, difficili da cambiare e costituiscono il risultato di stratificazioni successive di interventi realizzati in tempi diversi per rispondere a esigenze via via mutate;
- ✓ la cultura di un moderno approccio al *management* è spesso poco diffusa tra il personale dirigente sia di alto che di medio livello.

A causa dei fattori sopra elencati, qualsiasi progetto di miglioramento dell'efficienza può essere realizzato con successo unicamente adottando un metodo di lavoro *olistico* (che tenga conto dei diversi fattori coinvolti direttamente o indirettamente nelle azioni da intraprendere), *misurabile* (ossia *guidato* dai risultati raggiunti) e *flessibile* (adattabile a

[4] Cfr. *infra* §3.
[5] Cfr. Holweg e Matthias, (2007), Bailey e David (2008), Ohno e Taiichi (1988).
[6] Cfr. Radnor *et al.* (2008 e 2010).

contesti organizzativi complessi e mutevoli).

L'articolo ha l'obiettivo di dimostrare che l'analisi dei processi supportata da strumenti di simulazione, permette di individuare interventi organizzativi in grado di determinare consistenti recuperi di efficienza.

Il seguito del lavoro è articolato nel modo seguente: nel paragrafo 2 si riporta un'analisi della collocazione organizzativa dei processi di supporto; nel paragrafo 3 sono illustrate le macrofasi di un processo di miglioramento organizzativo e si focalizza l'attenzione sugli strumenti metodologici, la cui applicazione in un *case study* è descritta nel paragrafo 4; infine, nel paragrafo 5 sono sinteticamente riportate alcune conclusioni.

2. La collocazione organizzativa dei processi di supporto

Prima di affrontare gli aspetti più metodologici connessi ai processi di miglioramento dell'efficienza è opportuno fornire un quadro di come gli Enti, in genere, strutturano i servizi supporto, nel seguito anche indicati come servizi interni. L'organigramma di un Ente comprende unità organizzative dedicate alla fornitura dei servizi interni e unità dedicate alle funzioni istituzionali. Specialmente nelle Pubbliche Amministrazioni, sovente le prime unità sono più numerose delle seconde.

Nonostante, in termini quantitativi, le risorse impegnate nei processi di supporto non superino di norma il 30% del totale[7], le strutture organizzative sottostanti arrivano sovente a rappresentare più del 50% dell'organigramma. Questo è dovuto alla morfologia del «sistema dei servizi» nella maggior parte dei casi articolato in tre livelli. Il primo è costituito dai fornitori primari di servizi (gestione del personale, procurement, logistica, Information Technology, ecc.) che servono l'intera organizzazione; essi detengono la responsabilità dei servizi di propria competenza e decidono l'articolazione dei relativi processi, definendone le regole e il sistema dei controlli e dimensionandone le risorse. Il secondo livello è rappresentato dai centri servizi locali, entità distribuite nelle diverse aree di business (e sovente nelle diverse location geografiche dell'organizzazione) che offrono una gamma più o meno completa di servizi a un bacino di utenza circoscritto. Il terzo livello sono gli staff e le segreterie di Direzione, che svolgono prevalentemente un ruolo di collegamento fra il top /middle management e i centri servizi. In alcuni casi, tali strutture erogano direttamente porzioni di servizio alle strutture di cui fanno parte, sovrapponendosi ai centri servizi di secondo livello.

Il modello sopra descritto è complesso e presenta ridondanze e sprechi.

Un programma di miglioramento dell'efficienza deve muoversi lungo le seguenti linee guida:

- ✓ l'innovazione della gestione aziendale deve fare ampio ricorso alla tecnologia;
- ✓ l'orizzonte temporale non deve mai essere di breve termine; gli interventi devono essere articolati in più annualità (in genere un triennio);

[7] Non esiste alla data un'indagine completa sulla morfologia delle strutture di servizio nelle aziende. Per il sistema bancario indicazioni possono essere tratte dai rapporti della CIPA [cit.] e dell'ABI [cit.]. Per la PP.AA. alcune informazioni (seppure parziali) si trovano nelle pubblicazioni di AgID (già DigitPA, CNIPA, AIPA) e del Dipartimento della Funzione Pubblica. La stima effettuata dagli autori è basata anche sulle informazioni fornite da alcune importanti società di consulenza strategica impegnate in progetti di riorganizzazione di Enti pubblici.

- ✓ le possibilità di erogare i servizi attraverso fornitori esterni (outsourcing) devono essere investigate in modo approfondito;
- ✓ il contenimento dei costi operativi deve essere accompagnato da un'attenzione alla qualità dei servizi resi e dal monitoraggio dei rischi;
- ✓ considerando, in particolare, le strutture di servizio di secondo livello, ci si deve muovere dall'assioma che l'interposizione di una struttura fra il cliente e il fornitore del servizio deve sempre giustificarsi dall'aggiunta di valore al processo e mai deve essere un mero e improduttivo allungamento della catena (il cosiddetto effetto passacarte);
- ✓ le strutture di secondo livello devono quindi porsi come traguardo finale la loro completa trasformazione in moderni ed efficienti centri servizi orientati, da una parte, a conseguire la massima economicità di funzionamento (ottimizzazione delle risorse), dall'altra, a offrire agli utenti/clienti la qualità attesa. Se queste condizioni non sono verificate, la struttura di secondo livello deve essere soppressa.

3. Il processo di miglioramento organizzativo

3.1 Le macrofasi

Un programma di miglioramento dell'efficienza deve essere attuato per fasi successive, le cui principali sono:

- ✓ *ricognizione dei processi di supporto,*
- ✓ *analisi dei processi,*
- ✓ *definizione del piano azione,*
- ✓ *misurazione dei risultati.*

La *ricognizione dei processi di supporto* ha lo scopo di ricostruirne la tassonomia.

Spesso gli *Enti* si rappresentano al loro interno con viste organizzative differenti e, talvolta non congruenti. Di conseguenza, la tassonomia dei processi deve essere realizzata prendendo in considerazione tutte le classificazioni elaborate in precedenza per scopi diversi (ad esempio: sistema di contabilità analitica e controllo di gestione, sistema di governo dei rischi operativi - *Operational Risk Management*, sistema di rilevazione della produttività aziendale). I processi devono essere descritti individuandone le fasi, gli *input* e gli *output*, e determinandone le risorse assorbite, le quantità dei prodotti erogati e i principali fattori di qualità misurata e percepita.

L'analisi dei processi è orientata a determinare i fattori d'inefficienza e a progettare gli interventi correttivi di vario tipo: organizzativi, normativi, tecnologici.

L'insieme degli interventi confluisce nel *Piano di azione* la cui attuazione è demandata a piccoli gruppi operativi con il supporto di un *team* di progetto e sotto il diretto controllo del *Management*.

L'ultima fase è rappresentata dalla *misurazione dei risultati* che può anche fornire indicazioni per un parziale aggiornamento del modello dei processi.

La struttura delle fasi del programma sopra descritta rappresenta una versione

semplificata del modello DMAIC (*Define, Measure, Analyze, Improve, Control*) dell'approccio *Lean/Six Sigma*[8] che utilizza strumenti di tipo statistico, mutuati dalle tecnologie per il controllo della qualità e finalizzato all'implementazione e alla gestione operativa del *Total Quality Management*.

Ogni iniziativa di miglioramento ha bisogno di un meccanismo di *feedback* e di controllo (fase *Control*) per assicurare che non si torni nella situazione precedente al cambiamento introdotto o che le innovazioni apportate non producano effetti indesiderati e/o erratici. In questa fase si esegue un monitoraggio sull'impatto delle modifiche apportate.

3.2 Gli strumenti metodologici

I tre principali strumenti metodologici da utilizzare in un programma di miglioramento dell'efficienza dei servizi di supporto di un Ente sono:

1) il censimento dei processi per definirne la tassonomia e descriverne la struttura;
2) l'approccio *lean* per la riprogettazione dei processi;
3) la simulazione dinamica per la verifica delle soluzioni.

Essi trovano applicazione nelle diverse macrofasi del processo di miglioramento descritte nel precedente paragrafo, come rappresentato nella Tavola n.1.

Tavola n.1 – relazione strumenti/macrofasi

strumenti / macrofasi	*censimento dei processi*	*approccio lean*	*simulazione dinamica*
ricognizione dei processi	x		
analisi dei processi	x	X	x
definizione del piano azione		X	
misurazione dei risultati			x

3.2.1 Il modello dei processi: tassonomie a confronto

Il censimento dei servizi di supporto è complesso nell'ambito di un *Ente*. La difficoltà maggiore risiede nel fatto che possono coesistere viste organizzative parziali e non congruenti. Uno stesso servizio può essere visto da una visione organizzativa nella sua completezza, mentre da un'altra solo per una sua parte. I vocabolari e le terminologie possono differire.

Inoltre, i processi produttivi nella maggior parte dei casi sono sviluppati secondo logiche non integrate di tipo *stove pipe.* Non è infrequente che ciascun settore produttivo sviluppi i propri sistemi informativi locali e non faccia uso delle informazioni *corporate*.

[8] Cfr. Montgomery e Douglas C. (2009), Tennant e Geoff (2001). Il DMAIC si compone di cinque fasi: *Defining, Measuring, Analyzing, Improving, Controlling*.
Nella fase *Define* s'individua lo scopo del lavoro che si vuole svolgere, si determinano i miglioramenti da apportare al processo sotto esame e si fissano obiettivi realistici per quanto riguarda sia le tempistiche sia i costi condivisi con tutti gli *stakeholder*.
La fase *Measure* comprende la creazione di una mappa del processo *AS IS* e la raccolta di tutti i dati necessari per svolgere un lavoro efficace di analisi.
L'obiettivo della terza fase *Analyze* è verificare attraverso l'analisi dei dati se le cause identificate siano effettivamente quelle che hanno creato le criticità.
Lo scopo della fase *Improve* è progettare la soluzione adatta a risolvere il problema

Per superare le difficoltà di cui sopra, è necessario adottare un approccio basato sui seguenti capisaldi:

a) ci deve essere una forte e convinta *sponsorship* dell'operazione da parte del *Top Management* dell'*Ente*;
b) il consenso del *management* intermedio rappresenta un fattore chiave. Ciò ha due principali conseguenze: (b1) tutti gli *stakeholder* devono essere coinvolti nell'operazione; (b2) i risultati della ricognizione devono prevedere fasi di validazione e di discussione degli *output* intermedi;
c) la descrizione dei servizi deve essere realizzata adottando *framework* rigorosi e strutturati.

In particolare, la realizzazione del censimento dei processi è un passo necessario affinché l'*Ente* adotti un *modello unico* di azienda, integrando in una logica unitaria i *modelli diversi* sviluppati nel tempo, permettendo, altresì lo sviluppo di *modelli locali* che non confliggano con il modello unico e siano funzionali all'innovazione e all'efficienza organizzativa.

3.2.2 Il lean

Il *lean* è un approccio organizzativo ideato nel 1970 in Toyota[9] per gestire la produzione al ritmo degli ordini cliente e minimizzare gli sprechi (inefficienze) di produzione (*Just in Time).*

La parola chiave del *lean thinking* è MUDA (che in giapponese significa SPRECO). Si definisce spreco ogni attività umana che assorbe risorse senza creare valore, come ad esempio:

- beni e servizi che non incontrano i bisogni dei clienti;
- produzione al di là della richiesta;
- gruppi di persone che attendono perché un'attività precedente non è conclusa in tempo;
- scorte di componenti per sopperire ai *colli di bottiglia* produttivi;
- fasi di processo non necessarie;
- spostamenti di merci e/o personale senza scopo reale;
- produzione di pezzi con difetti (errori nelle fasi produttive) che richiedono rifacimenti e/o provocano scarti.

Il termine produzione snella (dall'inglese *lean manufacturing o lean production*) identifica una filosofia industriale, ispirata al Toyota Production System, che mira a minimizzare gli sprechi fino ad annullarli.

I principi *lean* sono:

- eliminare lo spreco;
- specificare precisamente il valore dalla prospettiva del cliente finale;
- identificare chiaramente il processo che consegna valore al cliente (cosiddetto value stream), ed eliminare le fasi che non aggiungono valore;
- svolgere le rimanenti fasi che aggiungono valore in un flusso senza interruzione, ottimizzandone le interfacce;

[9] Cfr. Ohno e Taiichi (1988).

- lasciare che sia il cliente a "tirare" il processo – non produrre niente fino a che non ce ne sia bisogno, poi produrre velocemente quanto richiesto;
- perseguire la perfezione tramite continui miglioramenti.

Le maggiori differenze dell'ambiente dei servizi rispetto a quello industriale, che influiscono sull'applicazione delle metodologie *lean*, si riferiscono principalmente alla natura *variabile* dei servizi:

- le richieste del cliente possono assumere sfaccettature diverse;
- la facilità di "copiatura" dei prodotti/servizi obbliga alla continua innovazione;
- lo svolgimento delle attività di erogazione dei prodotti/servizi dipende fortemente dalle valutazioni dei singoli;
- le reti di vendita e di erogazione hanno un'ampia distribuzione territoriale.

I criteri del *lean processing* sono comunque utilizzabili tenendo però presente che alcune misurazioni, e soprattutto la valutazione delle cause d'inefficienza, sono, di solito, meno *certe* che nel settore industriale.

L'applicazione nel concreto dell'approccio *lean* richiede un forte coinvolgimento della struttura organizzativa sulla quale s'interviene, a diversi livelli:

1) *Top management*: è richiesto un chiaro *committment* da parte dell'Alta Direzione che deve sponsorizzare in modo visibile l'iniziativa e soprattutto inquadrarla nella strategia generale dell'azienda.
2) *Middle management*: la Direzione locale deve essere coinvolta in modo diretto, partecipando allo *Steering Group* del progetto, verificandone i progressi e intervenendo per rimuovere vincoli o criticità.
3) *Supporto metodologico*: un gruppo di esperti deve supportare il progetto dal punto di vista metodologico per garantire la corretta applicazione dei metodi e delle teorie, curare le attività di *project management*, produrre la reportistica per il controllo di progetto e per il *management.* E' opportuno un pieno coinvolgimento della Funzione Organizzazione.
4) *Unità operative*: all'interno di ogni singola unità operativa elementare deve essere individuato un responsabile incaricato di coordinare le attività di raccolta dati, di analisi e di ricerca delle soluzioni nell'ambito dei processi (o sottoprocessi) di competenza. Tutto il personale deve essere informato del progetto, messo al corrente delle attività svolte e dei risultati conseguiti e coinvolto attraverso meccanismi che permettano di fornire idee e suggerimenti (*enterprise social network, collaboration tool*).

3.2.3. La simulazione dinamica

3.2.3.1. I simulatori

I modelli di simulazione (o simulatori) sono strumenti metodologici, nati soprattutto in ambito ingegneristico o chimico[10] che consentono di esplorare velocemente la situazione attuale (*AS-IS*) oppure i possibili sviluppi futuri (*TO-BE*) a fronte di possibili cambiamenti derivanti da decisioni o eventi esterni.

[10] Cfr. D'Amato, 1988.

Essi permettono di rappresentare, riprodurre e analizzare "in vitro" il funzionamento del sistema *AS IS* e di valutare a breve, medio e anche a lungo termine, le conseguenze delle decisioni applicate o dei cambiamenti che si vogliono apportare per la situazione *TO BE*; più in particolare, essi consentono di valutare *ex ante* le possibili conseguenze di determinate situazioni (eventi esterni, decisioni, circostanze possibili, ecc.) e di valutare l'efficacia delle decisioni, riducendo i rischi della fase d'implementazione operativa. La velocità dell'esplorazione è resa possibile dall'utilizzo di *software ad hoc* che implementano le prassi metodologiche.

I simulatori trovano utilizzo nel contesto della strategia, dell'organizzazione e dell'analisi dei processi, dove la complessità è fattore dominante. Essi sono applicati per la risoluzione di problemi sia strategici (analisi degli effetti delle politiche decisionali e valutazione dei rischi) sia operativi (pianificazione risorse e dimensionamento dei fabbisogni di un processo, prima e dopo una revisione; confronti di strutture organizzative).

La simulazione rappresenta quindi uno strumento veloce e flessibile - utilizzabile dal *management* aziendale - per verificare in brevissimo tempo gli effetti delle decisioni, prima di un successivo studio più dettagliato e specifico delle stesse; valutare i *savings* e il "ritorno" di eventuali investimenti.

Per un corretto utilizzo dei simulatori occorre tenere ben presente che non sono uno strumento operativo ma uno strumento esplorativo da utilizzare per ragionare e per verificare le diverse opzioni applicabili a un sistema complesso.

Un buon simulatore fornisce le indicazioni generali circa i possibili effetti delle decisioni e, volendo, anche alcuni aspetti di dettaglio (a seconda del livello di profondità richiesto); consente di navigare nella complessità per identificare "il percorso giusto"; consente una quantificazione iniziale di investimenti, costi, risparmi, livello di servizio, rischio.

In definitiva, la simulazione offre un approccio di grande valore per esplorare e valutare gli effetti di un cambiamento o di una decisione.

3.2.3.2. L'utilizzo in concreto dei simulatori

L'utilizzo in concreto dei simulatori necessita che l'organizzazione si doti di un metodo di lavoro strutturato affinché gli *input* con cui alimentare i modelli e le valutazioni sugli *output* siano sufficientemente supportati e robusti.

Le principali fasi dello schema di lavoro da adottare sono illustrate nella Figura 1.

Figura 1 - Metodo di lavoro per la simulazione dinamica

La costruzione di uno specifico modello di simulazione per un dato processo ha come *prerequisito* la raccolta degli *input* necessari, tra cui:

- ✓ la descrizione del *workflow* del processo;
- ✓ il calcolo del fabbisogno del personale e la definizione degli indici di utilizzo del personale (IUP) e dei tempi per l'espletamento dei servizi.

Naturalmente, la costruzione di un buon modello è un processo iterativo, che impara dall'esperienza, e che ha bisogno di un'attività costante di analisi e sull'efficacia, sull'efficienza delle scelte adottate e sui possibili sviluppi migliorativi che si possono introdurre.

4. Case study

Le aziende e le Pubbliche Amministrazioni eseguono ogni anno un notevole ammontare di pagamenti (prevalentemente, ma non esclusivamente, mediante bonifici). Tali pagamenti sono originati da attività interne all'Ente oppure effettuati (nel caso di alcune PP.AA.) per conto di altre PP.AA. o Istituzioni.

Il *Case Study* (CS) si riferisce alle fasi di *Back Office* del processo dei pagamenti. Si tratta di un'esperienza reale condotta nel 2012 in una Istituzione finanziaria italiana.

L'obiettivo del Case Study è: (i) verificare i margini di miglioramento del processo in termini di dimensionamento delle risorse, articolazione organizzativa delle strutture, livelli di servizio; (ii) determinare, mediante la simulazione, gli effetti di eventuali modifiche organizzative nelle diverse condizioni operative.

4.1 I processi di *Back Office* dei pagamenti

4.1.1 La definizione di Back Office

In assenza di una definizione *ortodossa* della funzione di *Back Office* (BO) la stessa analisi empirica sembra consigliare di astenersi dal tentativo di pervenire a un'accezione univoca, valida sempre e ovunque.

Si possono però individuare, come elementi che caratterizzano tale funzione, quelle attività deputate ad assicurare il perfezionamento e il controllo formale e sostanziale degli atti aventi rilevanza finanziaria che, compiuti da una distinta struttura (il *Front Office*), impegnano l'Ente giuridicamente e patrimonialmente.

Le attività di riscontro e regolamento svolte per ogni operazione, anche se con caratteristiche specifiche dovute alle peculiarità delle singole operazioni, possono essere schematizzate secondo il flusso operativo rappresentato nella Tavola n.2.

Tavola n.2 – Le attività di *Back Office*

riscontro dell'operazione nei sistemi interni e invio delle conferme alla controparte
spunta delle conferme provenienti dalla controparte
invio a regolamento, tramite movimentazione conti e/o *settlement* titoli
gestione anomalie e storni
controllo dell'adeguatezza dei *collateral* e dei margini
riconciliazione di tutti i conti movimentati
contabilizzazione delle operazioni e chiusura della giornata operativa

4.1.2 Il modello organizzativo del Case Study (CS)

Nel CS sono stati considerati i pagamenti di un Ente con riferimento sia a quelli domestici (eseguiti attraverso i sistemi di regolamento al dettaglio) sia quelli sull'estero.

I pagamenti sono disposti dalle strutture centrali e periferiche (es. filiali) dell'Ente attraverso un messaggio che contiene tutte le informazioni (nome del beneficiario, data di pagamento, banca beneficiaria, causale del pagamento, ecc.) necessarie per l'esecuzione.

Nel CS s'ipotizzano due Unità organizzative impegnate nel processo: la prima ("Unità di pagamento - UP") immette i dati dell'operazioni di pagamento nei sistemi informativi e ne avvia l'esecuzione dopo avere effettuato alcuni controlli, la seconda ("Unità contabile - UC") previa verifica della provvista, ne esegue la validazione.

Il *workflow* ipotizzato del processo prevede che le prime fasi di lavorazione siano di competenza della UP, mentre le fasi conclusive (*in primis* la contabilizzazione) siano di competenza della UC, che quindi inizia a lavorare sul singolo pagamento solo dopo che la prima Unità ha concluso il suo lavoro.

A fine giornata, la UP appronta per la firma della Direzione i prospetti riepilogativi delle operazioni effettuate e la UC effettua la quadratura contabile sia relativa alle

operazioni poste in essere e sia con riferimento alla complessiva attività del sistema di regolamento.

Nel CS s'ipotizza che le operazioni della specie siano 1.100 per ogni anno.

Le risorse nelle due Unità (pari a 33 addetti di cui 25 nella UP e 8 nella UC) lavorano su turni dalle 7,30 alle 19. S'ipotizzano tre tipi di addetti: Operatori, Specialisti, Coordinatori.

Esclusivamente i Coordinatori dell'UP autorizzano, con la propria firma sui relativi documenti, la trasmissione dei pagamenti all'UC per le successive fasi di lavorazione.

Le attività delle due Unità sono strettamente sequenziali.

Nel CS i pagamenti non sono l'unica attività svolta dalle due Unità, ma rivestono la massima priorità in quanto ogni pagamento trasmesso a inizio giornata deve essere completato entro un limite di tempo prefissato (h 17,30); se tale limite viene superato, l'Ente deve pagare una penale alla controparte.

Per questa ragione i pagamenti rappresentano la priorità assoluta delle due Unità, anche se, in condizioni normali, l'assorbimento di risorse si attesta intorno al 20%.

La variabilità del numero dei pagamenti elaborati ogni giorno è ipotizzata molto elevata. Generalmente circa il 50% dei bonifici è eseguito manualmente l'altro 50% è trasmesso attraverso sistemi informatici.

4.2. L'analisi del processo e le ipotesi evolutive

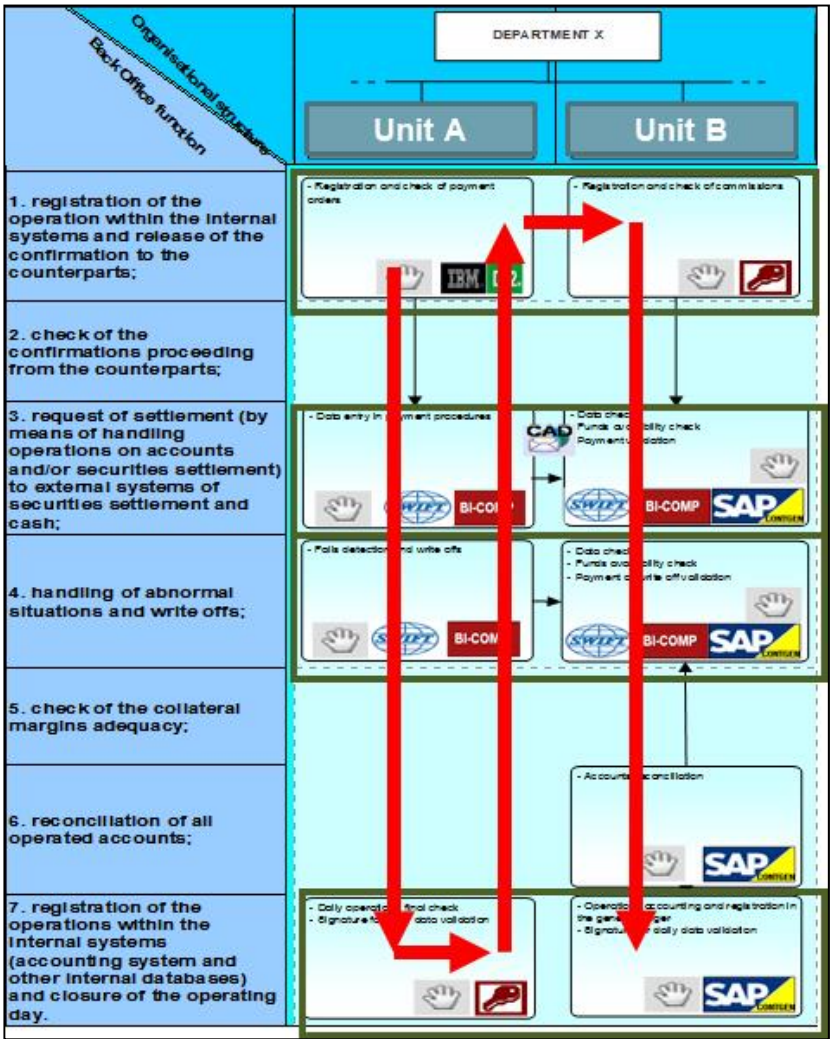
4.2.1. L'analisi del processo AS-IS

Il *workflow* del processo evidenzia le diverse fasi di lavorazione e il coinvolgimento delle due Unità. Dalla sua osservazione[11] (cfr. Figura 2) emerge che:

1) alcune sovrapposizioni nel lavoro delle due Unità portano alla duplicazione dei controlli[12];
2) la serializzazione delle operazione allunga e irrigidisce i tempi di lavorazione introducendo un fattore di rischio, tenuto conto del vincolo temporale;
3) in tutte le fasi sono previste lavorazioni di tipo manuale[13];
4) l'utilizzo di alcuni sistemi informativi è limitato, per la stessa attività, a una sola Unità organizzativa.

[11] Nell'analisi del processo è stato applicato l'approccio *lean* descritto nel paragrafo 3.2.2.
[12] I controlli sono eseguiti in tutti i passaggi di fase di lavorazione indicate con i riquadri verdi.
[13] Indicate con il simbolo della mano bianca. Quando le lavorazioni sono supportate da uno strumento informatico quest'ultimo è indicato con il relativo logo.

Figura 2 – Analisi del processo



4.2.2 La simulazione

Tra le possibili linee d'intervento, è stata presa in considerazione l'ipotesi di un accorpamento delle due Unità in una sola struttura. Per verificare l'impatto di un simile intervento sui tempi di lavorazione dei pagamenti, si è fatto ricorso a un modello di

simulazione[14] in grado di replicare il comportamento del sistema organizzativo riproducendone i meccanismi operativi e decisionali (es. poteri di firma). Attraverso il simulatore sono stati osservati:

- ✓ gli effetti dell'assenza di personale (valutando separatamente le conseguenze dell'assenza di diverse figure professionali) sui tempi di completamento del processo;
- ✓ gli effetti delle situazioni critiche (elevato numero di pagamenti da trattare in una giornata, elevata percentuale di pagamenti manuali) sui tempi di completamento e sul tasso di utilizzo del personale;
- ✓ i possibili vantaggi derivanti dall'accorpamento delle due Unità e/o dalla rimozione di vincoli normativi (es. estendere agli Specialisti il potere di firma riservato ai Coordinatori).

Nel modello sono stati inseriti i tempi del processo, le regole del *workflow* e le risorse. Nel cruscotto principale del simulatore sono stati inseriti cursori per variare il carico (quantità e tipologia di pagamenti elaborati) e uno "*switch*" per selezionare l'opzione "due Unità" o quella "una sola Unità".

4.2.3 I risultati della simulazione

Sono state simulate, fra tutte quelle possibili, le seguenti situazioni:

A. *AS-IS* / attività normale: si riproduce un giorno lavorativo nel quale le due Unità esistenti elaborano il numero medio di pagamenti.
B. *AS-IS* / assenza di personale: si riproduce un giorno lavorativo nel quale le due Unità elaborano il numero medio di pagamenti in assenza di una risorsa critica (coordinatore) per turno.
C. *AS-IS* / situazione critica: si riproduce un giorno lavorativo nel quale le due Unità esistenti elaborano un numero di pagamenti superiore del 300% a quello medio.
D. *TO-BE*: si riproduce un giorno lavorativo nel quale le due Unità sono accorpate in una.

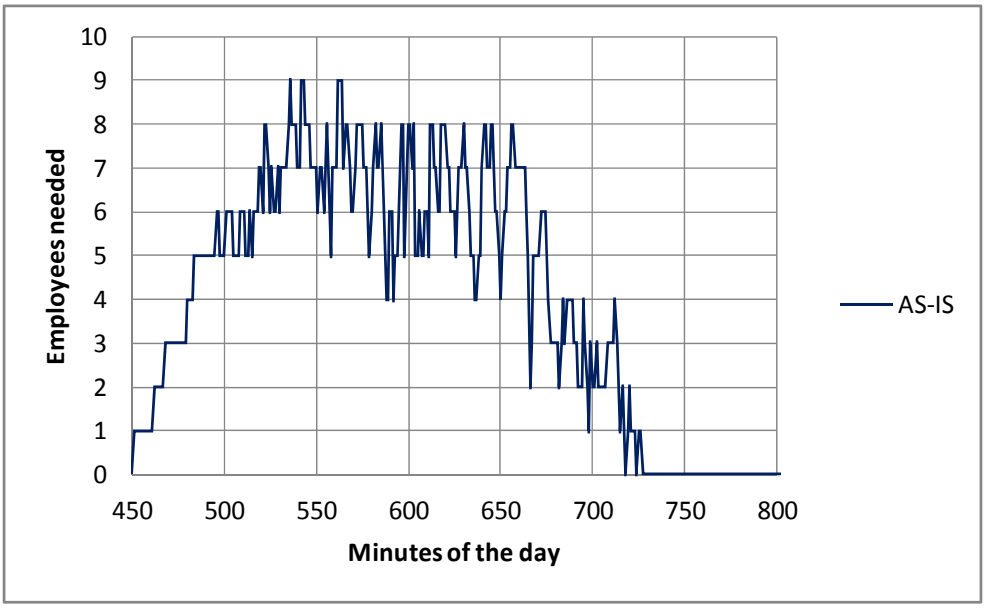
Le simulazioni sono rappresentate attraverso la curva di assorbimento delle risorse (sull'asse delle ordinate è rappresentato il numero di addetti richiesto dal processo) nella giornata lavorativa (sull'asse delle ascisse è rappresentato il tempo nell'intervallo compreso fra le h 7:30 e le h 19:30).

A. AS-IS attività normale

La situazione A (*AS-IS* / attività normale) riproduce un giorno lavorativo nel quale le due Unità esistenti elaborano il numero medio di pagamenti. Il grafico dell'assorbimento di risorse nella giornata si presenta così (Figura 3):

[14] Cfr. paragrafo 3.2.3.

Figura 3 – Attività normale



Il processo termina nel minuto 744 (h 12:24), le risorse complessivamente assorbite sono pari a 3,17 *Full Time Equivalent* (FTE) e il numero massimo di risorse assorbite è pari a 9. Il tempo richiesto per un pagamento (*payment lead time*) è pari a 134 minuti.

Nel caso considerato (carico standard, cioè numero di pagamenti da elaborare pari alla media) le *performance* sono in linea con le attese poiché il completamento delle attività è molto antecedente rispetto al tempo limite (h 17,30). Ciò significa che le risorse, completato il processo, possono essere dedicate ad altre attività.

B. AS-IS assenza di personale

La situazione B (*AS-IS* / assenza di personale) riproduce un giorno lavorativo nel quale le due Unità elaborano il numero medio di pagamenti in assenza di una risorsa critica (coordinatore) per turno.

Le attività terminano al minuto 860 (h 14:20), il numero di FTE è ancora pari a 3,17 e il numero massimo di risorse utilizzate è 10. Il tempo medio richiesto per un pagamento è pari a 210 minuti.

C. AS-IS situazione critica

La situazione C (*AS-IS* situazione critica) riproduce ciò che avviene in una giornata lavorativa nella quale il volume dei pagamenti da trattare è triplo rispetto alla media.

Figura 4 – Assenza di personale

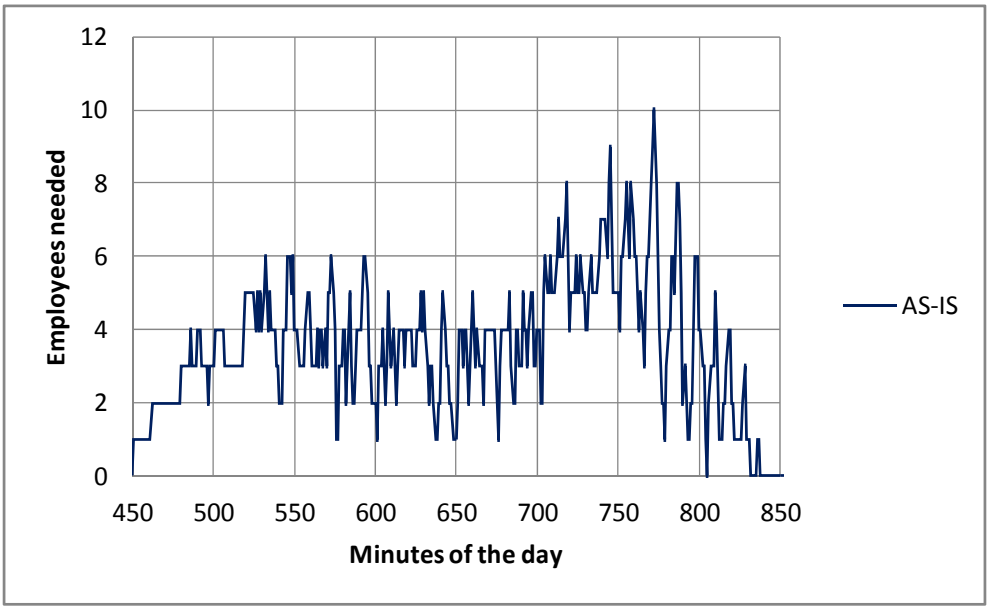
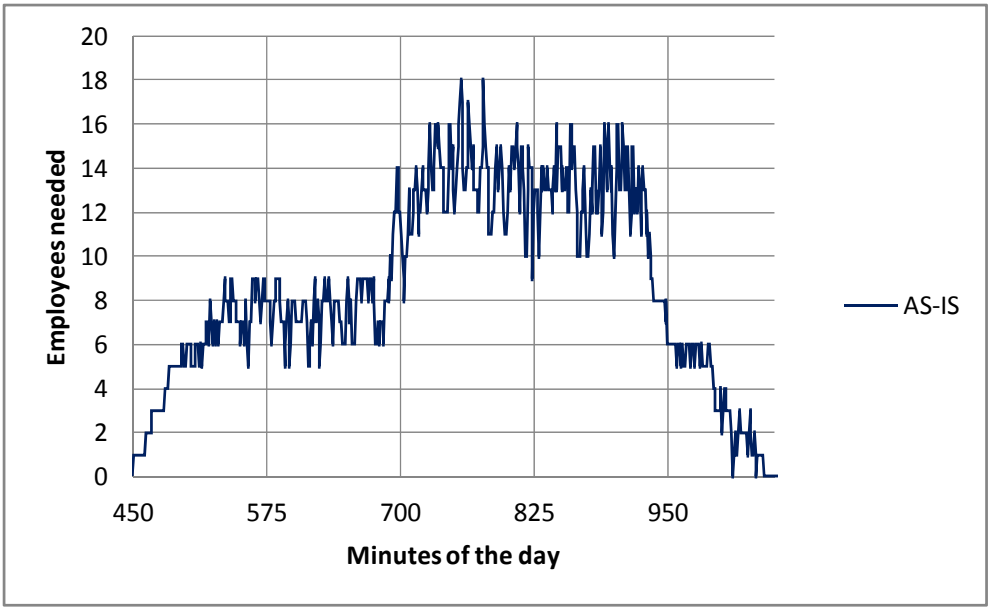


Figura 5 – Situazione critica (triplicamento del volume dei pagamenti)



Le attività terminano al minuto 1.061 (h 17:41), il numero di FTE è pari a 11,67 e il numero massimo di risorse utilizzate è 16. Il tempo medio richiesto per un pagamento è pari a 182 minuti. In questo caso l'attività presenta il forte rischio di superare la *deadline* fissata per il processo (la simulazione prevede una chiusura 11 minuti dopo il limite).

In questa situazione i tempi di lavorazione sono compressi al massimo, inducendo forti rischi di errore e generando una condizione di forte stress sulle risorse.

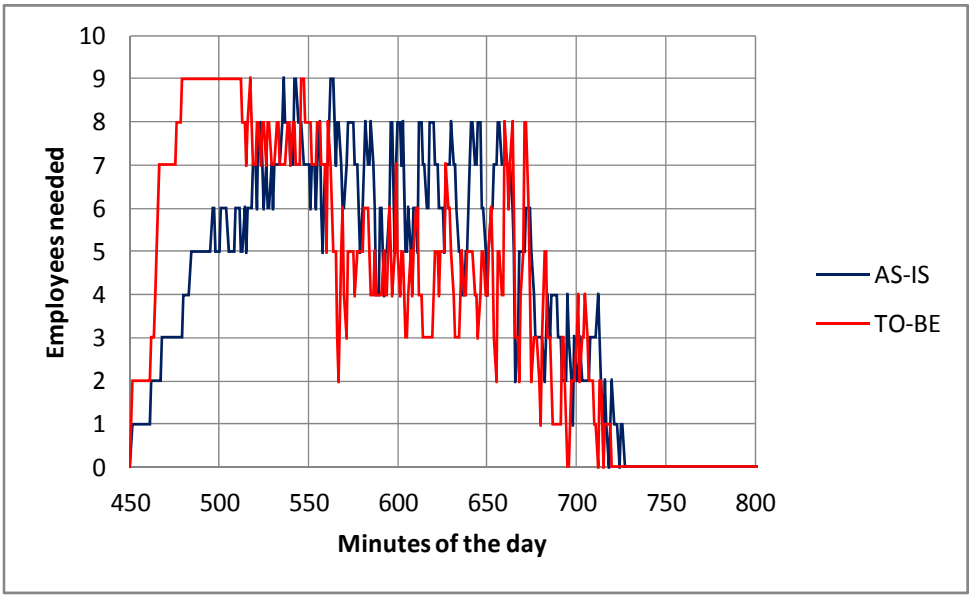
D. situazione TO-BE

La situazione D (*TO-BE*) simula ciò che accadrebbe nel caso in cui le due Unità fossero unificate. In questo caso tutte le risorse (indipendentemente dalla loro assegnazione originaria) possono essere usate nelle diverse fasi secondo necessità.

Nella successiva Figura 6 è rappresentata la simulazione di una condizione operativa normale: sono messe a confronto la nuova situazione (una sola Unità - curva blu) e la precedente (due Unità - curva rossa).

Adottando il nuovo assetto organizzativo (una sola Unità) le attività terminano al minuto 739 (h 12:19), il numero di FTE è ancora pari a 3,17 e il numero massimo di risorse utilizzate è 9. Il tempo medio richiesto per un pagamento è pari a 159 minuti. Non ci sono rilevanti differenze rispetto alla situazione *AS-IS*: il tempo di completamento del processo è lievemente inferiore (5 minuti in meno) mentre il tempo di elaborazione di un singolo pagamento è più elevato (25 minuti in più). Ciò è dovuto allo spostamento delle risorse sulle prime fasi del processo e alla maggiore parallelizzazione delle operazioni.

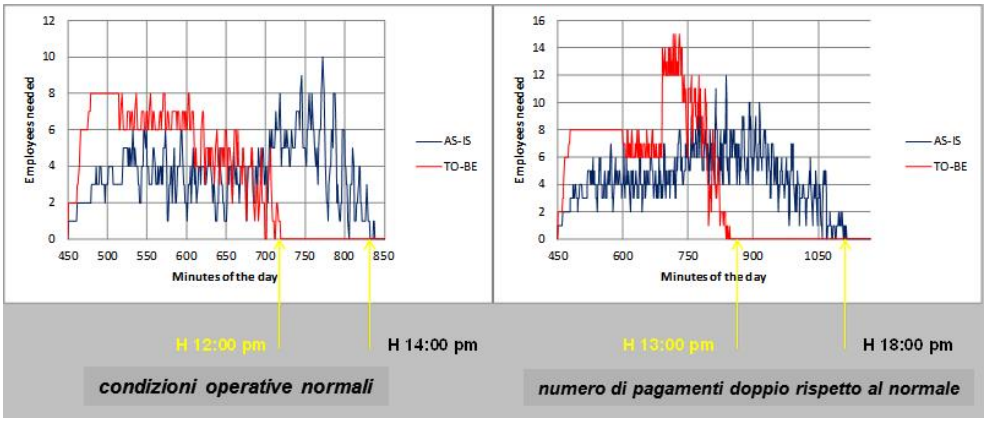
Figura 6 – Accorpamento delle Unità (situazione operativa normale)



Nelle situazioni critiche emergono invece differenze rilevanti fra le due situazioni (cfr. Figura 6). Nel caso di mancanza di personale (assenza di un Coordinatore per turno) la

fusione delle due Unità permette di chiudere le lavorazioni con due ore di anticipo rispetto alla situazione *AS-IS*; nel caso in cui il volume di transazioni è duplicato, la fusione delle due Unità permette di chiudere il processo con cinque ore di anticipo rispetto all'*AS-IS* e quindi ampiamente entro la *deadline* del processo.

Figura 7 – Mancanza di personale



4.3 Discussioni

In sintesi, la simulazione ha permesso di valutare l'impatto sulla *performance* (in particolare sul tempo di completamento del processo) di variazioni dell'*input* (numero dei pagamenti da effettuare), del numero e della composizione per ruolo delle risorse umane dedicate, nonché di possibili modifiche di carattere organizzativo. Ha reso anche possibile stimare il fabbisogno di risorse umane necessario a fronteggiare i possibili "picchi" di attività senza superare la soglia limite delle h 17:30.

In base all'analisi effettuata nel CS, è possibile affermare che la fusione fra le due Unità coinvolte nel processo, con la conseguente fungibilità mansionistica degli addetti, crea un assetto organizzativo più robusto che, grazie alla parallelizzazione dei flussi di lavoro, riesce ad affrontare con efficienza le situazioni critiche.

La fungibilità mansionistica deriva dall'utilizzo delle risorse su un ventaglio più ampio di attività (le caso delle strutture sperate ciascuna unità concentrava la creazione delle competenze degli addetti sui segmenti di processo di propria stretta competenza); ciò permette, inoltre, di dimensionare l'organico delle strutture su consistenze inferiori (di circa il 10%) rispetto alle attuali, perché rende possibile gestire i picchi strutturali di lavorazione attraverso la riallocazione temporanea di risorse.

Ciò porterebbe - a parità di *input* e di risorse umane assorbite - al sensibile accorciamento dei tempi di lavorazione nelle situazioni critiche e permetterebbe di fronteggiare efficacemente improvvisi picchi di operatività, limitando il rischio di rinviare una quota di

pagamenti al giorno successivo incorrendo nelle penali[15].

5. Conclusioni

Nel presente lavoro è stato proposto un approccio integrato di analisi e progettazione organizzativa, finalizzato ad aumentare l'efficienza dei processi interni degli Enti.

Tale approccio può trovare applicazione in qualsiasi realtà aziendale e permette di individuare efficaci azioni di miglioramento dell'efficienza dei processi di supporto.

Il *case-study* presentato nel paragrafo 4 dimostra come, attraverso l'analisi della morfologia dei processi e con l'ausilio degli strumenti di simulazione, è possibile determinare con precisione i più opportuni interventi organizzativi da adottare.

Le modifiche organizzative rappresentano solo una delle possibili leve d'intervento; a esse si affiancano la semplificazione dei processi, lo snellimento delle normative, l'esatta calibrazione dei controlli e l'introduzione di tecnologie informatiche per automatizzare le fasi operative e dematerializzare i flussi documentali.

E' opinione degli autori che tali modelli e approcci, nati in un contesto industriale maturo, possono giocare un ruolo determinante nel cammino intrapreso dalla Pubblica Amministrazione italiana nella direzione dell'efficienza, della qualità, della misurazione delle *performance* e dell'eliminazione degli sprechi. In un quadro di tagli crescenti ai *budget* pubblici, l'approccio qui illustrato (basato su un metodo di lavoro *olistico*, *misurabile* e *flessibile*) è particolarmente rilevante in quanto consente il recupero di efficienza. Vi sono, tuttavia, evidenti implicazioni in termini di *policy* in quanto azioni non ben meditate possono incidere anche sui servizi istituzionali che gli *Enti* erogano al pubblico. Diventa quindi necessario dotarsi di *framework* metodologici e concettuali, come quelli proposti nel lavoro, che consentano di migliorare l'efficienza, senza incidere sulla qualità dei servizi forniti. Tuttavia, l'estensione al settore pubblico fa emergere la necessità di ulteriori riflessioni dal punto di vista concettuale e metodologico in quanto gli enti pubblici presenta una serie di complessità di cui è necessario tenere conto nei modelli adottati, come la difficoltà oggettiva a definire il valore di mercato dei servizi prodotti, sia quelli esterni che quelli di supporto, la frequente mancanza di una rigorosa contabilità analitica ed il fatto che i modelli organizzativi adottati sono spesso sviluppati per rispondere a normative e regolamentazioni e non sono di tipo *output oriented*.

Riferimenti bibliografici

Bailey, D., *Automotive News calls Toyota world No 1 car maker*, Reuters.com., 24 January 2008, *http://www.reuters.com/article/businessNews/idUSN2424076820080124*.

D'Amato, *Analisi dinamica dei sistemi e dei modelli di simulazione per le strategie aziendali*, Franco Angeli, 1988.

[15] Per approfondire le caratteristiche e le potenzialità del simulatore descritto nel *case study*: *http://www.fairdynamics.com/*, *http://www.xjtek.com/*

Flake, G.W., *The Computational Beauty of Nature: Computer Explorations of Fractals, Chaos, Complex Systems, and Adaptation*, The MIT Press, Cambridge MA, 1998.

Geppert, L., Fiocca, R., *Reverse marketing and business paradigms*, *SIM Italian Marketing Society Congress*, Firenze, 2009.

Giachetti, R.E., *Design of Enterprise Systems, Theory, Architecture, and Methods*, CRC Press, Boca Raton, FL, 2010.

Holweg, M., *The genealogy of lean production, Journal of Operations Management* 25, 2007.

Montgomery, D.C., *Statistical Quality Control: A Modern Introduction*, Hoboken, New Jersey: John Wiley & Sons, 2009.

North, M.J., Macal, C.M., *Managing Business Complexity*, Oxford University Press, 2007.

Ohno, T., *Toyota Production System*, Productivity Press, 1988.

Radnor, Z., Walley P., Stephens A. Bucci G., *Evaluation Of The Lean Approach To Business Management And Its Use In The Public Sector*, scotland.gov.uk., 2008, *http://www.scotland.gov.uk/Publications/2006/06/13162106/0* .

Radnor & Bucci, *Analysis of Lean Implementation in UK Business Schools and Universities. Association of Business Schools*, 2010, *http://www.wbs.ac.uk/downloads/news/2011/03/abs-lean-report-exec-summary-march-2011-13008.pdf*.

Russo, D., Passacantando, F., Geppert, L., Manca, L., *Business Process Modeling and Efficiency Improvement Trough an Agent-based Approach*, *"The 8th International Symposium on Management, Engineering and Informatics: MEI 2012", in the context of "16th Multi-Conference on Systemics, Cybernetics and Informatics* – WMSCI 2012", Orlando FL, 2012. *http://www.iiis2012.org/wmsci/Website/AboutConfer.asp?vc=12*

Russo, D., "*Processi di IT Governance, qualità dei servizi IT e struttura organizzativa della funzione informatica: un modello integrato di indagine*", Università La Sapienza di Roma, Dipartimento di Informatica, *Master su "Audit e Governance nell'ICT"*, Roma, luglio 2011. *http://w3.uniroma1.it/mastersicurezza/index.php/master-itgov/*

Russo, D., *Workshop*: *L'evoluzione del back office nelle banche: lo sviluppo di efficienza ed efficacia operativa al servizio del cliente*, *FB Finance & Banking*, aprile 2012, *http://www.asseffebi.eu/*

Russo, D., *La gestione efficiente dei processi di supporto*", Università Cattolica di Milano /CeTIF, Milano, giugno 2012, *http://www.cetif.it/CM/main.aspx*

Russo, D., *Un approccio organizzativo integrato all'analisi e alla misurazione dei processi di supporto*, Università La Sapienza di Roma, Dipartimento di Informatica, *Master su "Audit e Governance nell'ICT"*, Roma, luglio 2012, *http://w3.uniroma1.it/mastersicurezza/index.php/master-itgov/*

Russo, D., *Strumenti metodologici per l'analisi delle organizzazioni IT*, *FB Finance & Banking*, dicembre 2009, *http://www.asseffebi.eu/*

Sterman, J.D., *Business Dynamics: Systems Thinking and Modeling for a Complex World*, Irwin-McGraw Hill, Boston, 2000.

Tennant, G., *SIX SIGMA: SPC and TQM in Manufacturing and Services*. Gower Publishing, Ltd., 2001, ISBN 0-566-08374-4.

The Open Group, *TOGAF standard*, *http://www.opengroup.org/togaf/*

What do Italian consumers know about Economic Data? Evidence from the Istat Consumer Survey[1]

Enrico Giovannini[2], Marco Malgarini[3], Raffaella Sonego[4]

Abstract

Standard theory describes economic decisions as result of optimising behaviour of well-informed agents. However, according to the "rational inattention" hypothesis, individuals may deliberately choose not to update their information set. The aim of our paper is study whether Italian consumers are adequately informed about economic data and to test if information is homogenously spread across the population. For this scope, we build a measure of knowledge of economic data at the individual level, and estimate a model relating knowledge to individual characteristics. Our main finding is that knowledge is relatively low and depends on the perceived costs and benefits of acquiring information. Results confirm one of the main postulates of the rational inattention hypothesis, i.e. that knowledge is highly differentiated across different population groups.

Keywords: Rational inattention, Household information acquisition, information and knowledge, consumer confidence, statistical literacy, media exposure.

1. Introduction

Mainstream economic theory describes policy decisions as the result of optimising behaviour of rational agents; on similar grounds, according to the public choice school, voters are also supposed to be well informed agents who base their decisions on utility maximisation. More generally, mainstream macroeconomics assumes that economic agents rationally elaborate on their full information set in order to form their savings or consumption decisions[5].

However, whether citizens are really well informed and rationally behaved is still highly disputed. Indeed, a number of studies have recently shown that agents are far from being fully informed about key economic variables; among them, Blinder and Kruger (2004) stressed the importance of determining *how* a society knows about statistics. They found that ideology is the strongest determinant in shaping public opinion: given the apparent inclination to use ideology, combined with the difficulty in building knowledge oneself, they find that US citizens tend to follow "ready-made" beliefs that society has chosen for

---

[1] The views expressed in this paper are solely those of the authors and do not involve the responsibility of their Insitutions.

[2] Università degli Studi di Roma "Tor Vergata", e-mail: enr.giovannini@gmail.com

[3] ANVUR, Roma, e-mail: malgmarco@gmail.com

[4] ISTAT, e-mail: rsonego@istat.it

[5] See on this Blinder and Krueger (2004).

them. According to Curtin (2008) people may be interested in knowing about how inflation affects their shopping trolley, or the unemployment rate in their specific labour market, but are less interested in learning about the performance of the whole country or in aggregated macro indicators which are difficult to apply to their daily life. In such circumstances, private information derived from neighbourhoods or local communities may be better appreciated by some than public information stemming from official sources. Reis (2006, 2009) interprets this kind of finding arguing that costs associated with the acquisition and use of information may generate "rational inattention", with widespread "knowledge inequalities" among the population.

In this respect, official statistics have an increasingly important role in the development of a common knowledge about the state and the evolution of a society: according to Giovannini et al. (2008) the value added of statistics critically depends on what people know about the world they live in[6]. Following this strand of literature, since 2007 the Italian Consumer survey[7] has incorporated once a year a number of questions on the degree of knowledge about economic data[8]. Questions concern knowledge about recent trends in GDP growth, inflation and the unemployment rate; consumers also have to report their opinions on the reliability of economic information and to indicate the main channels they use to acquire them. Finally, since 2009 they also have to report whether they use this kind of information in their decision process.

The aim of this paper is to analyse survey results in order to reach a better understanding of the level of knowledge of economic data and on if and how this knowledge is spread across the population. More specifically, we contribute to the existing literature testing on a brand new dataset the rational inattention hypothesis, checking if knowledge on economic data is homogenously spread across the population or rather if it is more concentrated in some socio-economic groups, depending on individual characteristics of the respondents. After having briefly introduced the consumer survey in section 2, section 3 presents a first description of the results obtained at the aggregate level. Hence, in section 4 we develop a new indicator of individual knowledge, the knowledge score, aimed at measuring the overall level of economic knowledge of each consumer. In section 5 we then estimate a probit model in order to test whether the probability of replying to survey questions and the level of the knowledge score are influenced by socio-demographic factors such as the age, gender, area of residence, professional status and education of the respondent; moreover, we also consider the possible role of opinions on the importance of this kind of information and of desire to be informed about economic issues. Concluding remarks are presented in Section 6.

[6] See also Giovannini (2013).

[7] The survey is part of the Harmonised Project of Business and Consumers survey coordinated by the European Commission; for details, see http://ec.europa.eu/economy_finance/db_indicators/surveys/index_en.htm.

[8] See Fullone et alii (2008) and D'Urzo et alii (2009).

2. The ISTAT Consumer survey

2.1 The Sample and the questionnaire

The ISTAT Consumer survey consists of qualitative questions on the personal situation of the consumer and the country. It is conducted monthly with a Computer Assisted Telephone Interviewing (CATI) system, on the basis of a random sample of 2.000 Italian consumers, changing each month, without any panel dimension[9]. The sample is selected in two stages (subscriber to the telephone register, in the first stage; individual consumer within the household of the subscriber, in the second stage), being proportional to the population of reference, represented by the Italian adult (aged 18 or more) population (about 50 million statistical units). The survey is stratified by geographical partitions and demographic width of municipalities, for a total of 42 strata (see Table 1). The sampling list is represented by the public fixed telephone directory (containing about 18 million units)[10]; the sampling unit in the first stage is the subscriber to the fixed telephone directory, randomly selected within the stratum; the statistical unit is the individual consumer, intended as the adult person chosen within the household of the subscriber. Random selection is used in the first stage; in the second stage, quota selection according to gender applies (48,5% males, 51,5% females)[11]. The response rate of the survey ranges from 45 to 66%, depending whether we consider among the eligible cases all the potentially eligible contacts (i.e. including cases in which the telephone is busy or there is no answer), or only the effectively eligible cases (i.e. excluding the two cases reported above and hence including among the non responses only refusals, unreachable contacts and automatic repliers; see on this table 8, page 20 in Fullone, Martelli, 2008). A response rate falling in a range of about 60 to 65% is usually considered as acceptable in the literature concerning this kind of surveys[12]; appropriate CATI techniques (i.e., high number of contact attempts, personal call-backs) are currently used in order to minimize distortions. However, in the analysis of survey results reported in the paper the reader should be aware of possible bias arising from the non-negligible share of non responses. In order to take into account possible selection biases and changes over time in the households composition and age structure, in this paper we will use a system of probability and post-stratification weights, based on Fullone and Martelli (2008). According to official ISTAT figures available on the EU website, the sampling error of the estimates is equal on average to 0,7 percentage points: i.e., all the estimates reported below should be considered as comprised between a confidence interval equal to ±0,7% with respect to the central estimate.

[9] As it is common in the EU-Harmonised Consumers' Opinion surveys; in the US experience, on the other hand, a fixed proportion of the sample is re-interviewed after six months (see on this Curtin, 2015).

[10] The use of fixed telephone directories as the sampling frame can generate an increasing bias, since their coverage of the reference population is diminishing over time; however, as already recognized by UN (2014), fixed telephone registers still represent the most used framing lists in this field. In fact, possible alternatives resorting to mobile phones or internet connections also rise relevant problems in terms of coverage and selection bias (see on this Curtin, 2003; 2015).

[11] See also ISTAT and EU metadata, respectively available at: http://siqual.istat.it/SIQual/visualizza.do?id=8888944&refresh=true&language=IT; http://ec.europa.eu/economy_finance/db_indicators/surveys/metadata/index_en.htm.

[12] See on this McKenzie, 2005.

Table 1 – The sample (number of units and percentage shares)

	SIZE OF MUNICIPALITIES (number of inhabitants)							
GEOGRAPHICAL AREAS	up to 5,000	5,001 - 10,000	10,001 - 20,000	20,001 - 50,000	50,001 - 100,000	100,001 - 500,00	500,001 +	Total
North – West (number of units)	56	22	23	34	19	4	51	209
(percentage share)	2.8%	1.1%	1.2%	1.7%	1.0%	0.2%	2.6%	10.5%
North – Centre (number of units)	70	63	55	55	26	14	45	328
(percentage share)	3.5%	3.2%	2.8%	2.8%	1.3%	0.7%	2.3%	16.4%
North – East (number of units)	64	67	81	56	18	100	0	386
(percentage share)	3.2%	3.4%	4.1%	2.8%	0.9%	5.0%	0.0%	19.3%
Centre (number of units)	42	39	52	79	50	45	93	400
(percentage share)	2.1%	2.0%	2.6%	4.0%	2.5%	2.3%	4.7%	20.0%
South (number of units)	75	58	77	94	81	43	31	459
(percentage share)	3.8%	2.9%	3.9%	4.7%	4.1%	2.2%	1.6%	23.0%
Islands (number of units)	34	26	30	48	28	31	21	218
(percentage share)	1.7%	1.3%	1.5%	2.4%	1.4%	1.6%	1.1%	10.9%
Total (number of units)	341	275	318	366	222	237	241	2000
(percentage share)	17.1%	13.8%	15.9%	18.3%	11.1%	11.9%	12.1%	100%

Source: ISTAT

The first part of the questionnaire provides structural information about the consumer and her household, including age, gender, the area of residence, level of education and working status of the respondent (see Table 2); the second part gathers consumers opinions on the general economic situation of the country (including questions on unemployment and price dynamics) and on that of the economic conditions of the household and of the individual consumer. The survey also asks Italian consumers about their income; more precisely, the respondent is asked to assign family income to one out of 22 classes, rather than providing a precise estimate.

Table 2 – Structural information about the individual and the household

Information about the individual	Modalities of reply
Gender	Male; Female
Region of residence	20 Italian administrative regions
Size of the municipality of residence	7 classes, see table 1
Relationship with the head of the household	Head of the household; Husband, wife; Son, daughter; Grand Parent; Other relative; Other
Age	18-20 years; 21-29; 30-39; 40-49; 50-59; 60-64; >64
Occupation	Full time; part time; unemployed; Pensioner; Student; Renter; Other (housemaid)
Professional category	Independent worker; agricultural worker; White collar employee; Specialised blue collar; non-specialised blue collar
Open ended / permanent worker	Open ended contract; permanent contract
Education (completed)	University degree; Tertiary education; Secondary education; Primary education; no cycle completed
Information about the household	**Modalities of reply**
Number of people in the household	Number
Total monthly family income, net of taxes, including capital income and transfers	22 brackets, from <350 euros to >6.000 euros

Source: ISTAT

2.2 Questions about knowledge of economic data

The first survey on the knowledge of Italian consumers about economic data has been administered in 2007 in close collaboration with OECD Statistics Directorate; the survey has become yearly since 2009[13]. The main goal is to verify the degree of knowledge of Italian consumers about the recent developments – as registered by official statistics – of key economic variables such as GDP growth, inflation and the unemployment rate. Every question contains three core elements: a brief definition of the key statistical variable, a reference to the agency responsible for its publication and a question about the most recently published figure. Participants may choose to: 1) answer, 2) indicate that they do not know the exact figure, or 3) refuse to answer. Failure to report official data could imply that participants are not aware of the most recent figure or that they do not know it, possibly because they have not recently heard about it in the media. In this respect, a scarce knowledge of the most recent data associated with a general knowledge of the phenomenon may imply a process of "staggering updates", in which people infrequently update their knowledge because of high costs and relatively low return. On the other hand, if the consumer has not recently heard about official data releases, he/she may well be considered

[13] The three questions read as follows:

Unemployment rate: As you may know, every quarter the Italian National Institute of Statistics publishes figures on the unemployment rate in Italy. In other words, every three months ISTAT officially reports the percentage of people unemployed with respect to the active population. Can you please tell us the most recent rate of unemployment published by ISTAT?

Inflation rate: Another important economic indicator that is published by ISTAT on a monthly basis is the consumer price index, commonly used to calculate the annual inflation rate. Can you please tell us the most recent rate of inflation published by ISTAT?

GDP growth: ISTAT has recently published figures on all final goods and services produced in Italy in 2008. This figure is known as the Gross Domestic Product (GDP) of the country. Can you please tell us the percentage of change of the Italian GDP recently published by ISTAT?

to be unaware of the existence of such data and of its use. Following Curtin (2008), in order to try to distinguish among these two cases, a follow-up question was introduced a first time in the 2009 and then regularly since 2012 for each of the previous questions, asking if the consumer has recently heard of a public announcement concerning official statistics on GDP, inflation and the unemployment rate.

The questionnaire also collects answers about the importance of being informed on such issues, asking about the desire to be more informed and the media channels used to acquire information (possible media considered in the question are the television; radio; internet; newspapers and magazines; scientific publications; contacts with friends and relatives, with experts and politicians). Two further questions ask for an assessment on the quality of economic information provided by the media and the quality and reliability of official statistics. Indeed, a previous study based on the Eurobarometer survey (Papacostas, 2008) has shown that there is a significant relationship between trust in official statistics and trust in the transparency of political decisions, confirming the important role of sound and accountable statistics in modern democracies. A final question asks the consumer if she uses information about GDP growth, inflation and the unemployment rate in her everyday life, in order to make strategic decisions about consumption and saving; in fact, as shown in Blinder and Kruger (2004) and Curtin (2009), people that make use of economic data in everyday life are expected to update more frequently their information set and hence to be better informed about those issues.

3. Aggregate results[14]

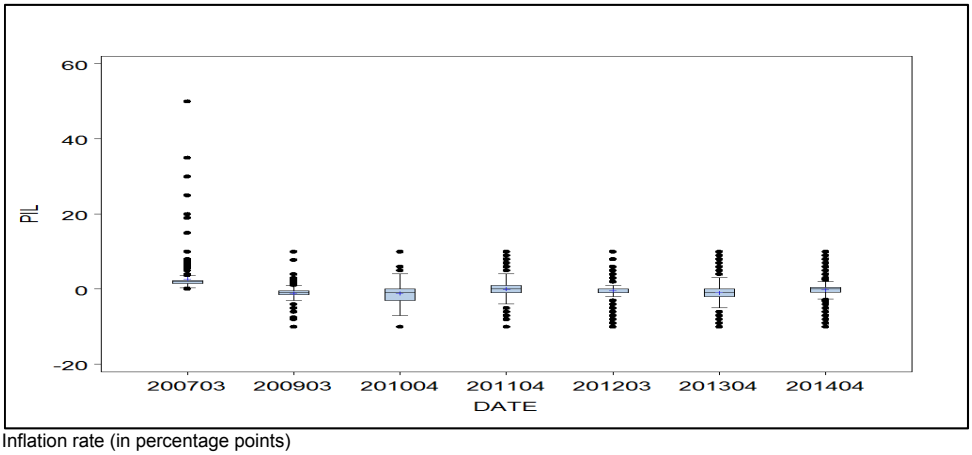
According to the ”rational inattention” approach, citizens follow more closely available information when it is perceived to be particularly relevant; in this sense, it is possible to assume (Curtin, 2008) that the economic crisis started in 2008 may have generated an increased sensibility to economic data. This hypothesis seems to be broadly confirmed by aggregate survey results. Figure 1 presents the Box-plot distribution of quantitative replies about the subjective knowledge on the statistics of interest. The box represents the answers' distribution around the median value (continuous line within the box), distinguishing among the answers comprised between the 75^{th} and 25^{th} percentile (respectively, the upper and lower margin of the box), answers immediately below and above the threshold (answers comprised within the segments above and below the threshold) and outliers, represented as dots.

Number of outliers for consumers' knowledge about GDP and the unemployment rate decrease over the years; moreover, in the case of the answers about GDP growth, in the last two years the 25^{th} and 75^{th} percentiles are much closer to each other, a result that may be interpreted as a decrease in the level of uncertainty about this variable. On the other hand, public knowledge about inflation does not seem to have changed much: outliers remain much more frequent than for the other two variables and the inter-quartile difference remains broadly stable. Indeed, it may be considered that during the economic crisis attention of the media and the general public was rather focussed on growth and unemployment than on inflation, thus these results may be interpreted as preliminary evidence of a “rational inattentive” behaviour of Italian consumers over the last few years.

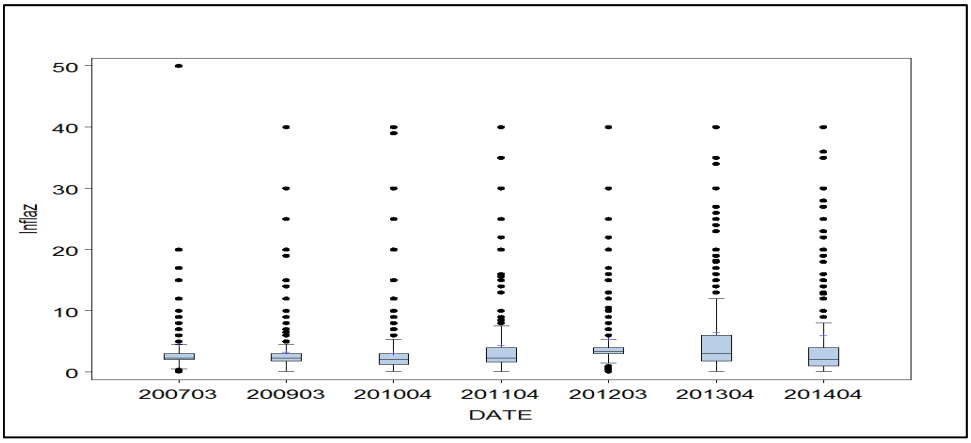
[14] This section is based on the data published each year by ISTAT (see http://www.istat.it/it/archivio/164177).

Figure 1 – Distribution of the answers

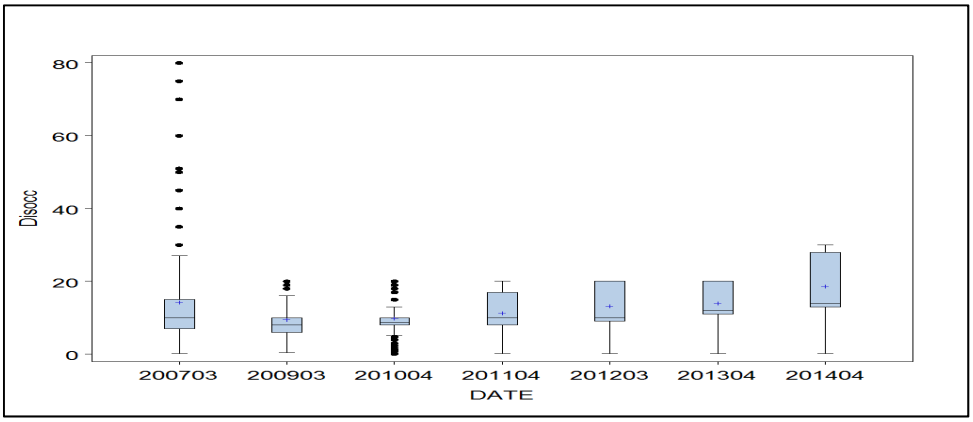
GDP growth (in percentage points)



Inflation rate (in percentage points)



Unemployment rate (in percentage points)



Source: Authors' elaboration on ISTAT data

Response rates vary between a minimum of 17% for the question about inflation in 2010 to a peak of 59% for the question about the unemployment rate in 2014 (Table 3). Among non-respondents, those appearing to be inattentive rather than completely unaware of economic data do prevail: the quote of those reporting to have heard about the data without being able to report the latest figure varies between 23% (question about GDP in 2014) and 51% (question about inflation in 2014), while the share of those not having heard at all about the data recently (i.e. those that we deem not having any knowledge of the statistic at hand) varies between a minimum of 7% for the unemployment rate in 2013 and 2014 and a peak of 28% for data about GDP growth in 2009.

As shown in Figure 1, the distribution of the replies is characterized by the presence of relevant outliers; more precisely, outliers are defined as the values lying above and below, respectively, the upper and lower whiskers of the box plots derived from the data[15]. On the basis of this evidence, median value may be considered as a more accurate measure of the distribution than the mean. Considering median values, Italian consumers are quite accurate regarding GDP developments for the years 2007, 2009 and 2012; on the other hand, they strongly underestimate the severity of recessions for the years 2010, 2013 and 2014, providing instead figures worse than the true ones for 2011. Median values for replies concerning the unemployment and the inflation rate are always well above actual values.

[15] Denoting with Q1 and Q3, respectively, the first and third quartile of the distribution, the upper whisker of the box plot is equal to Q3+1.5*(Q3-Q1), while the lower whisker is defined as Q1-1.5*(Q3-Q1).

Table 3 – Knowledge about GDP growth, the inflation rate and the unemployment rate

GDP growth

Percentage shares	2007	2009	2010	2011	2012	2013	2014
Consumers reporting a figure	26%	23%	20%	34%	34%	37%	56%
Consumers non reporting a figure	72%	73%	79%	64%	65%	62%	43%
Of which:							
- I have heard about it, but I do not remember the exact figure		44%			42%	45%	23%
- I have not heard about it recently		28%			22%	16%	19%
- Don't know		2%			1%	1%	1%
Refuse to answer	3%	4%	1%	3%	2%	1%	1%
(In percentage points)							
Average	2.7	-1.4	-1.0	0.1	-0.4	-1.0	-0.3
Median	2.0	-1.0	-1.0	0.0	0.0	-1.0	0.0
First quartile	1.5	-1.8	-3.0	-1.0	-1.0	-2.0	-0.8
Third quartile	2.4	-0.5	1.0	1.0	0.0	0.0	0.6
Standard deviation	3.7	2.2	3.2	3.0	3.1	3.2	3.7
Official data (a)	1.9	-1.0	-5	1,3	0,4	-2.4	-1.9

Inflation rate

Percentage shares	2007	2009	2010	2011	2012	2013	2014
Consumers reporting a figure	32%	24%	17%	26%	29%	33%	26%
Consumers non reporting a figure	66%	74%	73%	62%	64%	58%	72%
Of which:							
- I have heard about it, but I do not remember the exact figure		49%			43%	47%	51%
- I have not heard about it recently		23%			20%	11%	20%
- Don't know		2%			1%	1%	1%
Refuse to answer	2%	3%	9%	12%	7%	9%	2%
(In percentage points)							
Average	4.7	3.2	3.5	4.7	5.6	7.3	7.7
Median	2.4	2.5	2.0	2.4	3.3	3	2.7
First quartile	2.0	1.8	1.2	1.8	3.0	2	1.2
Third quartile	3.0	3.0	3.0	4.0	4.5	10	10
Standard deviation	8.9	3.5	6.3	6.4	7.1	9	10.5
Official data (b)	1.8	1.6	1.4	2.4	3.3	1.7	0.4

Table 3 segue – Knowledge about GDP growth, the inflation rate and the unemployment rate

Unemployment rate							
Percentage shares	2007	2009	2010	2011	2012	2013	2014
Unemployment rate	31%	22%	27%	39%	44%	46%	59%
Consumers reporting a figure	66%	75%	66%	55%	53%	52%	40%
Consumers non reporting a figure							
Of which:							
- I have heard about it, but I do not remember the exact figure		50%			42%	44%	33%
- I have not heard about it recently		24%			11%	7%	7%
- Don't know		1%			1%	1%	0%
Refuse to answer	3%	3%	7%	6%	3%	3%	1%
(In percentage points)							
Average	14.6	10.0	10.2	11.8	13.4	14.1	19.9
Median	10.0	8.0	9.0	10.0	12.0	13.0	20.0
First quartile	7.0	6.0	8.0	8.0	9.0	11.0	13.0
Third quartile	18.0	12.0	11.0	20.0	20.0	20.0	30.0
Standard deviation	13.0	5.7	4.6	6.2	5.7	5.2	7.8
Official data (c)	6.8	6.7	8.2	8.6	9.3	11,6	13

Source: authors' elaboration on ISTAT data.

(a) Official data is the most recent release of the yearly GDP growth rates available at the time of the survey, referred to the year before with respect to the one in which the survey was realized.

(b) Official data is the most recent release of the monthly inflation rate available at the time of the survey, referred to the month before with respect to the one in which the survey was realized.

(c) For the period 2007-2009, the official data is the most recent quarterly release of the unemployment rate available at the time of the survey, referred to the third or fourth quarter of the year before the one in which the survey was realized; for the period 2010-2014, official data is the most recent monthly release of the unemployment rate, referred to the month of January of the year in which the survey was realized.

Tables 4 reports the opinions of respondents about the quality of the public debate about these data. The increase in knowledge that has emerged from the analysis of the replies goes together with better assessments about the quality of the public debate and of official statistics in general. However, the share of respondents deeming that the quality and reliability of information published by the media is "good" is still largely lower than that of those considering it as "bad".

Table 4 – Quality of economic information

In your opinion, during the recent economic and financial crisis, the quality and reliability of the information on the economic situation published by the media and the public debate on these issues has been: Good/Sufficient/Bad?					
Percentage share of respondents answering:	2010	2011	2012	2013	2014
Good	8.5	7.2	14.8	17.3	19.3
Sufficient	32.5	36.1	39.4	31.5	30.4
Bad	47.3	43.9	38.4	43.5	43.5
Don't know	11.0	10.2	6.5	6.9	6.7
Refuse to answer	0.7	2.6	0.8	0.8	0.1

Source: authors' elaboration on ISTAT data.

The survey also shows (see Table A in the Statistical Appendix) an increase over the years of the importance assigned by the respondents to economic data: the share of those deeming they are very or fairly important rose from 71% to 83%, with a larger gap between the share of those deeming that the data are "very important" and the share of those judging them as "fairly important". Indeed, the increased importance of economic information also stimulated the desire to be more informed, expressed by almost half of the sample. Moreover, in the last two years the share of people using ("a lot" or "a bit") this kind of information as a support for relevant decisions concerning consumption and savings behaviour has significantly risen from 7% to over 17% (see Table B in the Statistical Appendix). On the other hand, those not using at all economic information has fallen from 77% of the sample in 2010 to 63% in 2014. Overall, the data seem to suggest that during the economic crisis the increased importance of economic data has gone hand in hand with a better assessment on the quality of the information, an increased desire to be informed and a growing use of the information for strategic decisions.

Finally, the survey also provides information about the media mostly used to gather this kind of data; the inclusion of this question was suggested by the Blinder and Kruger (2004) paper, where the source used to acquire information had a significant influence on the overall level of knowledge of US consumers about economic phenomena. According to our results, similarly to the findings obtained in the US, television is by far the most important channel for Italian consumers, while the internet is now considered, together with newspapers and magazines, the second most important channel of information, followed by the radio. More "private" channels of information as the contacts with friends and relatives are much less important; a minority of respondents also uses scientific press in order to acquire information about economic data.

4. The Knowledge Score

In order to assess the overall individual knowledge of economic data, we adopt the methodology already introduced in Fullone et al. (2008), D'Urzo et al. (2009) and Giovannini and Malgarini (2012). For each question we first calculate the absolute value of the relative error with respect to the official data available at the moment of the interview and then compute the Mean Absolute Relative Error (MARE) for the three questions, where

a higher score indicates a lower knowledge of economic data:

$$MARE_{i,t} = \frac{\sum_{J=1}^{3} \left| \frac{R_{i,j,t} - ISTAT_{j,t}}{ISTAT_{j,t}} \right|}{3} \qquad (4.1)$$

In (4.1) $R_{i,j,t}$ is the reply of an individual consumer i to each question j for each time t; $ISTAT_{i,t}$ is the official data pertaining to question j for time t. Hence, we calculate two different raw scores for each survey: in order to fully exploit their information content, the first score is calculated by considering also the *"don't know"* answers and excluding only those refusing to respond. To those answering *"don't know"*, we impute a score equal to the maximum value reached by the score of those having answered the question in each wave, augmented by a unit. In other words, we "penalise" those answering of not knowing about the statistic under discussion assigning them the maximum error committed by those having indeed provided a reply, augmented by a unit; the respondent should have answered at least one question to be considered in the score[16]. In this case we have a total of 4.923 observations available for the analysis.

The second score is calculated using also the information provided in the follow-up question asking whether consumers have publicly heard of such official statistics; in this case, data are available for the year 2009 and then yearly since 2012. We interpret this evidence as a measure of "rational inattention", i.e. we consider that those not being able to answer but having heard about the indicator of interest are subject to staggered updates, either because of the high cost or because of the low benefit of acquiring information. For this reason, we assign them a score equal to the maximum score available augmented by one; furthermore, we augment the maximum score by two units to those reporting that they have not heard recently about the data. Those refusing to answer are still excluded from the calculation, reducing the total availability to 4.492 observations. In the following, we shall use a linear transformation of individual MAREs, standardising them with respect to the mean and standard deviation of their distribution and calculating two z-scores, having the advantage of holding useful linear mathematical properties:

$$Z - score_{i,t} = \frac{MARE_{i,t} - Mean(Mare)}{Standard\ Deviation\ (Mare)} \qquad (4.2)$$

[16] We experimented with different possible values to be assigned to the "don't know" answers, e.g. assigning the maximum error committed from those having provided a reply; overall results are not influenced by the arbitrary choice concerning the quantification of the "don't know" replies.

5. Statistical knowledge, socio-demographic factors, desire to be informed and the media

As already pointed out in section 1, according to Mankiw and Reis (2002) acquiring, absorbing and processing information is costly, and hence consumers may rationally choose to update their information set only sporadically; as a consequence, information propagates slowly through society, level of individual knowledge being greatly heterogeneous across demographic groups (Souleles, 2004). According to this view, the main sources of heterogeneity are the level of education (Reis, 2006) and individual economic conditions (Blinder and Kruger, 2004), followed by other socio-demographic factors including age, race and gender. In the ISTAT survey, this hypothesis may be tested using data concerning age (4 classes, from <30 years to 65+), gender and education attainments (3 possible outcomes, from lower intermediate to University level) of the respondent. The survey also comprises data regarding self-reported income levels (expressed in quartiles) and other possible proxies for the economic situation, including employment status (4 categories, employees, self-employed, unemployed and inactive people), zone of residence (North West, North East, Centre and South) and number of inhabitants of the city of the respondent (5 categories from small town with less than 5,000 inhabitants to big cities with more than 500,000 people). Blinder and Kruger (2004) also pointed out that the level of knowledge may be influenced by the desire to be informed on the issues at stake and by the channels used to acquire the information; in our case, those data are available at least for some of the waves. All those variables define a vector of possible correlates for the probability to reply and the level of the score, the vector being denoted as Z_{it}.

5.1. Probability of answering knowledge questions

We define an ordinal discrete categorical variable assuming values comprised between 0 and 3 on the basis of the number of replies to the survey questions, $x \in \{0;3\}$; we consider that the respondent is wishing to reply if she has indeed updated her information set, and hence we interpret this variable as a proxy of the frequency of updating. We assume that the probability of the number of replies being equal to x may be influenced by the vector of controls Z_{it}:

$$\Pr(q_{it} = x \mid Z_{it}) = F(\beta' Z_{it}) \tag{5.1}$$

In (5.1), F is the cumulative function of the normal distribution, and the model is estimated as an Ordered probit, an extension of the standard binary probit model used when the dependent variable takes the form of a ranked and multiple discrete variable. In $Z_{i,t}$ we also add a set of time dummies in order to test for possible differences in knowledge among the various waves. Finally, we consider as individual weights the probability of inclusion for each respondent, calculated according to the methodology described in Fullone and Martelli (2008); unobserved error terms are assumed to be heteroscedastic and hence the model is estimated with robust methods.

Estimation results for model (5.1) are presented in Table 5. The three columns respectively report the results for the whole sample, those obtained using also information

on quality and use of statistical information (available only since 2010) and those obtained taking into account also the importance and desire to be informed (available for the whole sample, but 2010). In the estimation, we normalise with respect to male respondents, being independent workers, in the first income quartile, under 30 years of age, living in the North West with the lowest education attainment, deeming (when these opinions are available) that economic information is not important, of bad quality, not used and with no desire to be more informed. Therefore, the statistical significance, sign and magnitude of estimated parameters have to be interpreted as differentials with respect to this control group.

Estimation results broadly confirm the main findings of the rational inattention literature; considering a confidence level of 5%, probability to reply to the knowledge questions is growing with level of education and economic conditions. In fact, those with an high school or University degree, lying in the top two quartiles of the income distribution, living in the areas with higher per capita income levels (the regions of the North) and being independent workers have an higher probability of answering the knowledge questions. We also found a statistically significant positive effect of the size of the city the respondent lives in, which may be interpreted as a further proxy for economic conditions, and possibly also as a proxy for accession to different information sources. Similarly to what has been previously found in other countries (see for instance Bryan and Venkatu, 2004) men are found to update their information set more frequently than women. The probability to reply is also found to be influenced by age, with those between 50 and 65 years being more able to reply to the survey questions.

Empirical results confirm another seminal intuition of the rational inattention literature, namely that the frequency of updating is higher when information is considered more important. In fact, the probability of giving an higher number of replies is higher for those deeming that those information are important, which use these information in their decision making process and which are willing to be more informed on these issues. Information channels also matter, with those using other media on top of TV being more willing to reply to the survey questions. Finally, willingness to reply also grows during the time span of the survey: the latter result is also consistent with the rational inattention hypothesis, since economic information is usually considered to be more important in difficult times like those of the period 2009-2014. It should also be considered that in the period under examination, and especially since 2010, the Italian National Institute of Statistics has dramatically renewed its communication strategies, with an increase in the amount of information made available to the public trough press releases, the publication of a new open online database (http://dati.istat.it/) and a renewed website (www.istat.it). These initiatives led to a remarkable increase in the media coverage of Italian statistics. On the other hand, no effect is found for the opinion on the quality of the public debate on the media about economic information.

Table 5 – Ordered Probit model on the probability of replying to the questions about GDP growth, the inflation rate and the unemployment rate

Independent variables (value of the coefficient and statistical significance) (a)	Estimation period		
	Whole sample	2010-2014	Whole sample, except 2010
Socio-demographic variables			
Professional category (control group: independent workers)			
Dependent workers	-0.271***	-0.262***	-0.291***
Unemployed	-0.253*	-0.255	-0.292**
Inactives	-0.297***	-0.332***	-0.302***
Age (control group: 18-29 years)			
30-49 years	0.089	0.084	0.163**
50-65 years	0.313***	0.332***	0.372***
>65 years	0.075	0.177*	0.163**
Gender (control group: men)			
Women	-0.530***	-0.537***	-0.531***
Area of residence (control group: North West)			
North East	-0.051	-0.049	-0.042
Center	-0.151***	-0.130**	-0.156***
South and Islands	-0.187***	-0.191***	-0.204***
Number of inhabitants (control group: <5000 inhabitnants)			
From 5.001 to 20000	0.120**	0.114*	0.129**
From 20001 to 100000	0.167***	0.182***	0.173***
From 100001 to 500000	0.190***	0.180**	0.202***
Over 500000	0.192***	0.180**	0.187***
Education attainment (control group: primary school)			
Secondary school	0.321***	0.245***	0.274***
University	0.562***	0.484***	0.503***
Income (control group: first quartile)			
II Quartile	0.096*	0.030	0.066
III Quartile	0.204***	0.158**	0.165***
IV Quartile	0.224***	0.169**	0.203***
Information channels			
Tv Only	-0.039	0.054	-0.035
Radio	0.112**	0.137**	0.069
Newspapers	0.345***	0.307***	0.311***
Internet	0.287***	0.266***	0.265***
Political leaders	0.408***	0.357***	0.340***
Friends and relatives	0.150***	0.172**	0.115*
Reliability and use of information			
Quality of information (control group: bad)			
Good		-0.016	
Sufficient		0.000	
Use of information (control group: no use)			
Use		0.422***	
Importance of information (control group: not important)			
Important			0.415***
Desire to be informed (control group: no desire)			
Desire			0.325***
Time control (control group: 2007)			
2009	-0.200***		0.159
2010	-0.302***		
2011	0.257***	0.477***	0.629***
2012	0.415***	0.608***	0.755***
2013	0.570***	0.788***	0.918***
2014	0.648***	0.901***	1.005***
Number of observations	9,594	6,342	8,268

*** p<0.001; ** p<0.05; * p<0.1

Source: authors' elaborations on ISTAT data

5.2. Estimation results: the knowledge score

In this section we turn to the analysis of the possible relationship among the level of knowledge as measured by the two alternative definitions of the z-score reported in section 4 and the same vector Z_{it} of possible correlates as in section 5.1. The estimated model is the following:

$$K_{i,t} = f(Z_{i,t}) + u_{i,t} \quad (5.2)$$

where K_{it} is the individual knowledge score in each time t. We estimated the model with OLS, considering probabilities of inclusion as individual weights and accounting for possible heteroscedasticity in the unobserved error term with robust methods. Table 6 reports the results obtained; similarly to table 3, the first three columns respectively report the results for the whole sample, those obtained using also information on quality and use of statistical information (available only since 2010) and those obtained taking into account also the importance and desire to be informed (available for the whole sample, but in 2010). The fourth column reports the results obtained using the alternative definition of the knowledge score described in session 4, in which we include also the replies to the follow up question, administered only in the 2009 and 2012-2014 waves. As already stated in section 4, in this case we explicitly consider the possibility that the respondent has indeed heard about the data, but has decided not to update her information according to the "rational" inattention hypothesis. Once estimating the model for the whole sample, and hence without considering the follow up questions available only in 2009 and 2012-2014, the number of available observations vary between 4.923 and 3.659, depending on the availability of the controls used in the analysis. We use the same normalisations adopted in table 5: therefore, the constant term may be interpreted as the average z-score for the control group, the coefficients of the various dummies representing – if significant – the increase/decrease in knowledge associated with the specific characteristic at hand.

Also in this case, results are broadly supportive of the theoretical framework we have adopted for the analysis, even if some differences do emerge with respect to the estimation of model (5.1). In particular, higher education attainments looks correlated not only with a higher frequency of updating, but also with a higher level of knowledge (i.e., with a lower z-score), On the other hand, the effect of economic conditions on the z-score is limited or absent, with only a mildly significant negative effect for those living in the South, characterised by lower average income levels. No effect is found for self-reported income levels and the size of the municipality the respondent lives in. A strong effect of age emerges, with the level of knowledge steadily growing as the respondents gets older. This is a relatively new finding in this kind of literature, since both Blinder and Kruger (2004) and Curtin (2008) found only a small effect of age, respectively, on the desire of being informed and on the probability of replying to a knowledge question similar to the ones used in this study. Moreover, those using newspapers and the internet have a higher knowledge score; likewise, those willing to be more informed about these issues, deeming that economic information is important and using this kind of information in their decision-making process have a better knowledge than the control group. No significant differences in the level of knowledge are found according to the opinions on the quality of the media debate. Overall, the regression explains over 50% of the total individual variability of knowledge levels and results seem to be quite robust across different specification of the control variables and over time. Similar results are found if we also consider the follow up question.

Table 6 – The level of knowledge on economic data and its possible determinants

Independent variables (value of the coefficient and statistical significance) (a)	Estimation period			
	Whole sample	2010-2014	Whole sample, except 2010	2009; 2012-2014
Socio-demographic variables				
Professional category (control group: independent workers)				
Dependent workers	0.053	0.029	0.052	0.050
Unemployed	-0.099	-0.171	-0.111	-0.107
Inactives	0.049	0.037	0.040	0.046
Age (control group: 18-29 years)				
30-49 years	-0.129**	-0.140*	-0.150**	-0.151**
50-65 years	-0.245***	-0.239***	-0.260***	-0.270***
>65 years	-0.286***	-0.330***	-0.312***	-0.332***
Gender (control group: men)				
Women	0.253***	0.263***	0.272***	0.273***
Area of residence (control group: North West)				
North East	-0.043	-0.056	-0.060	-0.060
Center	-0.002	-0.030	-0.006	-0.007
South and Islands	0.071*	0.074	0.067	0.069*
Number of inhabitants (control group: <5000 inhabitnants)				
From 5.001 to 20000	-0.043	-0.069	-0.053	-0.051
From 20001 to 100000	-0.056	-0.096	-0.066	-0.064
From 100001 to 500000	0.0072	0.012	-0.001	0.009
Over 500000	-0.070	-0.058	-0.078	-0.080
Education attainment (control group: primary school)				
Secondary school	-0.110***	-0.086*	-0.094**	-0.094**
University	-0.248***	-0.276***	-0.232***	-0.232***
Income (control group: first quartile)				
II Quartile	-0.033	-0.048	-0.030	-0.026
III Quartile	-0.073	-0.071	-0.072	-0.070
IV Quartile	-0.081	-0.105	-0.088	-0.085
Information channels				
Tv Only	-0.028	-0.069	-0.031	-0.028
Radio	-0.016	-0.000	0.0032	-0.007
Newspapers	-0.107***	-0.098**	-0.101***	-0.107***
Internet	-0.138***	-0.131***	-0.142***	-0.158***
Political leaders	-0.042	-0.006	-0.031	-0.03
Friends and relatives	-0.023	-0.037	-0.019	-0.028
Reliability and use of information				
Quality of information (control group: bad)				
Good		0.043		
Sufficient		0.015		
Use of information (control group: no use)				
Use		-0.125***		

Table 6 – The level of knowledge on economic data and its possible determinants

Importance of information (control group: not important)				
Important			-0.154*	0.332***
Desire to be informed (control group: no desire)				
Desire			-0.073**	-0.090***
Time control (control group: 2007)				
2009	-0.292***		-0.435***	
2010	-0.247***			
2011	-0.423***	-0.151***	-0.563***	
2012	-0.245***	0.024	-0.380***	0.081***
2013	-0.414***	-0.149***	-0.549***	-0.090***
2014	1.364***	1.612***	1.222***	1.681***
Constant	0.306***	0.165	0.502***	0.050
Number of observations	4,923	3,659	4,492	
R^2	0.508	0.531	0.512	0.502

*** p<0.001; ** p<0.05; * p<0.1

Source: authors' elaborations on ISTAT data

6. Conclusions

Surveys conducted since 2007 indicate that the level of knowledge of economic data of Italian consumers is relatively low: response rates are most of the times below the 50% threshold and accuracy of response is seldom assured. Results available from similar surveys (see for instance Curtin, 2008; 2009, and Papacostas, 2008) show that these findings are similar to those emerging on average in EU and the US. The analysis performed in this paper suggests, however, a high variability of the level of individual knowledge: this finding may be interpreted as a confirmation of the rational inattention hypothesis of Mankiw and Reis (2002) and Reis (2006), according to which information is costly and hence agents may rationally choose to update it only sporadically, the frequency of updating and the level of individual knowledge depending on the interactions with various factors, including the level of education, economic conditions, the importance assigned to information and the media used to acquire it. Econometric findings show that the frequency of updating and the level of individual knowledge grow with the level of education, the importance assigned to statistical information and the use of newspapers and the internet. On the other hand, no evidence of the importance of private channel of information (contacts with friends and relatives) emerge from the analysis. Knowledge also increases with age and is higher for men than for women, while economic conditions seem to have a significant effect on the frequency of updating, but not on the level of knowledge itself. No effect is found for opinions on the quality of the statistical and economic debate on the media.

These results have interesting implications for economic theory, policy makers and statistical producers alike. From a theoretical point of view, the data support the "rational inattention" hypothesis, providing evidence of deviation from the standard approach of full rationality. If agents are not always fully rational, possible delays in information acquisition patterns have to be taken into account by policy makers in designing appropriate

interventions, for example using appropriate communication tool to inform citizens about important policy economic decisions or taking into account the lack of knowledge of statistical data when estimating expected result from them. Results provide also very interesting evidence for official statistical agencies: first of all, it clearly emerges that an increase in the media exposure (as it was the case in the aftermath of the economic crisis) favours an increase in individual knowledge of the data. Moreover, in order to ensure a better translation of information available in effective knowledge statistical education programmes should be promoted at all school levels, but especially in the elementary school. Finally, statistical agencies should largely use innovative visualisation tools, in order to help the users to understand the "message" emerging from data without being obliged to go through complex and dense statistical tables.

References

Blinder A. S., Krueger A. B., (2004). "What Does the Public Know About Economic Policy, and How does It Know It?", *Nber Working Papers, n. 10787, September*.

Bryan M.R., G. Venkatu G. (2001), "The Curiously Different Inflation Perspectives of Men and Women", *Economic Commentary*, Federal Reserve Bank of Cleveland.

Carroll C.D. (2006), "The Epidemiology of Macroeconomics Expectations", in *The Economy as an Evolving Complex System III*. Larry Blume and Steven Durlauf (eds.) Oxford University Press.

Curtin, R. T. (2003). Current Research and Development Agenda for the U. S. Consumer Surveys. OECD workshop, Brussels, Belgium, November 21.

Curtin R. (2008), "What US Consumers know about Economic Conditions", in *Statistics, Knowledge and Policy 2007: Measuring and Fostering the Progress of Societies*, OECD, Paris.

Curtin R. (2009), "What US Consumers Know About the Economy: the Impact of Economic Crisis on Knowledge", paper presented at the III OECD World Forum on "Measuring the Progress of Societies", Busan, Republic of Korea, October.

Curtin, R. T. (2015). Surveys of Consumers – Sample Design. University of Michigan, last accessed August 26, 2015, at https://data.sca.isr.umich.edu/fetchdoc.php?docid=24773.

D'Urzo A., Gamba M., Giovannini E., Malgarini M. (2009), "What do Italian Citizens Know About the Progress of their Country? Results from the 2009 ISAE/OECD Survey", paper presented at the III OECD World Forum on "Measuring the Progress of Societies", Busan, Republic of Korea, October.

Fullone F., Gamba M., Giovannini E., Malgarini M., (2008), "What do Citizens Know About Statistics? The Results of an OECD/ ISAE Survey on Italian Consumers", in *Statistics, Knowledge and Policy 2007: Measuring and Fostering the Progress of Societies*, OECD, Paris.

Fullone F., Martelli B. (2008), "Re-Thinking the ISAE Consumer Survey Processing", *Documento di Lavoro*, ISAE n. 92, February.

Giovannini E., J. Oliveira Martins, M. Gamba (2008), "Statistics, Knowledge and governance", paper presented at the Workshop "Committing Science to Global Development", Lisbon, 29-30 September.

Giovannini E., Malgarini M. (2012), "What do Italian Consumers think of Economic data? An Analysis based on the ISTAT Consumers' survey", MPRA Working Paper n. 54125, Munich.

Giovannini E. (2014), *Scegliere il futuro. Conoscenza e politica al tempo dei Big Data*, Il Mulino, Bologna.

ISTAT (2014) La conoscenza dei dati economici da parte dei consumatori italiani, Statistiche Focus, giugno, Roma.

Mac Kenzie, R. (2005), "Assessing and Minimizing the Impact of Non-Response of Survey Estimates". Analysis of Key Issues and Main Findings, OECD Taskforce on Improvement of Response Rate and Minimisation of Respondent Load, Paper presented at the Joint European Commission – OECD Workshop on International Development of

Business and Consumer Tendency Surveys, Brussels, 14-15 November, available at: http://www.oecd.org/dataoecd/56/16/356340123.pdf.

Mankiw, N.G., R. Reis, 2002. "Sticky Information Versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve," *QJE* 117(4), 1295-1328.

Papacostas S. (2008), "Special Eurobarometer: European Knowledge on economical indicators", in Statistics, Knowledge and Policy 2007: Measuring and Fostering the Progress of Societies, OECD, Paris.

Reis R. (2006), "Inattentive Consumers", *Journal of Monetary Economics*, vol. 53, (8), pp. 1761-1800.

Reis R. (2009), "A sticky information general equilibrium model for policy analysis", *NBER Working Paper*, n. 14732, February, available at http://www.nber.org/papers/w14732.

Souleles, N. (2004) Expectations, heterogeneous forecast errors and consumption: micro evidence from the Michigan consumer sentiment surveys, *Journal of Money, Credit and Banking*, 36, 39–72.

United Nations (2014), Handbook Economic Tendency Surveys – Draft, available on line at: https://unstats.un.org/unsd/nationalaccount/consultationDocs/draftETS-Handbook-May2014.pdf.

Statistical Appendix

Table A – Importance and desire of being informed

Importance of being informed						
Percentage share of respondents answering:	2007	2009	2011	2012	2013	2014
Very important	Na	23.0	30.4	34.0	37.4	42.8
Fairly important	Na	48.2	39.4	40.1	39.7	40.2
Not important, nor unimportant	Na	17.7	14.7	14.8	9.7	6.7
Relatively not important	Na	4.7	5.0	4.8	6.9	4.9
Not important at all	Na	5.3	5.6	4.7	4.2	4.0
Don't know/no opinion	Na	0.9	4.8	1.6	2.1	1.4
Desire of being more informed						
Yes	51.5	40.7	40.6	46.6	43.4	48.7
No	43.8	55.6	52.5	51.2	53.5	46.9
Don't know	4.7	3.7	6.9	2:1	3.2	4.4

Source: authors' elaborations on ISTAT data.

Table B – Use of information for strategic decisions

In your private life, do you use the information we have talked about for your economic decisions about financial investments, relevant purchases and others?					
Percentage share of respondents answering:	2010	2011	2012	2013	2014
A lot	0.9	1.0	2.1	2.0	4.7
A bit	6.0	14.8	13.2	14.3	12.6
Not much	10.4	20.2	21.2	18.2	17.3
Not at all	76.9	56.9	61.1	62.0	62.6
Don't know	4.6	4.3	1.9	2.8	2.5
Refuse to answer	1.0	2.7	0.5	0.7	0.3

Source: authors' elaborations on ISTAT data.

Table C – Information channels

Percentage share of respondents answering:	2007	2009	2010	2011	2012	2013	2014
Television	82.7	91.2	86.9	84.9	87.9	86.9	82.7
Radio	17.2	17.7	16.4	17.2	16.6	14.0	20.6
Newspapers, magazines	49.4	49.1	47.6	44.2	39.6	33.9	44.2
Internet	20.6	24.8	31.0	35.9	30.9	35.5	43.4
Political and opinion leaders	8.2	4.3	5.3	4.5	3.5	5.2	5.2
Friends and relatives	9.9	7.5	11.1	10.8	9.9	10.1	14.2
Scientific publications	nd	nd	nd	3.4	1.9	0.6	5.5
Dont'know	3.1	1.0	4.0	1.0	0.7	0.4	1.3
Refuse to answer	2.0	0.2	0.8	2.4	0.2	0.3	0.5

Source: authors' elaborations on ISTAT data.

Fecondità e maternità: un sistema integrato per la misurazione di fenomeni sanitari e socio-demografici[1]

Tiziana Tuoto, Marina Attili, Alessandra Burgio, Rossana Cotroneo, Claudia Iaccarino, Sabrina Prati, Francesca Rinesi, Fabio Rottino, Laura Tosco, Luca Valentino

Sommario

Il lavoro descrive i passi iniziali di progettazione e sperimentazione di un complesso progetto di integrazione tra fonti, finalizzato alla realizzazione del "Sistema integrato sugli esiti del concepimento". L' integrazione delle fonti è indispensabile per ottenere un quadro completo e dettagliato sui principali aspetti demografici e socio-sanitari degli esiti dei concepimenti, dati i profondi cambiamenti degli ultimi decenni nella regolamentazione sulla raccolta dei dati amministrativi, legati a questioni di semplificazione e di privacy. Nel lavoro si delinea una strategia di integrazione, mettendone in rilievo gli aspetti metodologici e prediligendo soluzioni che possano essere facilmente estese, portate a regime ed inserite in processi di produzione corrente.

Parole chiave: integrazione di dati, record linkage, indicatori demo-socio-sanitari

Abstract

This paper describes the initial steps of design and testing of a complex and ambitious integration project between sources aimed at the implementation of a system called "Integrated system on pregnancy outcome". Given the deep changes in the regulation on administrative data collection occurred in the last decades and related to simplification and privacy issues, the integration of sources is essential to obtain a complete and detailed picture on the main social, health and demographic aspects of pregnancy outcome. The paper outlines a strategy for integration, highlights the methodological issues and proposes solutions that can be easily extended and included into current production processes.

Keywords: data integration, record linkage, demographic, social and health indicators.

[1] Gli articoli pubblicati impegnano esclusivamente gli Autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat. Sebbene il lavoro sia frutto dell'opera di tutti gli autori, sono da attribuire: i paragrafi 1, 3, 3.2, 4, 5 a Tiziana Tuoto; il paragrafo 2, 2.2, 2.3, 2.4 a Alessandra Burgio e Sabrina Prati, il paragrafo 2.1 a Marina Attili Alessandra Burgio, Rossana Cotroneo, Claudia Iaccarino, Tiziana Tuoto; il paragrafo 3.1 a Laura Tosco e Tiziana Tuoto; il paragrafo 4.1 a Alessandra Burgio, Rossana Cotroneo, Francesca Rinesi, Laura Tosco, Tiziana Tuoto e Luca Valentino, il paragrafo 4.2 a Marina Attili, Fabio Rottino, Claudia Iaccarino; il paragrafo 4.3 a Marina Attili, Claudia Iaccarino e Tiziana Tuoto. Per contattare gli autori: tuoto@istat.it.

1. Introduzione

L'integrazione di dati provenienti da fonti amministrative e/o da indagini statistiche è diventata negli ultimi anni una priorità per la statistica ufficiale: essa permette di migliorare la qualità dei dati e l'efficienza delle analisi attraverso un poderoso utilizzo di dati di fonte amministrativa e quindi disponibili senza ulteriori carichi economici e riducendo il carico statistico sui rispondenti. Tra le tecniche di integrazione di dati, quella del record linkage ha come obiettivo l'identificazione della stessa unità statistica, tipicamente memorizzata in archivi diversi e descritta con chiavi identificative non perfettamente coincidenti. Le soluzioni ai problemi di record linkage, studiate in letteratura e adottate nella pratica, si rifanno a svariati approcci e metodologie, che coinvolgono soluzioni euristiche, metodi probabilistici, approcci bayesiani, soluzioni basate sulle tecniche di data-mining e/o machine learning. Tuttavia i problemi di record linkage sono fortemente caratterizzati dalla natura dei dati da abbinare e dagli obiettivi dell'abbinamento e nessuna delle metodologie o delle tecniche proposte è la più efficace o la più efficiente, per tutte le diverse applicazioni, ma è necessario adattare il processo di linkage agli specifici requisiti dei dati in esame.

Il presente lavoro descrive un complesso progetto di integrazione tra fonti, finalizzato alla realizzazione di un "Sistema integrato sugli esiti del concepimento" il cui obiettivo è quello di misurare i principali aspetti socio-demografici e sanitari degli esiti dei concepimenti. Tale integrazione permetterà di superare il gap informativo venutosi a creare alla fine degli anni '90 con l'entrata in vigore delle leggi sulla semplificazione amministrativa e sulla privacy, che hanno generato profondi cambiamenti nella raccolta dei dati per le statistiche sanitarie e socio-demografiche. Solo attraverso l'integrazione di fonti di dati diverse è ora possibile fornire una visione completa e allo stesso tempo dettagliata dei fenomeni legati alla maternità e alla fertilità. Tale operazione però presenta diversi elementi di complessità, legati essenzialmente a due fattori: in primo luogo la disomogeneità delle fonti coinvolte, che sono di diversa natura (amministrativa, sanitaria, statistica), fanno riferimento a universi solo parzialmente sovrapposti, riportano dati raccolti con differente livello di accuratezza; in secondo luogo la mancanza, per la tutela della privacy, di informazioni "forti" per l'identificazione degli individui è un fattore estremamente rilevante per l'applicabilità stessa e l'efficacia delle tecniche di integrazione. In ogni caso, l'operazione di integrazione e di creazione di questo sistema integrato è irrinunciabile per la determinazione di misure sulle caratteristiche delle donne in gravidanza, sulla salute delle donne e dei bambini durante la gravidanza e nel periodo post-parto e sulla salute delle madri e dei bambini. In particolare, il sistema integrato permetterà di produrre indicatori sul tipo di parto, sui parti pretermine e multipli, sulla distribuzione del peso alla nascita e dell'età gestazionale. Il sistema renderà possibile studiare le principali relazioni tra tali fenomeni e la relazione tra questi e le grandezze che li determinano (stato civile, cittadinanza, livello d'istruzione e condizione lavorativa).

Il presente contributo è organizzato come segue: nel paragrafo 2 sono descritte le esigenze conoscitive legate al sistema integrato degli esiti dei concepimenti e definiti alcuni dei principali output attesi, insieme ad una sintetica ricognizione delle fonti e della loro qualità; nel paragrafo 3 vengono fornite la descrizione del sistema integrato, la definizione del fenomeno di riferimento adottata nel resto dell'articolo e le scelte alla base delle metodologie di integrazione sperimentate, evidenziando i requisiti della procedura di abbinamento e lo strumento utilizzato; il paragrafo 4 riporta i primi incoraggianti risultati

ottenuti con alcune sperimentazioni effettuate sui dati del 2007 per la regione Emilia-Romagna. Infine, nel paragrafo 5 sono raccolte alcune conclusioni, si prospettano i passi operativi e futuri indirizzi di ricerca.

2. L'esigenza conoscitiva

Il principale obiettivo del Sistema integrato sugli esiti dei concepimenti (SIEC) è consentire la costruzione di una molteplicità di indicatori su parti, nascite, interruzioni di gravidanza, mortalità perinatale mediante l'integrazione delle fonti demografiche e sanitarie. L'integrazione delle fonti permetterà di disporre di una base dati affidabile e di buona qualità attraverso la validazione e la eventuale correzione delle informazioni disponibili nei singoli flussi. Gli output dell'integrazione sono molteplici e differenziati a seconda che si assuma come unità di analisi la donna, il parto o l'esito del concepimento (nato vivo, nato morto, interruzione volontaria di gravidanza -IVG- e aborto spontaneo -AS), oppure un'ottica trasversale (più fonti con uno stesso riferimento temporale) o longitudinale (la stessa fonte in periodi diversi) per "seguire" nel tempo le storie riproduttive di differenti coorti di donne. Gli output del sistema possono essere inoltre classificati secondo una logica prevalentemente di processo o di prodotto.

Per output di processo si intende la possibilità di migliorare la qualità e la completezza delle informazioni rilevate in ciascuna fonte, nonché di armonizzare e rendere compatibili le informazioni desumibili per le stesse unità da diversi flussi produttivi. Esiste una gerarchia del livello di qualità delle informazioni che vengono rilevate nelle diverse fonti amministrative. In ogni fonte la qualità è più alta per quelle informazioni che sono disponibili nei registri di base e/o che sono necessarie per le finalità istituzionali del titolare della fonte. Ad esempio, per le rilevazioni di fonte anagrafica la qualità e la copertura delle informazioni demografiche desumibili direttamente dai registri di popolazione (data di nascita, stato civile, cittadinanza, luogo di nascita e di residenza) è massima. Nelle fonti amministrative sanitarie, ad esempio le SDO, la qualità è massima per tutte le variabili che descrivono il ricovero (età e sesso, diagnosi alla dimissione, procedure o interventi effettuati, tipo di ricovero, ecc.). Nel caso dei dati sulle interruzioni volontarie di gravidanza e sull'abortività spontanea sono di elevata qualità tutte le informazioni di carattere demografico e socio-sanitario che vengono desunte direttamente dalla cartella clinica.

La qualità è inoltre variabile nel tempo e nello spazio. Esemplare è a questo proposito il caso della rilevazione sui Certificati di assistenza al parto (CEDAP). La rilevazione ha avuto un percorso faticoso, a distanza di 10 anni dall'inizio dell'acquisizione di questa preziosa fonte informativa si registra ancora una situazione molto eterogenea sul territorio, con punte di eccellenza in alcune regioni e situazioni meno soddisfacenti in altre.

Per output di prodotto si intende tutto ciò che è rilasciato, o rilasciabile, all'utenza esterna tramite l'accesso al sistema. Nei successivi sottoparagrafi 2.2, 2.3 e 2.4 si evidenzieranno alcuni output di prodotto propri dello sfruttamento integrato delle fonti reso possibile dal sistema e quindi aggiuntivi rispetto a quelli già elaborabili attingendo ad ogni singola fonte.

2.1 Ricognizione delle fonti

Affinché i risultati di un processo di integrazione siano di buona qualità, è necessario prevedere una preliminare e approfondita analisi dei dati provenienti dalle distinte fonti che si intende integrare. Infatti, per scegliere in modo efficiente le informazioni da utilizzare ai fini del record linkage, è fondamentale conoscere la natura dei dati, il numero di record da trattare, la presenza in essi di identificatori univoci e di variabili con alto potere discriminante, la presenza più o meno consistente di dati mancanti e/o con codifiche errate.

In prima istanza sono state considerate le seguenti fonti di dati, costituite da indagini, archivi amministrativi o registri di varia natura:

CEDAP – la rilevazione sui certificati di assistenza al parto (distinguendo ove possibile tra parti, nati vivi, nati morti);

P4 – la rilevazione Istat degli iscritti in anagrafe per nascita;

IVG – l'indagine Istat sulle interruzioni volontarie di gravidanza;

AS - l'indagine Istat sugli aborti spontanei;

SDO - le schede di dimissione ospedaliera (distinguendo parti, nati vivi, nati morti, interruzioni volontarie di gravidanza, aborti spontanei).

Le tavole successive riportano i principali risultati della ricognizione per le fonti considerate a questo stadio di progettazione del sistema. Nella tavola 2.1 per ogni fonte è indicato se l'Istat è l'ente titolare dei dati, mentre la successiva tavola 2.2 specifica il periodo di riferimento, la dimensione in termini di numero di eventi registrati in Italia, la presenza e la qualità di variabili identificative per ciascuna fonte di dati. Per motivi di privacy, le fonti sanitarie giungono all'Istat completamente prive degli identificativi delle persone coinvolte dall'evento.

Tavola 2.1 - Tipo di fonte ed ente gestore

	Nome della fonte	Titolarità dell'Istat	
		Si	No
1a	CEDAP parto		X
1b	CEDAP nato vivo		X
1c	CEDAP nato morto		X
3	P4	X	
4	IVG	X	
5	AS	X	
6 a	SDO parto		X
6 b	SDO nato vivo		X
6 c	SDO nato morto		X
6 d	SDO ivg		X
6 e	SDO as		X

Tavola 2.2 - Numerosità e qualità dell'informazione

	Nome della fonte	Anno di riferimento	Numero di unità in Italia (annuali approx)	Presenza di variabili identificative (*)	Qualità (**) delle variabili disponibili	
					Bassa	Buona
1a	CEDAP parti	2002-2007	520.000	NO	X	
1b	CEDAP nati vivi	2002-2007	525.000	NO	X	
1c	CEDAP nati morti	2002-2007	1.500	NO	X	
3	P4	1999-2009	570.000	SI		X
4	IVG	1982-2009	125.000	NO		X
5	AS	1982-2009	77.000	NO		X
6a	SDO parti	2001-2009	550.000	NO		X
6b	SDO nati vivi	2001-2009	560.000	NO		X
6c	SDO nati morti	2001-2009	1.700	NO		X
6d	SDO ivg	2001-2009	120.000	NO		X
6e	SDO as	2001-2009	85.000	NO		X

Note: (*) per variabili identificative di intendono i Codici Fiscali, nomi e cognomi (eventualmente standardizzati o codificati), codici univoci di evento …

(**) per qualità delle variabili si fa riferimento a valori mancanti, errori nella codifica, incompatibilità con altre variabili.

2.2 Linkage CEDAP e iscritti in anagrafe per nascita (P4)

L'integrazione tra CEDAP e iscritti in anagrafe per nascita consente, per il sottoinsieme di record comuni, ovvero i nati vivi della popolazione residente, di validare le informazioni demografiche di base del nato e dei genitori (output di processo). Un esempio è il caso della cittadinanza, informazione che in termini di qualità e completezza è maggiore nella rilevazione di fonte anagrafica.

Dal lato degli output di prodotto, invece, l'integrazione tra le due fonti consente di recuperare completamente il debito informativo su alcune relazioni tra nascite e parti creatosi a partire dal 1999 con la soppressione della rilevazione individuale dei nati effettuata dall'Istat fin dal 1926 presso lo stato civile. Infatti, mentre nelle attese del legislatore il contenuto informativo dei CEDAP avrebbe dovuto sostituire completamente le rilevazioni soppresse, le difficoltà di fatto incontrate nella rilevazione dei CEDAP, di cui si accenna al paragrafo precedente, hanno reso lacunoso il patrimonio informativo acquisito.

E' pertanto possibile elaborare e diffondere con periodicità annuale i principali indicatori sulla gravidanza, il parto e le caratteristiche dei nati della popolazione residente, distinti per le principali caratteristiche demografiche e sociali delle madri (stato civile, cittadinanza, titolo di studio, condizione professionale).

E' inoltre possibile mettere a disposizione dell'utenza una base di micro dati validati, priva di elementi che consentano di risalire agli individui (file standard o file per la ricerca) che fornisce agli studiosi strumenti per analizzare gli effetti delle principali determinanti socio-demografiche e sanitarie rispetto a diversi tipi di esiti (nato vivo, nato morto) oppure nati a rischio (fortemente pre-termine, fortemente sottopeso), nascite gemellari, nascite da parti non fisiologici (in particolare parti cesarei).

2.3 Linkage CEDAP e SDO

Le due rilevazioni (CEDAP e SDO), per quanto concerne l'evento parto, afferiscono alla stessa popolazione. In alcune regioni esiste una corrispondenza uno a uno in quanto se non viene compilato il CEDAP non viene rimborsato, o viene rimborsato in misura minore, il ricovero in caso di parto. L'integrazione può essere effettuata assumendo come unità di analisi il parto oppure il nato vivo. Il primo output di processo di rilievo dell'integrazione

SDO-CEDAP consiste nella possibilità di ricondurre la SDO parto alla SDO nato (attualmente i due flussi sono separati e non riconducibili allo stesso evento), ampliando così notevolmente il ventaglio informativo.

Si possono inoltre sfruttare tutte le informazioni delle SDO e dei CEDAP per effettuare controlli di coerenza tra le diverse fonti e correggere situazioni di incompatibilità dovute ad errori o scarsa qualità: si possono utilizzare le due fonti integrate sia per recuperare eventuali mancate risposte parziali, sia per recuperare le mancate risposte totali utilizzando le SDO come universo di riferimento.

Dal lato degli output di prodotto, l'integrazione permette di calcolare i principali indicatori sui parti. Infatti, le SDO arricchiscono i CEDAP con informazioni di tipo clinico (diagnosi principale e secondarie, procedure diagnostiche, interventi principali e secondari, informazioni sul ricovero e la dimissione) che permettono di studiare in maniera più dettagliata la medicalizzazione del percorso gravidanza-parto nonché di evidenziare l'adozione dei diversi protocolli di assistenza nei punti nascita in relazione alle principali caratteristiche dei centri nascita.

Di particolare rilievo sono le possibilità offerte dal sistema per lo studio dei parti cesarei. Ad esempio, è possibile procedere al calcolo delle classi di Robson[2] a livello regionale o sub-regionale per evidenziare le aree di maggiore criticità rispetto all'eccessivo ricorso al cesareo e metterle in relazione con le informazioni di contesto sul centro nascita.

Inoltre l'integrazione SDO-CEDAP, insieme ad una terza fonte di cui è titolare il Ministero della Salute e che si chiama "Struttura e attività degli Istituti di cura" permette di disporre di informazioni di contesto sul centro nascita, per sapere quanti posti letto sono disponibili nel reparto di ostetricia-ginecologia e quante culle nei nidi. Un tema estremamente attuale se si pensa alle chiusure previste per i centri di nascita che effettuano un numero di parti annui al di sotto di una soglia prefissata. nonché informazioni relative alla struttura dove è stato effettuato il ricovero ospedaliero.

Inoltre, l'integrazione permetterà di calcolare i principali indicatori sulla salute perinatale: ad esempio, gli indicatori raccomandati dal Progetto Europeristat (European Perinatal Health, 2008)., che consentono il confronto con gli altri Paesi europei, potranno essere calcolati anche a livello regionale e sub-regionale con periodicità annuale in modo da monitorare il fenomeno della salute riproduttiva sul territorio secondo quanto è richiesto dalle raccomandazioni internazionali.

Sarà infine possibile creare una base dati individuale anonima che consenta agli esperti che operano nel settore e agli studiosi di analizzare le relazioni tra le principali caratteristiche individuali e di contesto e l'evento nascita con i suoi possibili esiti.

2.4 Linkage SDO e IVG, SDO e AS

Il principale output di processo derivante dall'integrazione tra SDO-IVG-AS è la possibilità di validare alcune informazioni socio-demografiche di rilievo per l'analisi dei

[2] Questa classificazione, prendendo in esame la precedente storia ostetrica (parità), l'età gestazionale, la presentazione e le modalità del travaglio, propone 10 categorie mutuamente esclusive che quantificano 10 sottopopolazioni, entro ciascuna delle quali è possibile analizzare la frequenza di ricorso al parto cesareo (Can we reduce the caesarean section rate? Michael Stephen Robson, Best Practice & Research Clinical Obstetrics & Gynaecology Vol. 15, No. 1, pp. 179-194, 2001Harcourt Publishers Ltd).

fenomeni come lo stato civile e il titolo di studio, rilevate da tempo e affidabili nelle fonti Istat sull'abortività; viceversa per la fonte SDO sono da considerarsi di buona qualità le informazioni sul ricovero (in regime ordinario con pernottamento o in day hospital senza pernottamento, durata della degenza e intervento effettuato).

La fonte SDO viene attualmente utilizzata come universo di riferimento per la correzione della mancate risposte totali presenti nei flussi IVG e AS. Con l'integrazione delle fonti in esame sarà possibile analizzare i protocolli utilizzati negli ospedali per codificare l'IVG e l'AS (analisi della diagnosi principale e delle diagnosi secondarie, dell'intervento/procedura principale e degli interventi/procedure secondari). Questo può fornire dei criteri più mirati per selezionare i casi di IVG e AS nelle SDO e quindi migliorare la costruzione dell'universo di riferimento per la stima dei dati mancanti.

In termini di output di prodotto, l'integrazione permette di arricchire e dettagliare le informazioni medico-sanitarie rilevate con i moduli Istat e calcolare indicatori sulle caratteristiche del ricovero per le donne che subiscono una interruzione della gravidanza o un aborto spontaneo: interventi, terapie e procedure effettuate, presenza di eventuali complicazioni durante il ricovero.

3. Le metodologie di integrazione per il sistema integrato sugli esiti dei concepimenti

Per la realizzazione del prototipo del sistema integrato sugli esiti dei concepimenti, è stata individuata come unità di riferimento l'evento "esito del concepimento": per tale fenomeno il sistema è in grado di distinguere le seguenti modalità

- nato vivo;
- nato morto;
- aborto spontaneo;
- interruzione volontaria di gravidanza.

In questa prima fase di progettazione e sperimentazione della fattibilità dell'intero sistema, eventi riferiti a concepimenti diversi che coinvolgono la stessa donna, nell'arco dello stesso anno, sono distinti e non immediatamente riconducibili. Si è deciso quindi di costruire il sistema integrato coinvolgendo principalmente le basi di dati riguardanti i certificati di assistenza al parto (CEDAP), distinguendo ove possibile tra parti, nati vivi, nati morti; gli iscritti in anagrafe per nascita (P4); le indagini Istat sulle interruzioni volontarie di gravidanza e sugli aborti spontanei (IVG); le schede di dimissione ospedaliera (SDO), distinguendo le interruzioni volontarie di gravidanza, gli aborti spontanei, i parti, i nati vivi e i nati morti. Come esposto nel paragrafo 2, è possibile ampliare le potenzialità conoscitive del sistema collegando eventi che vedono protagonista la stessa donna, anche in tempi diversi; a tal fine è necessario un approfondimento successivo che studi la possibilità di agganciare eventi diversi riferiti alla stessa donna, anche in un approccio longitudinale, per la ricostruzione della vita riproduttiva.

Il sistema sugli esisti del concepimento può essere definito secondo diversi livelli di integrazione, ad esempio è possibile prevedere integrazioni di tipo macro (a livello di indicatori e di aggregati) o di tipo micro (a livello di singolo record per ogni evento considerato). L'integrazione a livello micro, laddove possibile, è quella che salvaguarda il patrimonio informativo più ampio, quindi si è deciso di privilegiare inizialmente questo tipo

di metodologia, mettendo in campo tutte le risorse e le competenze per valutare la fattibilità del sistema in cui ogni fonte è integrata a livello del singolo record. Opportune valutazioni relative ad integrazioni di tipo macro e ad eventuali interazioni tra le due operazioni possono essere ulteriormente prese in considerazione.

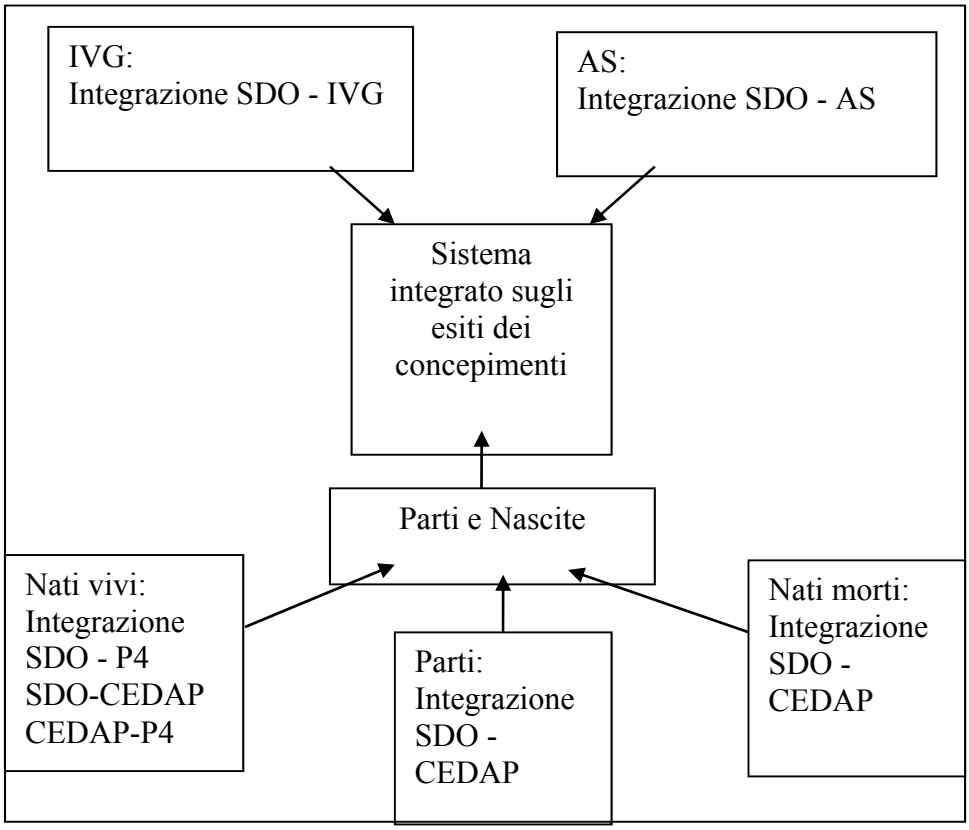
Nella scelta delle metodologie da testare per lo studio di fattibilità e la realizzazione, almeno in fase prototipale, del sistema integrato sugli esiti del concepimento, si persegue l'obiettivo di mettere a punto una strategia di integrazione che possa essere facilmente estesa, portata a regime ed inserita in un processo di produzione corrente.

Nell'ambito delle metodologie per l'integrazione dei microdati, il record linkage probabilistico permette di corredare il sistema integrato con una serie di informazioni e misurazioni relative al processo di integrazione che i dati hanno subito, tra cui ad esempio la probabilità di corretto abbinamento per ogni singolo record considerato nel sistema e altre misure di qualità complessive relative all'intero processo di integrazione. Le applicazioni di record linkage devono essere corredate da informazioni sulla qualità del linkage da utilizzare con apposite metodologie di stima, volte ad assicurare la qualità delle analisi condotte sui dati abbinati. Infatti, trattare tali dati ignorando il processo di integrazione e gli eventuali errori da questo generati, può portare, in generale, a stime distorte. Tali informazioni sono di fondamentale importanza per la corretta analisi e interpretazione statistica dei fenomeni legati alla fecondità e alla maternità che attraverso il sistema si vogliono studiare.

Con l'ottica di risolvere il problema dell'integrazione delle fonti relative agli esiti dei concepimenti in modi facilmente riproducibili in contesti di produzione dei dati, è stato utilizzato il toolkit RELAIS (RELAIS, 2011), come strumento privilegiato per la realizzazione dei processi. Tale strumento mette a disposizione una serie di metodi e tecniche diverse per l'esecuzione di ogni singola fase di un processo di record linkage. Lo strumento è quindi particolarmente indicato per la sperimentazione e realizzazione di progetti di integrazione che coinvolgono fonti di dati così diverse per quantità e qualità, e quindi per la creazione di un sistema complesso e articolato come quello integrato per gli esiti del concepimento. In fase sperimentale, una ulteriore valutazione della robustezza dei risultati dell'abbinamento probabilistico è stata effettuata attraverso il confronto con i risultati di una procedura di abbinamento deterministico, (descritta nel successivo paragrafo 4.2) che è stata sviluppata *ad hoc* nel contesto delle attività di abbinamento deterministico di dati demo-sociali. Tale procedura è stata arricchita e perfezionata nel corso degli anni e ha dimostrato di perseguire ottimi risultati, testati soprattutto in situazioni analoghe (Attili e Valentino, 2010). In ogni caso, per la costruzione del sistema integrato degli esiti dei concepimenti è più opportuno l'utilizzo di strumenti di integrazione che siano basati su metodologie statistiche ben consolidate e validate dalla comunità scientifica.

In questa fase sperimentale finalizzata alle valutazioni della fattibilità del sistema integrato, tutte le fonti che riportano informazioni relative allo stesso fenomeno-evento sono state abbinate a coppie e, successivamente, sono stati considerati i risultati di ogni integrazione per individuare sia l'insieme minimo che quello massimo di eventi abbinati. Queste operazioni, seppure più dispendiose rispetto ad abbinamenti successivi in sequenza o a cascata, permettono di tenere sotto controllo il potere informativo di ogni fonte e predisporre le operazioni successive avvalendosi del massimo dell'informazione possibile. Il progetto di integrazione per la costruzione a livello micro del prototipo del sistema integrato dei concepimenti prevede i seguenti abbinamenti come riportati nello schema 3.1.

Schema 3.1. Sintesi delle integrazioni tra fonti per la costruzione del sistema sugli esiti dei concepimenti



Operativamente, seguire questo schema significa procedere parallelamente con gli abbinamenti:

1. SDO – IVG, il cui risultato intende costruire la parte del sistema integrato relativa alle interruzioni volontarie di gravidanza
2. SDO – AS, il cui risultato intende costruire la parte relativa agli aborti spontanei per cui è stato necessario rivolgersi ad una struttura ospedaliera
3. SDO - P4, il cui risultato intende costruire la parte del sistema integrato relativa ai nati vivi
4. CEDAP - P4, il cui risultato intende costruire la parte relativa ai nati vivi
5. SDO - CEDAP, il cui risultato intende costruire la parte relativa ai parti, ai nati vivi e ai nati morti.

I risultati degli abbinamenti 3), 4) e 5) andranno a loro volta integrati e analizzati per verificare:

- i casi linkati in tutte le applicazioni
- i casi linkati solo da una o due applicazioni

L'integrazione dei risultati parziali 3), 4) e 5) permetterà di ricostruire l'archivio dei parti e delle nascite e a sua volta andrà integrato con i risultati 1) e 2) per costituire il sistema integrato sugli esiti dei concepimenti.

3.1 Il record linkage e lo strumento RELAIS

Come noto, il record linkage indica un processo di abbinamento di record che ha come obiettivo l'identificazione della stessa unità statistica, memorizzata in archivi diversi o presente più volte nella stessa lista, anche in assenza di identificatori univoci o quando questi sono affetti da errori. L'identificazione dell'unità in archivi di diversa natura avviene attraverso chiavi comuni, presenti nei vari file; le chiavi possono essere anche non perfettamente corrispondenti. La complessità del record linkage dipende da molteplici aspetti, principalmente legati all'assenza di identificatori univoci o alla presenza di errori negli identificatori stessi.

Formalmente, l'obiettivo del linkage è identificare un'unità che può essere rappresentata in maniera differente in due diverse fonti dati A e B. In generale, le coppie che si intende classificare come abbinamenti (ossia a e b sono la stessa unità) o non abbinamenti (a e b sono due differenti unità) sono quelle dell'insieme Ω, prodotto cartesiano di A e B. Tale insieme ha cardinalità $n_A \times n_B$ ed è costituito da tutte le possibili coppie a, b (a,b | a $\in$ A, b $\in$ B). Per individuare le coppie che si riferiscono alla stessa unità, gli abbinamenti, si ricorre al confronto tra k variabili, "variabili di match", comuni alle due fonti di dati e associate alle unità. Tali variabili identificano in maniera univoca le unità, a meno, ovviamente, di errori o valori mancanti nelle variabili stesse; proprio a causa delle imperfezioni nelle variabili di match, l'abbinamento non può essere risolto attraverso l'utilizzo di un semplice "join" fra le due liste in esame. Il confronto tra le variabili viene effettuato per mezzo di un'opportuna funzione, scelta in base al tipo di variabile e alla sua qualità (in termini di completezza e correttezza). Per ogni coppia (a,b) $\in \Omega$, si definisce un vettore γ, detto "vettore dei confronti", i cui k elementi sono il risultato del confronto tra le variabili di match. Nel modello probabilistico per l'individuazione degli abbinamenti, si ipotizza che la distribuzione del vettore dei confronti sia una mistura di due distribuzioni, una generata dalle coppie (a,b) che effettivamente rappresentano la stessa unità, distribuzione *m*, e una generata dalle coppie (a,b) che rappresentano unità diverse, distribuzione *u*. A partire dalla stima di tali distribuzioni, è possibile costruire il peso composto di abbinamento (Fellegi and Sunter, 1969), dato dal rapporto delle verosimiglianze

$$r = \frac{m(\gamma)}{u(\gamma)} = \frac{\Pr(\gamma \mid M)}{\Pr(\gamma \mid U)},$$

dove *M* è l'insieme delle coppie che rappresentano degli abbinamenti e *U* è l'insieme delle coppie che non rappresentano degli abbinamenti, con $M \cup U = \Omega$ e $M \cap U = \varnothing$. In generale, la stima dei parametri delle distribuzioni viene generalmente ottenuta per mezzo

dell'applicazione dell'algoritmo EM (Jaro 1989). La stima dei parametri delle distribuzioni risulta molto complessa o addirittura impraticabile quando la dimensione dello spazio dei confronti Ω è dell'ordine dei milioni. A tali dimensioni si arriva rapidamente dato che lo spazio Ω, creato come prodotto cartesiano dei file da abbinare, cresce in maniera quadratica rispetto alla dimensione dei file di partenza. In tali casi si procede usualmente alla riduzione dello spazio di ricerca delle coppie attraverso l'applicazione di metodi di bloccaggio, sorting o indexing (Cibella, Tuoto, 2012).

Sulla base del rapporto r, le coppie sono ordinate e sottoposte ad un processo di classificazione negli insiemi M ed U:

- se il peso r è maggiore di una certa soglia T_m allora la coppia viene classificata come match;
- quando il suo peso è inferiore alla soglia T_u la coppia viene classificata come non match;
- per le unità il cui peso cade nell'intervallo $I=(T_u , T_m)$ non è possibile stabilire lo stato di abbinamento ma è necessario procedere ad un'ispezione manuale o comunque ad ulteriori analisi.

Secondo lo schema di decisione impostato da Fellegi e Sunter le due soglie, T_u e $T_{m,}$, sono fissate in modo che siano minimizzati sia gli errori di classificazione che la dimensione dell'area tra le soglie per cui non viene presa una decisione.

In numerose applicazioni, attraverso il linkage si mira ad individuare tra le coppie solo legami del tipo 1 a 1, in cui, cioè, una unità del file A viene abbinata con una sola unità del file B; in questi casi è necessario introdurre metodi di ottimizzazione che consentano di selezionare, tra tutte le coppie che coinvolgono le stesse unità della lista A e della lista B, quelle che rispettano il vincolo 1:1 e massimizzano la somma dei pesi r.

RELAIS (REcord Linkage At IStat) è un toolkit sviluppato presso l'Istat che mette a disposizione un insieme di tecniche per affrontare e risolvere problemi di record linkage (Cibella *et al.* 2007). RELAIS si basa sull'idea che un processo di record linkage può essere visto come costituito da diverse fasi per ognuna delle quali possono essere adottate diverse tecniche risolutive afferenti a diverse aree di conoscenza. La scelta della tecnica più appropriata da applicare dipende dal dominio di applicazione. RELAIS fornisce diverse tecniche per le diverse fasi di un processo di record linkage, consentendo di combinare tali tecniche in modo da ottenere il processo lavorativo ottimale per la specifica applicazione.

Nella costruzione del sistema integrato dei concepimenti è stata usata la versione 2.3 beta di RELAIS. Le caratteristiche specifiche del sistema possono essere trovate nel manuale utente, disponibile all' indirizzo http://www.istat.it/it/strumenti/metodi-e-software/software/relais. Rispetto alla versione 2.2, la versione 2.3 beta è stata rilasciata in maniera informale specificatamente per le attività di questo lavoro, poiché per le caratteristiche specifiche dei dati in esame è stata arricchita con la funzione di distanza, denominata window equality, ideata in particolare per il confronto tra numeri interi e definita secondo la regola: se (|x-y| <= w) i due numeri sono considerati uguali altrimenti i due numeri sono considerati diversi, dove w è la dimensione della finestra definita dall'utente e x e y sono i due numeri da confrontare.

3.2 I requisiti della procedura di abbinamento

La realizzazione del sistema integrato sugli esiti del concepimento ha l'obiettivo principale di permettere di misurare i più importanti aspetti socio-sanitari degli esiti dei

concepimenti attraverso un insieme di indicatori. Ciò impone che la base di dati di riferimento sia la più ampia possibile e che non presenti "distorsioni" rispetto alle fonti originarie che la compongono. In termini di linkage, questi requisiti si traducono innanzitutto in un "alto" livello del match rate (o tasso di abbinamento), definito, nel caso dell'abbinamento tra due fonti, come il rapporto tra numero dei record abbinati e il numero dei record presenti nella più piccola delle fonti considerate[3]. Per questo motivo, particolare attenzione in questo studio di fattibilità deve essere prestato alle sperimentazioni relative all'abbinamento delle fonti di dimensioni più contenute, IVG e AS e al sottoinsieme dei nati morti, poiché, in questi casi, un basso valore del match rate potrebbe comportare a livello integrato una eccessiva riduzione delle osservazioni relative al fenomeno in esame.

D'altro lato, a patto di raggiungere un "buon" livello del match rate, è possibile valutare meno severamente l'eventuale perdita di ulteriori veri link, a meno che questi non siano caratterizzati in modo tale da rendere distorte le successive analisi sui dati integrati. Ciò significa poter tenere in diversa considerazione gli errori di falso abbinamento e quelli di mancato abbinamento[4]: i primi saranno considerati più gravi, in quanto introducono eventi non esistenti nei dati, mentre i mancati abbinamenti possono essere considerati meno gravi, anche se comunque da evitare, sotto l'ipotesi, da verificare, che non comportino distorsione nei risultati, ossia che i mancati abbinamenti non si concentrino in maniera evidente in particolari categorie o sotto-popolazioni. Le procedure di abbinamento, sperimentate e descritte nel seguito, hanno quindi posto vincoli più stringenti sul tasso di falso abbinamento e meno restrittivi sul tasso di mancato abbinamento. Per analizzare l'eventuale introduzione di un effetto distorsivo dovuto al mancato abbinamento di alcuni record, gli esiti dell'abbinamento probabilistico sono comparati a livello di singola coppia individuata con le risultanze dell'abbinamento deterministico proposto nel successivo paragrafo 4.2; inoltre verranno comunque confrontate le distribuzioni di frequenza dei dati abbinati e di quelli di partenza rispetto alle principali variabili di interesse per la costruzione di indicatori socio-sanitari.

[3] Il tasso di abbinamento o match rate, che è una delle misure di qualità di un processo di linkage, è propriamente definito come il rapporto tra il numero totale di record abbinati e il numero vero (ignoto) di abbinamenti. Tale indicatore viene usualmente calcolato sostituendo al totale ignoto di veri abbinamenti la sua stima fornita dal modello o la dimensione del più piccolo tra i file da abbinare, che costituisce comunque il massimo degli abbinamenti possibili, sotto l'ipotesi che tutti i record debbano essere abbinati. In questo modo si fornisce una misura del minimo del tasso di abbinamento, ed è questa la valutazione cautelativa della procedura di linkage che è stata adottata nel seguito del presente lavoro.

[4] Gli errori di classificazione nel modello di linkage proposto da Fellegi e Sunter sono di due tipi: gli abbinamenti errati (o falsi abbinamenti – false matches nella più diffusa terminologia inglese), quando vengono abbinate unità che corrispondono a entità differenti e gli abbinamenti mancati (false non-matches), quando record corrispondenti ad una stessa entità non vengono abbinati. In generale, gli abbinamenti errati si suddividono, a loro volta, in: accoppiamenti tra due unità che non dovrebbero essere abbinate tra loro ma con altri record e accoppiamenti tra unità che non dovrebbero essere abbinate affatto. Gli errori di abbinamento, sia abbinamenti errati che abbinamenti mancati, giocano un ruolo fondamentale per la valutazione della bontà dei risultati delle procedure di linkage e devono essere tenuti nella massima considerazione nelle successive analisi sui dati linkati, in quanto possono influire significativamente su di esse. Misure sintetiche della qualità del linkage, basate su tali errori, sono i tassi di mancato abbinamento e di falso abbinamento, definiti, il primo, come il rapporto tra numero stimato di mancati abbinamenti e il totale stimato di veri abbinamenti e, il secondo come il rapporto tra il numero stimato di falsi abbinamenti e il totale di abbinamenti individuati.

4. Primi risultati: una sperimentazione sull'Emilia Romagna

La dimensione del fenomeno oggetto di studio è rilevante in termini di numero di individui coinvolti. Come riportato nella tavola 2.2 del paragrafo 2, alcune delle fonti principali coinvolgono un numero di record molto elevato (circa 500 mila unità, annualmente a livello Italia). Le analisi e le sperimentazioni per lo studio della fattibilità sono state avviate sulla regione Emilia Romagna, come riferimento territoriale per gli eventi da considerare. La selezione di questa regione è legata principalmente a due fattori:

- la numerosità degli eventi presi in esame è cospicua, in Emilia Romagna sono concentrati il 10% circa degli eventi nazionali relativi alle nascite;
- la buona qualità delle fonti disponibili, per quanto riguarda i dati sanitari che arrivano all'Istat dopo essere stati raccolti a livello regionale.

La scelta di restringere, in fase sperimentale, il campo di osservazione ad un numero ridotto di casi tende a far crescere il potere identificativo di variabili come le date di nascita per quanto riguarda gli individui e le date relative agli eventi, che per alcune fonti sono le informazioni con più alto potere identificativo disponibili.

Per quanto riguarda il riferimento temporale, la sperimentazione del sistema integrato è stata avviata selezionando come anno di riferimento degli eventi il 2007. La scelta dell'anno di riferimento è stata dettata dall'opportunità di scegliere un periodo in cui la gestione e la qualità delle fonti coinvolte abbia raggiunto una certa stabilità; inoltre, si prevede la possibilità, in una fase successiva, di incrementare il sistema con gli eventi relativi agli anni seguenti, così da gettare le basi per la creazione di un sistema che permetta anche di seguire nel tempo i comportamenti riproduttivi e i relativi esiti in un'ottica longitudinale.

Si evidenzia che la scelta dell'anno è stata anche dettata dal fatto che le fonti hanno una tempistica di rilascio dei dati diversa per cui il 2007 era l'anno più recente comune a tutte le fonti disponibili nel momento in cui è iniziata la progettazione delle attività descritte in questo documento (primo semestre 2012). Nella successiva tavola 4.1 si riportano le numerosità relative alla regione Emilia Romagna per gli eventi occorsi nell'anno 2007 e la percentuale delle numerosità di tale regione rispetto al totale Italia.

Tavola 4.1. Numerosità degli eventi in Emilia Romagna nel 2007

	Nome della fonte	Numero eventi in Emilia Romagna	Percentuale sugli eventi in Italia
1a	CEDAP parti	39.792	7.65
1b	CEDAP nati vivi	40.022	7.63
1c	CEDAP nati morti	114	5.19
3	P4	39.744	7.15
4	IVG	11.267	9.01
5	AS	5.872	7.61
6a	SDO parti	40.242	7.22
6b	SDO nati vivi	41.637	7.44
6c	SDO nati morti	154	9.16
6d	SDO ivg	11.661	8.85
6e	SDO as	6.448	7.25

E' importante notare che, nell'anno 2007 e in una regione come l'Emilia Romagna che rappresenta un'eccellenza per la gestione e la qualità delle fonti oggetto di analisi, alcune di

queste riportano dati in maniera ancora poco precisa ed affidabile sia in termini di numerosità che di informazione registrata. E' il caso ad esempio della fonte CEDAP, soprattutto per la rilevazione della nati-mortalità. Queste criticità emergono chiaramente anche dagli indicatori di qualità delle procedure di linkage, come si può notare nel successivo paragrafo 4.1. Per questo motivo è importante che le analisi successive sui dati abbinati tengano conto e includano le misure della qualità dell'integrazione effettuata. Infatti, a queste criticità si può ovviare in diverse fasi del processo di produzione dell'informazione statistica. In fase di stima, si devono adottare stimatori che tengono conto esplicitamente della scarsa qualità delle fonti coinvolte. Questo implica che la qualità delle fonti sia valutata anche in termini quantitativi. In questo senso, il processo di integrazione di tipo probabilistico è il più indicato perché fornisce delle misure quantitative (la probabilità di corretto abbinamento e quella di mancato abbinamento) che possono essere direttamente inserite nella fase di stima degli indicatori forniti dal sistema integrato. A livello organizzativo, d'altro canto, è necessario che queste evidenze sulle cadute di qualità per particolari fonti o sottopopolazioni, siano sfruttate per identificare i campi di azione e gli attori su cui intervenire prioritariamente per aumentare la qualità dell'input.

4.1 I risultati del linkage probabilistico

Le analisi per la valutazione della fattibilità comprendono anche numerose applicazioni di integrazione delle fonti. I dettagli delle procedure di integrazione sperimentate sono descritti in un rapporto tecnico a diffusione interna dal titolo "Progettazione e realizzazione del prototipo del sistema integrato degli esiti del concepimento" (Istat, 2012). Di fatto si è proceduto abbinando singole coppie di fonti che riguardano lo stesso fenomeno, secondo lo schema 3.1. Le variabili disponibili per l'abbinamento possono essere raggruppate in 3 gruppi:

- le variabili riguardanti la donna protagonista dell'evento: l'età (in poche fonti la data di nascita completa), la provincia e/o il comune/stato estero di residenza, la provincia e/o il comune/stato estero di nascita;
- le variabili riguardanti la localizzazione geografica dell'evento: il codice dell'ospedale, il comune e la provincia;
- le variabili riguardanti la data dell'evento: giorno e mese.

Le strategie di abbinamento più efficaci sono risultate quelle in cui nella selezione delle variabili di linkage e delle variabili utilizzate per la riduzione dello spazio di ricerca delle coppie candidate è considerata almeno una variabile da ciascun gruppo. In realtà, i diversi processi di integrazione applicati alle varie coppie di dati hanno richiesto metodi di riduzione delle spazio di ricerca differenti (di bloccaggio o di ordinamento) basati su variabili diverse, in funzione soprattutto della dimensione dei dati da trattare. Conseguentemente, per le varie procedure di abbinamento sono state selezionate variabili di linkage diverse, evitando di scegliere all'interno dello stesso gruppo quelle con potere identificativo aggiuntivo troppo scarso, perché fortemente correlate (come ad esempio il codice dell'ospedale e il comune dell'evento). Per alcune variabili, quali l'età della donna e il giorno dell'evento, sono state prese in considerazione funzioni di confronto che, accettando concordanze meno stringenti dell'esatta uguaglianza, consentono di definire degli intervalli di somiglianza così da aumentare la probabilità di aggancio tra record.

La tavola 4.2 riporta i risultati delle sperimentazioni che hanno fornito gli esiti migliori, in base ai criteri definiti nel paragrafo 3.2, in termini di match rate e probabilità di corretto

abbinamento.

Tavola 4.2. Quadro riassuntivo degli abbinamenti ottenuti con record linkage probabilistico

Abbinamenti	Dimensione file di partenza	Numero abbinamenti da RL probabilistico	Match rate
AS – SDO as	5.872 – 6.448	3.864	0.66
IVG - SDO ivg	11.267 – 11.661	10.938	0.97
P4 - SDO parti	39.744 – 40.242	27.711	0.70
CEDAP natimorti – SDO natimorti	114 - 154	80	0.70
CEDAP parti - SDO parti	39.792 – 40.242	33.622	0.84
P4 - CEDAP nati	39.744 – 40.370	37.469	0.94

I risultati degli abbinamenti tra le fonti considerate sono sicuramente incoraggianti. Infatti, per tutte le procedure di abbinamento considerate, le probabilità medie stimate che ciascuna coppia individuata sia un vero match è generalmente superiore al 95% (Tavola 4.3).

Tavola 4.3. La qualità degli abbinamenti in termini di falso e mancato abbinamento

Abbinamenti	Numero record abbinati	Probabilità media di corretto abbinamento	Numero stimato di falsi abbinamenti	Numero massimo stimato di mancati abbinamenti
AS – SDO as	3.864	0.97	116	1314
IVG - SDO ivg	10.938	0.93	766	328
P4 - SDO parti	27.711	0.99	277	8.313
CEDAP natimorti – SDO natimorti	80	0.99	1	15
CEDAP parti - SDO parti	33.622	0.99	336	5.380
P4 - CEDAP nati	37.469	0.99	375	2.248

La eventuale presenza di falsi abbinamenti nel sistema integrato sugli esiti dei concepimenti costituisce una criticità per lo studio e l'analisi dei fenomeni che da questo si vogliono interpretare soprattutto se i falsi abbinamenti introducono distorsione nei dati. Di fatto, non ci sono ragionevoli motivi per ritenere che il processo di abbinamento sia distorto rispetto a qualcuna delle variabili di riferimento per le stime degli indicatori di interesse. In ogni caso il tasso di falso abbinamento e la probabilità di corretto abbinamento sono misure di cui bisognerà tenere conto esplicitamente nelle analisi successive sui dati abbinati, per una corretta modellizzazione del processo di generazione dei dati stessi.

Il numero di mancati abbinamenti è anch'esso contenuto, dati gli alti valori del match rate.

Riguardo alla stima del numero di mancati abbinamenti, si deve considerare che non tutti i record dei file di partenza che non si abbinano sono mancati abbinamenti, dato che nella maggior parte dei casi non si verifica una perfetta sovrapposizione delle popolazioni di riferimento. Ad esempio, i record riportati nelle due fonti P4 - SDO parti non sono riferiti esattamente allo stesso universo di riferimento, per varie ragioni, tra cui:

- la fonte P4 riporta tanti record quanti sono i nati vivi, indipendentemente dal fatto che provengano da parto singolo o multiplo, quindi, nel caso di parti singoli, c'è coincidenza con le informazioni sul parto riportate nei record della fonte SDO parti mentre i record del P4 generati da parti multipli "eccedono" rispetto ai record SDO parti per tutti i gemelli dal secondo in poi;
- nella fonte SDO ci sono record di parti di bambini nati morti che la fonte P4 non

registra;
- nella fonte SDO ci sono record relativi a parti avvenuti in Emilia Romagna di bimbi iscritti in anagrafe fuori dalla regione o anche non iscritti, come nel caso delle donne straniere;

Di conseguenza, è attesa una buona sovrapposizione dei file, d'altra parte è ammesso che un numero di mancati abbinamenti non costituisca errore del processo ma sia dovuto al diverso universo di riferimento delle fonti trattate. Per quanto riguarda il fenomeno delle nascite, la fonte SDO nati potrebbe avere una maggiore aderenza alla fonte P4 rispetto a SDO parti; purtroppo, dalla ricognizione delle fonti effettuata, è emerso che allo stato attuale le variabili identificative utili ai fini dell'abbinamento riportate in SDO nati sono troppo esigue e nella maggior parte dei casi non valorizzate.

Infine, occorre ricordare che sia nel record linkage tra le fonti considerate, che in quello effettuato sui residui si è scelto di operare un abbinamento di tipo 1:1: in caso di parto plurimo che ha dato luogo a due o più nati vivi ci si aspetta che una stessa scheda di dimissione ospedaliera per parto si agganci a più schede di Iscrizione in Anagrafe per nascita. L'abbinamento di tipo 1:1 consente, per questa specifica coppia di fonti, di creare un sistema integrato che ha come unità di riferimento la singola donna e non il risultato dei suoi concepimenti. Il fenomeno "esito del concepimento" per ciascuna donna può essere ricostruito evitando la riduzione degli abbinamenti da 1:n a 1:1, integrando con la fonte SDO nati morti e prendendo in considerazione l'integrazione con le fonti SDO nati (quando sarà maggiormente popolata di informazioni essenziali per il linkage) e CEDAP.

Infine, è ragionevole ritenere che ulteriori abbinamenti saranno recuperati nel momento in cui verranno prese in esame tutte le regioni italiane, dato che mancate coincidenze nella variabile che indica la regione di evento impediscono di confrontare record e riconoscere ulteriori coppie. Nel momento in cui si provvederà alla messa a punto del sistema per l'intera Italia, si può immaginare di confrontare a livello nazionale i record che non si abbinano all'interno della singola regione, così da recuperare queste ulteriori coppie.

4.2 L'applicazione di linkage deterministico

Sui dati in esame è stata applicata anche una procedura di abbinamento deterministico, che è stata sviluppata nel contesto delle attività di abbinamento deterministico di dati demo-sociali e nel tempo è stata arricchita e perfezionata, dimostrando di perseguire ottimi risultati. Il metodo viene brevemente descritto in questo paragrafo e può definirsi un modello a "chiavi integrate" corredato di un indicatore di "qualità degli abbinamenti" ottenuti.

Il metodo consente l'utilizzo contemporaneo, in maniera integrata all'interno della stessa procedura, di una serie di chiavi di linkage, che si ottengono dal "concatenamento testuale" di un gruppo di variabili di linkage, opportunamente scelte, nell'ottica del problema trattato.

È opportuno precisare che una delle peculiarità del metodo in questione è quella di prevedere due diversi tipi di chiavi di linkage:
- la chiave di linkage completa, che è la chiave più restrittiva che si possa costruire, ossia quella che si ottiene concatenando tutte le variabili di linkage individuate;
- le "altre" chiavi di linkage o "derivate", che sono chiavi alternative alla chiave completa e in un certo modo da essa "derivate" e per questo "meno restrittive" della stessa. Esse si ottengono dalla chiave completa togliendo una variabile di linkage alla volta.

Per il calcolo dell'indicatore di qualità degli abbinamenti, occorre definire le variabili "di controllo", anch'esse comuni alle due fonti, meno discriminanti delle variabili di linkage, ma lo stesso utili per la valutazione della qualità dei risultati ottenuti. Infatti, il suddetto indicatore di qualità si ottiene conteggiando il numero di concordanze totali che si verificano tra i valori assunti dalle variabili comuni alle due fonti (sia di linkage che di controllo) sull'abbinamento individuato. Le concordanze vengono conteggiate solo per i valori diversi dai valori mancanti su entrambe le fonti. Questo indicatore costituisce uno strumento prezioso in due momenti di applicazione del metodo:

- in una fase intermedia del processo, quando una unità di una delle due fonti si lega a più unità dell'altra; in questo caso il metodo sceglierà l'abbinamento in cui l'indicatore di qualità presenti il valore massimo (massima concordanza delle informazioni disponibili);
- nella fase finale di valutazione, quando occorre decidere se gli abbinamenti dichiarati dal metodo debbano essere giudicati veri, falsi o dubbi.

Per le sperimentazioni relative al sistema integrato degli esiti dei concepimenti, le diverse chiavi su cui si basa il metodo sono state costruite a partire dalle stesse variabili utilizzate per l'applicazione del linkage probabilistico.

4.3 Analisi e confronto dei risultati

In questo paragrafo si analizzano in maniera comparativa i risultati delle due tecniche di linkage sperimentate: quella probabilistica e quella deterministica. Dal confronto emergono punti di forza e di debolezza di entrambi gli approcci, ma soprattutto tale analisi mette in luce il lavoro ancora da fare per il miglioramento della qualità delle fonti di partenza, soprattutto in un'ottica di utilizzo integrato dei dati. In ogni caso, il confronto dei risultati delle due procedure di abbinamento sperimentate conferma la validità della strategia delineata per la costruzione del sistema integrato sugli esiti dei concepimenti, poiché in generale, salvo poche eccezioni discusse nel seguito, la percentuale degli abbinamenti individuati da entrambe le procedure è superiore al 70%, con punte di 97%, a conferma della robustezza dei processi di linkage individuati.

Nella successiva tavola 4.4 si riassumono i risultati, per ogni coppia di fonti considerate: riportando ancora la dimensione dei file di partenza per ogni abbinamento, vengono presentati il totale di abbinamenti individuati dall'approccio deterministico, il numero di abbinamenti trovati dai modelli probabilistici e il numero di coppie comuni alle due procedure; l'ultima colonna della tavola riporta in termini percentuali il rapporto tra il numero di abbinamenti comuni ai due metodi e la dimensione del più piccolo insieme di abbinamenti individuati.

Tavola 4.4. Quadro riassuntivo degli abbinamenti secondo i metodi di record linkage sperimentati

Fonti abbinate	Dimensione file di partenza	RL deterministico	RL probabilistico	Link comuni	% Link comuni
AS - SDO as	5.872–6.448	4.746	3.864	720	18,6
IVG - SDO ivg	11.267–11.661	9.650	10.938	7.497	77,7
P4 - SDO parti	39.744–40.242	33.752	27.711	19.266	69,5
CEDAP natimorti - SDO natimorti	114–154	102	80	79	98.7
CEDAP parti - SDO parti	40.242–39.792	36.481	33.622	27.844	82.8
P4 - CEDAP nati	39.744–40.370	38.081	37.469	36.572	97.6

Da un primo esame della tavola risulterebbe che in generale il metodo deterministico tenda ad individuare un numero maggiore di abbinamenti rispetto ai modelli probabilistici sperimentati. In realtà, questo risultato necessita un'analisi più approfondita poiché dipende da diversi fattori. In primo luogo incide il metodo deterministico applicato, infatti questo non si limita a dichiarare abbinamenti tutte quelle coppie che coincidono sull'insieme completo di variabili di confronto selezionate ma procede ad abbinare anche le coppie che coincidono su un insieme più limitato di variabili, rimuovendo il vincolo di uguaglianza per una o più delle variabili di confronto selezionate, purché il numero di concordanze totali sia superiore ad una soglia prefissata. In secondo luogo, il minor numero di abbinamenti individuati dai modelli probabilistici è condizionato dal fatto che questi sono stati applicati in via sperimentale senza iterazioni successive sui record rimasti non abbinati al primo passo, al contrario della pratica consueta nell'uso di metodi probabilistici. Ciò è confermato in particolare dal risultato relativo all'abbinamento tra le fonti IVG e SDO-ivg, per le quali le procedure probabilistiche sono state impiegate come di prassi in più passi successivi, con relativo abbinamento dei record rimasti non abbinati nei passi precedenti. In questo caso in particolare sono state sperimentate due iterazioni e ciò ha consentito al metodo probabilistico di individuare un numero di abbinamenti maggiore rispetto al metodo deterministico. Tale constatazione suggerisce di estendere la pratica di applicazioni ripetute e iterate dei modelli di linkage probabilistico anche alle altre coppie di dati, poiché è ancora possibile processare i dati non abbinati per recuperare corretti abbinamenti. In ogni caso, restano valide le considerazioni conclusive del paragrafo 4.1, sulla possibilità di individuare nuovi abbinamenti estendendo il campo di osservazione dell'attuale sperimentazione di linkage probabilistico all'intero territorio nazionale.

Infine, al di là della valutazione comparativa sulle performance dei due diversi metodi sperimentati, il confronto mette in luce un aspetto fondamentale nell'ottica di messa a regime del sistema integrato sugli esiti dei concepimenti. A tal proposito, la tavola 4.4 evidenzia chiaramente come la comparabilità dei risultati sia fortemente legata alle fonti considerate nel linkage: prendendo in considerazione le fonti per cui la sovrapposizione degli eventi e degli universi di riferimento è più alta (P4 - CEDAP nati, CEDAP parti - SDO parti, CEDAP natimorti - SDO natimorti) i risultati delle due procedure sono simili, con percentuali di sovrapposizione fino al 98%. La comparabilità dei risultati scende fino al 70% circa per le fonti relative ad universi di riferimento solo parzialmente omogenei (si vedano le considerazioni conclusive del paragrafo 4.1 relative al linkage tra P4 - SDO parti). Un discorso a parte merita il confronto dei risultati degli abbinamenti AS - SDO e IVG - SDO: data la natura delle fonti e la qualità delle informazioni in esse riportate, ci si

aspettava una percentuale di sovrapposizione dei risultati analoga tra i due abbinamenti. La notevole differenza nelle percentuali di abbinamenti comuni individuati dalla procedura deterministica e da quella probabilistica e il numero contenuto di abbinamenti per le fonti AS - SDO sono da imputare alla non disponibilità per l'anno di riferimento della variabile "giorno di intervento". Ulteriori approfondimenti sono disponibili su altri lavori effettuati in Istat (Cotroneo, Tuoto, Loghi, 2012). In ogni caso la variabile "giorno di intervento" si è rivelata di fondamentale importanza per la costruzione dei modelli di linkage tanto che è stata tempestivamente attivata la procedura per il suo inserimento nella rilevazione Istat degli aborti spontanei.

5. Conclusioni e prospettive future

I risultati del complesso processo di integrazione per la valutazione della fattibilità della costruzione di un sistema integrato sugli esiti dei concepimenti sono molto incoraggianti. La sperimentazione sulla regione Emilia Romagna e per l'anno di riferimento considerato suggeriscono di continuare sul percorso individuato, dato l'elevato numero di record abbinati e l'alta qualità degli abbinamenti. Il sistema permette di produrre indicatori su temi quali la medicalizzazione del percorso gravidanza, nascita e allattamento in relazione al contesto socio-sanitario, alle caratteristiche socio-demografiche delle donne, al contesto socio-economico familiare; il percorso gravidanza, nascita e parto (fisiologico/naturale vs fortemente medicalizzato).

In questo paragrafo sono riportate alcune conclusioni sul lavoro svolto e si evidenziano delle proposte per possibili sviluppi, relativamente agli aspetti definitori, ai dati considerati e alle procedure di integrazione.

Per quanto riguarda gli aspetti definitori, la strategia sperimentata per la messa a punto del sistema integrato sugli esiti dei concepimenti fa riferimento al fenomeno "esito del concepimento", rispetto alle modalità: nato vivo, nato morto, aborto spontaneo, interruzione volontaria, come precisato nel paragrafo 3. Le potenzialità offerte dal sistema fin qui illustrate fanno riferimento quindi all'adozione di un'ottica trasversale. Come possibile sviluppo si vuole indagare nei prossimi passi la possibilità di ricondurre ciascun fenomeno a quelli ad esso correlati (ad esempio, parti multipli e anche parti, interruzioni di gravidanza e aborti relativi ad una stessa donna) ossia spostare l'attenzione dal fenomeno "esito" del singolo concepimento alla "donna", in quanto soggetto del concepimento. Il prototipo studiato e sperimentato in questo lavoro è comunque un passaggio obbligato dell'estensione proposta, come ampiamente illustrato nel paragrafo 2, per "seguire" nel tempo l'evoluzione delle storie riproduttive delle donne, anche in prospettiva longitudinale. Per quanto riguarda i dati da utilizzare per la costruzione del sistema integrato sugli esiti dei concepimenti, l'attuale sperimentazione ha condotto un'analisi puntuale della fattibilità dell'integrazione a livello micro per le varie fonti considerate, da cui è emerso un insieme minimo di variabili fondamentali a garantire abbinamenti di elevata qualità e la conseguente necessità di richiedere tali variabili per le fonti che ancora non ne sono provviste. Tale richiesta dovrebbe essere evasa facilmente, in quanto le variabili fondamentali sono già rilevate dall'ente che fornisce i dati e non sono soggette a vincoli legati alla privacy. E' il caso ad esempio della variabile "giorno di intervento" per la fonte AS, evidenziato nel paragrafo precedente.

Un ulteriore utilizzo dei dati disponibili, non ancora indagato in modo approfondito, prevede infine la possibilità di collegare tra di loro gli abbinamenti individuati tra le singole coppie di fonti attraverso identificativi esatti condivisi tra alcune fonti, laddove la corrispondenza sia del tipo uno a uno (si pensi ai nati vivi registrati nei CEDAP, negli iscritti in anagrafe per nascita e nelle SDO). Questo passaggio richiede la definizione di controlli di coerenza tra le diverse fonti per correggere situazioni di incompatibilità dovute ad errori o scarsa qualità.

Si rimanda ad approfondimenti successivi anche il trattamento dei record non abbinati delle varie fonti considerate dalle procedure di linkage. Queste valutazioni sono legate soprattutto alla definizione degli universi di riferimento delle fonti considerate e agli obiettivi conoscitivi del sistema integrato, oltre che alle procedure di integrazione messe in atto.

Per quanto riguarda gli aspetti strettamente connessi alle procedure di integrazione, le sperimentazioni effettuate hanno fornito risultati molto confortanti sia in termini di numero di record abbinati che di qualità degli abbinamenti, dichiarando quindi fattibile la realizzazione di un sistema integrato sugli esisti dei concepimenti. I lavori dovrebbero proseguire estendendo l'abbinamento all'intero territorio nazionale. In quest'ottica, sulla base delle sperimentazioni condotte, un ulteriore filone di attività consiste nella possibilità di introdurre nelle procedure di abbinamento dei vincoli che garantiscano un'elevata qualità dei risultati soprattutto rispetto a particolari fenomeni o sottopopolazioni su cui il sistema integrato voglia fornire dei focus (ad esempio, il fenomeno della natimortalità o la sotto-popolazione delle donne straniere).

Un successivo sviluppo dell'attuale sistema prevede, come anticipato nel paragrafo 2, l'integrazione di ulteriori fonti di dati, ad esempio quelle relative all'indagine campionaria sulle nascite e all'indagine Istat sulle cause di morte, con riferimento ai morti nel primo anno di vita e negli anni successivi. L'integrazione di queste ulteriori fonti di dati permetterebbe di arricchire gli obiettivi conoscitivi coperti dal sistema integrato e recuperare il grave vuoto informativo sulla salute perinatale. Al momento, infatti, non è possibile ad esempio ricondurre la mortalità nel primo anno di vita alle informazioni sulla gravidanza e il parto. Includere nel sistema anche l'indagine Istat sulle cause di morte permetterebbe di calcolare i principali indicatori sulla salute perinatale (mortalità neonatale precoce e tardiva), come raccomandato dal Progetto Europeristat, e consentire il confronto con gli altri Paesi europei oltre che fornire valutazioni anche a livello regionale e sub-regionale. Inoltre l'integrazione nel sistema della rilevazione delle cause di morte dopo il primo anno di vita consentirebbe di mettere in relazione la mortalità materna per cause connesse alla gravidanza o al parto con le informazioni sulla gravidanza e il parto contenute nei CEDAP. Anche questo tipo di indicatori è fortemente raccomandato a livello internazionale (European Perinatal Health Report, 2008). E' possibile includere nel sistema integrato anche le indagini campionarie sulle nascite e le madri, dove la popolazione di riferimento sono le nascite della popolazione residente riferite agli anni di calendario 2003 e 2009-2010. Tale operazione consentirebbe di:

a) verificare la qualità e la copertura delle indagini rispetto ad alcune variabili chiave come il livello di istruzione della popolazione;
b) valutare la bontà dell'ordine di nascita stimato utilizzando la variabile "numero di componenti minorenni" registrata nelle schede anagrafiche di iscrizione per nascita (previo opportuno trattamento statistico di validazione), con l'informazione diretta

fornita dalle intervistate e la parità desumibile dai CEDAP in modo da migliorare ulteriormente le procedure di stima;

c) studiare la qualità delle informazioni sanitarie sul parto rilevate attraverso le interviste Cati rispetto a quelle rilevate per le stesse donne dal personale sanitario che compila il CEDAP (per verificare, ad esempio, se le donne tendono a sovrastimare il peso dei figli alla nascita o a non riferire eventi sensibili come esiti negativi di precedenti concepimenti o nascite non viventi nel caso di parto gemellare con nati vivi) .

In generale sarebbe possibile valutare il grado di affidabilità delle informazioni raccolte con le indagini campionarie anche in vista di successive occasioni di indagini. Ciò consentirebbe di disporre di una base dati di amplissimo potere informativo per l'analisi delle nascite e dei parti in relazione alle determinanti socio-sanitarie e a quelle demo-sociali.

Infine, si vuole ricordare che le successive analisi statistiche basate sui dati riportati nel sistema integrato dovranno necessariamente tenere nella debita considerazione il processo di integrazione che ha generato tali dati. Infatti, sebbene i metodi di integrazione di microdati siano uno strumento estremamente potente e ampiamente utilizzato, il linkage, come in genere molte fasi del processo di produzione del dato statistico, produce risultati non sempre privi di errori. Come evidenziato nel paragrafo 3, le applicazioni di record linkage devono quindi essere corredate da opportune informazioni sulla qualità del linkage; tali indicatori di qualità, a partire da quelli riportati nel paragrafo 4, devono essere inseriti con opportune metodologie nelle successive fasi di stima basata su dati linkati, così da assicurare la qualità delle analisi finali. Infatti, trattare i dati abbinati ignorando il processo di integrazione e gli eventuali errori da questo generati, può portare, in generale, a stime distorte (Neter *et al,* 1965, Winkler e Scheuren, 1993, 1997). Bisognerà quindi studiare e implementare opportuni metodi di stima di indicatori e di relazioni statistiche tra variabili che tengano nella dovuta considerazione il processo di integrazione alla base della costruzione del sistema integrato.

Riferimenti bibliografici

Attili M., Valentino L. (2010). "Le informazioni sulle nascite e i parti. Esperienze di integrazione tra dati di fonte anagrafica e sanitaria", presentato al seminario "REcord Linkage At IStat: Applicazioni con RELAIS 2.0", 22 aprile 2010, Istat, Roma.

Cibella N., Fortini M., Spina R., Scannapieco M., Tosco L., Tuoto T. (2007). "Relais: An open source toolkit for record linkage", Rivista di Statistica Ufficiale n. 2-3/2007, pp.55-68

Cibella N., Tuoto T. (2012) "Statistical perspectives on blocking methods when linking large data-sets", in A. Di Ciaccio et al. (eds.), Advanced Statistical Methods for the Analysis of Large Data-Sets, Springer-Verlag Berlin Heidelberg.

Cotroneo R., Tuoto T., Loghi M. (2012). "L'importanza della scelta delle variabili nel record linkage: il caso delle Interruzioni volontarie di gravidanza e degli Aborti spontanei" presentato alle Giornate di Studio sulla Popolazione (GSP) 2013, Brixen, Febbraio 6 –8, 2013, disponibile all'indirizzo web http://www.sis-aisp.it/ocs-2.3.4/index.php/gsp2013/gsp2013/paper/view/233

European Perinatal Health Report. (2008). Disponibile all'indirizzo web www.europeristat.com

Fellegi, I.P., Sunter, A.B. (1969). "A Theory for Record Linkage", Journal of the American Statistical Association, 64, pp. 1183-1210.

Jaro, M. A. (1989). "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida", Journal of the American Statistical Society, 84 (406), pp.414–20.

Istat (2012) "Progettazione e realizzazione del prototipo del sistema integrato degli esiti del concepimento" relazione finale del Gruppo di lavoro "Metodi e tecniche di record linkage tra fonti demografiche e sociali"

Neter, J., Maynes, E.S., Ramanathan, R. (1965). The effect of mismatching on the measurement of response error. Journal of the American Statistical Association, 60, 1005-1027.

RELAIS (2011). User's guide version 2.2. Disponibile all'indirizzo web http://www.istat.it/it/strumenti/metodi-e-software/software/relais e http://joinup.ec.europa.eu/software/relais/release/22

Scheuren, F.,Winkler, W.E. (1993). Regression analysis of data files that are computer matched – Part I. Survey Methodology, 19, pp. 39-58.

Scheuren F.,Winkler W.E. (1997). Regression analysis of data files that are computer matched- part II, Survey Methodology, 23, pp. 157-165.

Gli stranieri residenti per genere e cittadinanza: la stima per comune negli anni successivi al censimento[1]

Mauro Albani[2], Maura Simone[3]

Sommario

I dati sulla popolazione straniera tratti dalle anagrafi comunali, validati e diffusi annualmente dall'Istat, forniscono l'ammontare dei cittadini comunitari ed extra comunitari regolarmente residenti in Italia, distribuito per genere e Paese di cittadinanza. Il processo di validazione dei suddetti dati, con l'allineamento alla popolazione rilevata con il censimento e aggiornata sulla base dei bilanci demografici annuali, rappresenta un esempio di utilizzo a fini statistici dell'informazione contenuta nei registri amministrativi. Obiettivo del presente lavoro è mostrare come, stanti le informazioni derivanti dalle fonti disponibili, negli anni immediatamente successivi al censimento del 2011 il metodo dei saldi indiretti netti abbia consentito la stima per comune della popolazione straniera residente distribuita per genere e cittadinanza, garantendo la transizione graduale dalla distribuzione rilevata al censimento verso le distribuzioni annuali tratte dai registi anagrafici, a valle delle operazioni di revisione anagrafica.

Parole chiave: stranieri, cittadinanze, popolazione residente.

Abstract

The data on foreign population currently captured from municipal population registers, are yearly validated and disseminated by the Italian National Institute of Statistics (Istat). They concern the number of EU and non-EU citizens regularly resident in Italy, distributed by gender and citizenship. The validation process - including its ex-ante consistency making step with the official census individual records - is a typical example of using administrative source information for statistical purposes. The aim of this work is to show that, given the information deriving from sources currently available, for the years immediately following the 2011 census the method of indirect net balances allowed to estimate foreign population distributed by sex and country of citizenship, ensuring the smoothest transition from the distribution detected via-census towards the post-censual yearly distributions of municipal population registers after their post-censual revision operations.

Keywords: foreigners, citizenship, resident population

---

[1] Sebbene il lavoro sia frutto dell'opera di tutti gli autori, la cura dell'articolo e tutti i paragrafi sono da attribuire a Mauro Albani, eccetto i paragrafi 2.3 e 2.6, che sono da attribuire a Maura Simone. Gli articoli pubblicati impegnano esclusivamente gli autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

[2] Istat, e-mail: albani@istat.it.

[3] Istat, e-mail: simone@istat.it.

1. Introduzione

Obiettivo del presente lavoro è descrivere il metodo utilizzato per stimare le distribuzioni per sesso e Paese di cittadinanza dei cittadini stranieri residenti in ciascun comune italiano negli anni immediatamente seguenti il XV Censimento generale della popolazione e delle abitazioni (2011) e verificarne la validità attraverso il confronto con gli altri metodi applicabili date le fonti a disposizione[4]. In particolare il metodo adottato ha consentito di armonizzare le suddette distribuzioni, ricavabili annualmente dai registri anagrafici, con il calcolo della popolazione derivante dai risultati del censimento, aggiornati anno per anno sulla base dei bilanci demografici comunali.

Occorre chiarire sin da subito in proposito che la procedura adottata risulta strettamente legata alle modalità innovative secondo cui si è svolto il censimento del 2011 (Crescenzi et al., 2009) e allo strumento altrettanto nuovo con cui è stata eseguita, negli anni successivi, la revisione anagrafica (Simone et al., 2012).

L'analisi per genere e cittadinanza della presenza straniera regolare in Italia contribuisce in modo decisivo allo studio del fenomeno migratorio nel Paese. Per la popolazione straniera infatti le caratteristiche demografiche e sociali (ad esempio la composizione per genere, la composizione dei nuclei familiari, ecc.), l'attività lavorativa, la distribuzione gli stessi modelli insediativi sul territorio risultano spesso strettamente collegati con la cittadinanza di origine. Le numerose comunità straniere residenti nel territorio italiano presentano caratteristiche tra loro anche molto diverse e non si può prescindere dall'analisi per cittadinanza per dipingere il quadro completo del fenomeno (Albani et al., 2010).

Le fonti di dati amministrativi rappresentano una preziosa sorgente di informazioni sulla presenza straniera e le migrazioni. Una fonte primaria è rappresentata dalle anagrafi comunali. Sin dal 1993 l'Istat ha condotto annualmente la rilevazione sul "Movimento e calcolo annuale della popolazione straniera residente e struttura per cittadinanza" (nel seguito: "indagine sulla popolazione straniera") che fornisce il bilancio demografico, il numero e la distribuzione per sesso e cittadinanza degli stranieri residenti in ciascuno dei circa otto mila comuni italiani alla fine di ciascun anno di calendario (Albani, Brandimarti, 2012)[5]. I dati trasmessi all'Istat dai comuni sono tratti dagli archivi anagrafici e riflettono la presenza straniera regolarmente documentata nei suddetti registri. Un'altra fonte primaria è il censimento della popolazione che, come noto, è stato sino ad oggi utilizzato anche per la verifica decennale delle informazioni contenute nei registri anagrafici. Tradizionalmente dopo ogni censimento il confronto tra risultanze anagrafiche e censuarie ha dato origine alle così dette operazioni di "revisione anagrafica". Il calcolo della popolazione straniera dopo il censimento viene fatto "ripartire" dalla popolazione censita, salvo poi essere corretto degli eventuali errori censuari (persone effettivamente residenti ma non censite e persone censite ma non effettivamente residenti) individuati con la revisione. Un'altra fonte per la verifica dei dati raccolti con l'indagine sulla popolazione straniera sono le Liste Anagrafiche Comunali (LAC), estratte anch'esse dagli archivi anagrafici e trasmesse all'Istat da ciascun

[4] Come è noto, i comuni italiani sono circa ottomila. Il numero esatto muta nel tempo, a seguito di variazioni amministrative comportanti l'istituzione di nuovi comuni e/o la soppressione di comuni preesistenti.

[5] La rilevazione è inserita, con codice IST-00202, all'interno del Programma statistico nazionale 2011-2013. L'Aggiornamento 2013 è entrato in vigore con la pubblicazione del D.P.C.M. 21 marzo 2013 sulla Gazzetta Ufficiale - serie gen. n. 138 - del 14 giugno 2013, Supplemento ordinario n. 47.

comune. L'acquisizione delle LAC, che contengono i principali dati demografici dei singoli individui residenti nel comune, è stata introdotta ai fini della predisposizione di liste di supporto e di confronto da utilizzare nelle operazioni di rilevazione del Censimento della popolazione e delle abitazioni del 2011, che si è connotato infatti come il primo "censimento assistito da lista" in assoluto.

Per una migliore comprensione del metodo di stima descritto nel presente lavoro è utile una descrizione preliminare delle fonti citate e delle loro caratteristiche e limiti.

2. Le fonti

2.1 Il censimento della popolazione e delle abitazioni

Come è noto, in Italia, sin dal 1861 quasi regolarmente ogni dieci anni si è tenuto il censimento generale della popolazione e delle abitazioni, indagine di natura esaustiva e universale con la quale vengono rilevati il numero e le principali caratteristiche demografiche, sociali ed economiche della popolazione residente.

Tradizionalmente il censimento della popolazione ha rappresentato il nuovo punto di partenza per il calcolo della popolazione straniera residente negli anni tra un censimento e il successivo[6]. Lo stock di popolazione straniera distribuito per genere e cittadinanza, negli anni successivi ai censimenti del 1991, 2001 e 2011 è stato ricalcolato a partire dagli ultimi dati censuari disponibili. A causa degli errori di cui sono affetti i registri, normalmente accade che, al termine di ciascun decennio tra un censimento e il successivo, per molti comuni si riscontri un disallineamento tra l'informazione sulla popolazione residente contenuta nel registro anagrafico comunale e la popolazione censita (Gesano et al., 1993)[7]. Il disallineamento tuttavia può dipendere non solamente dagli errori contenuti nei registri della popolazione, ma anche da errori di sotto/sovra copertura del censimento. La popolazione straniera censita con il XV Censimento della popolazione del 2011, riferita al 9 ottobre dell'anno, per l'intero territorio nazionale è risultata pari a 4.027.627 individui. Nonostante si sia trattato per la prima volta di censimento condotto su base anagrafica, il tradizionale scostamento tra risultati censuari e anagrafici si è verificato anche in questa occasione rendendo negli anni seguenti ancora una volta opportune le operazioni di revisione previste dal Regolamento anagrafico (cfr. paragrafo 2.5)[8]. La tecnica di rilevazione assistita da lista dell'ultimo censimento (che ha fatto uso delle Liste anagrafiche comunali per l'invio dei questionari alle famiglie) ha consentito, per la prima volta, la

[6] Come accennato nell'introduzione, nel 1993 venne lanciata l'indagine sulla popolazione straniera residente. Fu introdotto il Modello Istat P.3, per la raccolta delle informazioni sul bilancio annuale e lo stock a fine anno della popolazione straniera residente nei comuni italiani, per genere e cittadinanza. Si trattava della prima rilevazione di dati amministrativi sul fenomeno, le cui dimensioni erano allora ancora piuttosto limitate (il censimento del 1991 aveva rilevato poco più di 356 mila stranieri residenti nel Paese).

[7] Per gli stranieri si tratta principalmente di mancate cancellazioni. Mentre l'iscrizione in anagrafe garantisce allo straniero una serie di diritti, in particolare di natura socio-sanitaria, la cancellazione dal Registro della popolazione del comune di residenza per emigrazione all'estero non presenta alcun tipo di vantaggio. Si presume quindi che in diversi casi la dichiarazione trasferimento di residenza venga omessa da parte dell'interessato.

[8] D.P.R. 30 maggio1989, n.223, "Approvazione del nuovo regolamento anagrafico della popolazione residente", Pubblicato nella Gazzetta Ufficiale dell'8 giugno 1989, n. 132. In particolare, ci si riferisce qui al Capo VIII "Revisioni da effettuarsi in occasione dei censimenti; altri adempimenti statistici", Articolo 46 "Revisione delle anagrafi".

verifica "micro" ex-post degli scostamenti tra le LAC (alla data del censimento) e la lista delle unità effettivamente censite in ciascun comune (il così detto "confronto censimento-anagrafe, effettuato avvalendosi di un apposito portale denominato SGR "Sistema di Gestione della Rilevazione). La natura "micro-individuale" del confronto censimento-anagrafe ha permesso di progettare e rendere operativo il sistema per il monitoraggio delle derivanti operazioni di revisione degli archivi anagrafici (la così detta "revisione anagrafica, effettuata con l'ausilio del portale SIREA). A valle del processo questa impostazione ha comportato incomparabili miglioramenti nella qualità dei registri rispetto a qualsiasi precedente tornata censuaria, oltre a consentire il corretto calcolo della popolazione residente (cfr. paragrafo 2.3)[9].

2.2 La rilevazione annuale dei bilanci demografici e dello stock di popolazione straniera residente

I bilanci demografici annuali della popolazione straniera e la distribuzione di detta popolazione per genere e cittadinanza vengono elaborati sulla base dei dati raccolti dall'Istat presso ciascun comune con la rilevazione sulla popolazione straniera residente, come accennato in precedenza[10]. Il modello di rilevazione dei dati, somministrato via Internet, contiene il numero di iscrizioni e cancellazioni dall'anagrafe di stranieri e straniere per movimento naturale (nascite e decessi), migrazioni (trasferimenti di residenza tra comuni e da e verso l'estero), altri motivi[11].

I bilanci vengono utilizzati per l'aggiornamento annuale del calcolo della popolazione residente, a partire dalla popolazione censita con l'ultimo censimento disponibile e forniscono, comune per comune, lo stock di popolazione a fine anno, distinto per genere ("popolazione calcolata", nel seguito). Con il modello vengono raccolti anche i dati sugli stranieri iscritti nell'anagrafe del comune a fine anno, distinti per genere e singolo Paese di cittadinanza ("popolazione anagrafica", nel seguito). Quest'ultimo dato deve essere ricondotto ai totali per genere della popolazione calcolata attraverso i bilanci demografici. Per alcuni comuni infatti, anche a valle delle operazioni di revisione dei registri, la popolazione calcolata non coincide con la popolazione anagrafica. Il prodotto di questa riconciliazione nel periodo intercensuario rappresenta l'unica fonte di dati aggiornata, coerente con il calcolo della popolazione, disponibile fino al livello territoriale comunale,

[9] Il sistema SIREA si è giovato in input delle liste derivanti dal confronto a livello del singolo individuo tra i risultati censuari e i dati anagrafici, e ricavate come output del Sistema per la Gestione della Rilevazione (SGR).

[10] Le rilevazioni statistiche concernenti il movimento naturale della popolazione residente ed i trasferimenti di residenza sono previste, tra l'altro, tra i compiti degli Ufficiali di Anagrafe e degli Uffici di Statistica comunale proprio dal già citato Regolamento Anagrafico, in particolare agli articoli 48 e 50 e vengono annualmente regolamentate dall'Istat con un'apposita circolare. Negli anni 2011 e 2012 esse hanno subito nuove profonde trasformazioni, atte a renderle idonee a consentire la ripresa del calcolo della popolazione a partire dal censimento. I dati vengono raccolti con il modello Istat P.2&P.3, che a partire dalla rilevazione dati 2013 compendia le informazioni già contenute nei cessati modelli Istat P.2 e Istat P.3

[11] Il modello viene somministrato attraverso il sito per l'acquisizione dei dati dell'Istat, il sito Indata (https://indata.istat.it/). Tra gli "altri motivi" nel caso degli stranieri, citiamo le cancellazioni per irreperibilità ordinaria e le cancellazioni per mancato rinnovo del permesso di soggiorno. Le cancellazioni per irreperibilità e quelle per mancata dichiarazione della dimora abituale a seguito di rinnovo del permesso di soggiorno sono previste dagli artt.7 e 11 del D.P.R. n. 223/1989 e successive modificazioni o integrazioni.a Negli anni di svolgimento della revisione anagrafica, tra gli altri motivi sono state considerate anche le voci relative alle iscrizioni e le cancellazioni per rettifiche post-censuarie (cfr. paragrafo 2.3).

sul numero, il genere e il Paese di cittadinanza degli stranieri residenti in Italia.

E' opportuno specificare che con l'indagine sulla popolazione straniera residente viene rilevato tanto lo stock di cittadini stranieri comunitari, quanto quello di cittadini extra-comunitari[12]. Secondo i dati di questa fonte, al 1° gennaio 2015 gli stranieri residenti in Italia cittadini di un Paese dell'Unione erano circa 1.492 mila, il 29,8% del totale (circa 5.014 mila).

2.3 Il Sistema SIREA

La Circolare Istat n.15 del 13 dicembre 2011, di intesa con il Ministero dell'Interno, ha stabilito le modalità tecniche e i tempi che i comuni erano chiamati a rispettare nell'esecuzione delle attività di revisione dell'anagrafe a seguito del XV Censimento generale della popolazione, come previsto a norma di legge (D.P.R. 223/1989). Sulla base di tale disposizione di legge e delle successive circolari Istat di natura tecnica (circolari Istat n. 6/2012 e n. 12/2013), i comuni hanno effettuato la revisione anagrafica utilizzando gli elenchi informatizzati delle persone i cui esiti censuari e anagrafici non hanno trovato corrispondenza, disponibili in Istat e derivanti dal Sistema di Gestione della Rilevazione (SGR). Si tratta delle persone che pur iscritte in anagrafe non sono risultate censite nello stesso comune e delle persone censite ma non iscritte nel comune alla data di censimento.

La modalità con la quale si è svolto il XV Censimento generale della popolazione, con lista assistita, ha reso disponibili, per la prima volta, i dati a livello individuale delle persone risultate disallineate, ponendo le premesse per lo sviluppo di un sistema informatizzato *on-line*, che ha consentito ai comuni di documentare l'attività di revisione, utilizzando regole e strumenti condivisi e uniformi: il Sistema di Revisione delle Anagrafi (SIREA). Ogni comune ha potuto in tal modo individuare in modo puntuale le persone risultate disallineate, verificare per ciascuna di esse la situazione riportata nei registri anagrafici e aggiornare i dati, allineando la situazione registrata in anagrafe a quella di fatto, eliminando possibili doppi conteggi di eventi relativi a persone non censite o già censite. Le operazioni di documentazione della revisione delle anagrafi su SIREA sono iniziate a marzo 2012 e si sono concluse il 30 giugno 2014.

Per quanto riguarda la lista delle persone iscritte in anagrafe ma non risultate censite, la revisione è consistita nel verificare per ciascun individuo se era ancora iscritto in anagrafe, confermando la dimora abituale di coloro che nel periodo a ridosso del censimento hanno effettuato un accesso agli uffici demografici per richiedere certificati (carta di identità, cambio domicilio e simili), sono stati rintracciati al proprio domicilio in seguito a verifica del vigile o erano iscritti in liste aggiornate, previste in circolare, attestanti la loro presenza sul territorio comunale (liste scolastiche o di assistenza sanitaria). Allo stesso tempo, sono state documentate le cancellazioni avvenute per movimento anagrafico nel periodo intercensuario (decedute, emigrate in altro comune o all'estero) e quelle delle persone che, a seguito di verifica, sono risultate irreperibili al censimento.

[12] Si rammenta che per i cittadini comunitari non è necessario il possesso di un permesso di soggiorno ai fini della permanenza nel Paese nel medio-lungo periodo. I dati dell'indagine sulla popolazione straniera residente risultano pertanto gli unici dati di fonte amministrativa disponibili nel periodo intercensuario per questo segmento di popolazione. La libera circolazione ed il soggiorno dei cittadini dell'Unione e dei loro familiari nel territorio degli Stati membri è stata sancita con la Direttiva 2004/38/CE del Parlamento europeo e del Consiglio, del 29 aprile 2004. La direttiva è stata recepita dall'Italia con il Decreto legislativo n. 30, del 6 febbraio 2007.

Analogamente, per ciascun individuo risultato censito ma non ancora in anagrafe i comuni hanno indicato la conferma dell'iscrizione anagrafica nel comune (per nascita o immigrazione da altro comune o dall'estero) o la mancanza dei requisiti per la dimora abituale.

L'utilizzo del sistema SIREA ha, pertanto, permesso all'Istat di controllare e monitorare quotidianamente i risultati del lavoro di revisione dei comuni e di calcolare in modo uniforme le rettifiche da apportare al calcolo della popolazione. Infatti, per ciascuna posta documentata su SIREA veniva registrata una rettifica in aggiunta, in sottrazione o nulla. Le rettifiche in aggiunta hanno riguardato le persone che sono state confermate in anagrafe a seguito di verifica e coloro che erano già stati cancellati per movimento anagrafico con decorrenza giuridica post censimento, comportante conteggio. Specularmente, nelle rettifiche in sottrazione sono state registrate le revisioni delle persone che pur censite non avevano i requisiti per la dimora abituale e di coloro che erano stati iscritti in anagrafe nel comune con decorrenza giuridica post censimento, comportante conteggio. Effetti nulli sul calcolo della popolazione hanno avuto i movimenti anagrafici avvenuti con decorrenza giuridica precedente il 9 ottobre 2011 e le cancellazioni per irreperibilità censuaria.

Le rettifiche in aggiunta e quelle in sottrazione inserite nei modelli Istat P.2&P.3 alle voci "Iscritti per rettifiche post censuarie" e "Cancellati per rettifiche post censuarie" hanno, pertanto, concorso alla determinazione dei bilanci demografici annuali per gli anni 2012-2014.

2.4 Le Liste Anagrafiche Comunali

Le LAC sono collezioni di dati individuali, riguardanti i residenti e tratte dai registri anagrafici di ciascuno dei comuni italiani. Nelle LAC ciascun record è un singolo individuo. Oltre ai dati identificativi della persona (cognome, nome, codice fiscale) sono riportate, secondo un prefissato tracciato record, le principali informazioni di natura demografica e sociale (genere, luogo e data di nascita, stato civile, cittadinanza, indirizzo di residenza dell'individuo, ecc.). Queste informazioni costituiscono una fotografia del registro anagrafico comunale, in un certo istante di tempo (normalmente il 31 dicembre dell'anno); sono un'istantanea della popolazione che risulta iscritta in anagrafe in quel momento, secondo alcune sue caratteristiche.

Ai fini delle operazioni di censimento, le LAC sono state utilizzate per l'invio dei questionari ai residenti (Crescenzi et al., 2009). Integrate nel Sistema di Gestione della Rilevazione (SGR), esse hanno inoltre costituito la base per il così detto "confronto censimento-anagrafe", operazione preliminare alla revisione anagrafica (cfr. paragrafo 2.3).

La raccolta delle LAC, tuttavia, non è stata limitata all'occasione del censimento. Negli anni seguenti l'Istat ha continuato a richiedere ai comuni le liste anagrafiche. Le LAC rappresentano una fonte innovativa di dati sugli individui residenti, le famiglie e le convivenze, fonte che è stato possibile utilizzare, per esempio, per la verifica dei dati statistici altrimenti raccolti o elaborati dall'Istat[13].

[13] Con riferimento all'oggetto specifico di questo lavoro i dati individuali relativi alle variabili genere e Stato estero di cittadinanza contenuti nelle LAC, opportunamente trattati e aggregati, sono stati utilizzati come termine di confronto per l'integrazione/correzione delle informazioni sui Paesi di cittadinanza degli stranieri residenti contenute nei modelli Istat P.2&P3. L'applicazione dei piani di check previsti dalla procedura di validazione dei dati dell'indagine sulla

2.5 Il confronto tra fonti: il "paradosso delle tre popolazioni" e la distribuzione degli stranieri per cittadinanza

Una pluralità di fonti disponibili rappresenta sicuramente una ricchezza dal punto di vista informativo. Nel caso specifico essa pone nondimeno alcune problematiche, su cui vale la pena di soffermarci, soprattutto in riferimento all'ultima esperienza censuaria. Lo scopo è quello fornire un quadro più preciso del contesto all'interno del quale il processo di stima esaminato in questo lavoro si è collocato.

L'apparente "paradosso delle "tre popolazioni" si appalesa, ciclicamente, in occasione di ogni censimento della popolazione, in corrispondenza della cui data di esecuzione è possibile accertare l'esistenza di tre diverse misure della popolazione abitualmente dimorante:

1. la popolazione calcolata (dall'Istat)[14];
2. la popolazione censita (legale);
3. la popolazione anagrafica (vale a dire il numero di schede individuali contenute nello schedario anagrafico, corrispondente – in pratica – alla LAC).

A causa di errori di varia natura, imputabili alla stessa esecuzione della rilevazione censuaria e a errori conteggio e di tenuta dei registri anagrafici nel corso del decennio intercensuario, si può verificare che per lo stesso aggregato (la popolazione abitualmente dimorante, cioè la popolazione residente), alla data del censimento vengano riscontrate fino a tre misurazioni. Al 31 dicembre 2010 gli stranieri residenti, secondo il calcolo della popolazione effettuato sommando ai dati del censimento precedente (2001) le iscrizioni e le cancellazioni anagrafiche rilevate nei bilanci demografici intercensuari, erano circa 4 milioni e 570 mila. Le schede anagrafiche, secondo quanto riportato nei modelli Istat P.3 del 2010, erano circa 4 milioni e 644 mila.

La popolazione straniera censita al 9 ottobre 2011 è risultata pari a 4.027.627 unità, inferiore sia alla popolazione calcolata che a quella anagrafica riferite a circa nove mesi prima[15].

Alla data del nuovo censimento il calcolo della popolazione basato sul censimento del 2001 e sui dati dell'indagine sulla popolazione straniera residente aveva aggiornato il livello dello stock di popolazione straniera valutandolo in 4.790.405 individui. Lo scarto con la popolazione censita era di oltre 760 mila individui, con una variazione percentuale pari a -15,9%. Se come termine di confronto con la popolazione censita si considera lo stock di popolazione anagrafica al 9 ottobre 2011, il divario aumenta ulteriormente (quasi 816 mila unità, pari a -16,8%). Questi risultati, con i relativi scarti in valore assoluto e percentuale, sono riportati nella Tavola 1.

popolazione straniera residente ha infatti evidenziato la presenza di mancate risposte parziali o incongruenze per alcuni comuni. Le LAC sono state utilizzate inoltre per la validazione dei dati sulle cancellazioni per acquisizione della cittadinanza italiana (cfr. anche nota 24).

[14] Aggiungendo, come si è detto (cfr. paragrafo 2.2), le poste del movimento demografico naturale e migratorio e del movimento per altri motivi alla popolazione censita (legale) in occasione del precedente censimento.

[15] L'Indagine di copertura del XV Censimento generale della popolazione e delle abitazioni (PES), condotta tra l'Aprile e il Luglio 2012, ha peraltro individuato un errore censuario (sotto copertura netta) pari a circa 650 mila individui, di cui circa 500 mila stranieri (Istat, 2012).

Tavola 1 - Popolazione straniera residente in Italia calcolata e anagrafica e popolazione censita al 9 ottobre 2011, per genere *(valori assoluti e differenze assolute e percentuali)*

	Popolazione al 9 ott 2011			Differenze assolute			Differenze percentuali		
	Calcolata (a)	Anagrafica (b)	Censita (c)	(b)-(a)	(c)-(a)	(c)-(b)	[(b)-(a)]/(a) *100	[(c)-(a)]/(a) *100	[(c)-(b)]/(b) *100
Uomini	2.302.228	2.334.961	1.881.030	32.733	-421.198	-453.931	1,4	-18,3	-19,4
Donne	2.488.177	2.508.764	2.146.597	20.587	-341.580	-362.167	0,8	-13,7	-14,4
Totale	**4.790.405**	**4.843.502**	**4.027.627**	**53.097**	**-762.778**	**-815.875**	**1,1**	**-15,9**	**-16,8**

Fonte: elaborazione su dati Istat (Censimento della popolazione 2011, Rilevazione sul Movimento e calcolo della popolazione straniera residente e struttura per cittadinanza)

Si osserva in particolare che le differenze in questione hanno riguardato, in tutti i casi, soprattutto gli uomini. Questo fatto potrebbe testimoniare una maggiore difficoltà di conteggiare gli stranieri non raggruppati in famiglie.
Se si vuole approfondire l'analisi per le diverse cittadinanze, l'ultimo dato pubblicato è quello al 31 dicembre 2010[16]. Nella Tavola 2 si riporta la distribuzione della popolazione straniera residente a tale data per le prime venticinque cittadinanze in ordine di importanza numerica, accanto alle analoghe distribuzioni della popolazione straniera iscritta nelle anagrafi e di quella censita.

Nel complesso, i venticinque Paesi raggruppano poco più dell'85% del totale degli stranieri residenti. Come si osserva dalla tabella, per alcuni Paesi l'ordine in graduatoria e l'importanza percentuale sono differenti nelle tre distribuzioni e le differenze appaiono più marcate tra la distribuzione al censimento e quella della popolazione calcolata (o anche dell'anagrafica) all'inizio del 2011 di quanto non risulti dalla calcolata e dall'anagrafica. Tra le cittadinanze più rappresentative osserviamo in particolare nella distribuzione al censimento una riduzione di 0,8 punti percentuali della quota dei cittadini Rumeni, che si sostanzia in un minore ammontare per oltre 145 mila unità. Nella stessa distribuzione si osservano, all'inverso, incrementi delle quote percentuali spettanti alla maggior parte delle altre principali cittadinanze.

I dati della tavola 2, in conclusione, mostrano che le differenze tra le tre popolazioni non si limitano all'ammontare complessivo degli stock rilevati, ma interessano in misura a volte non irrilevante anche la distribuzione per cittadinanza.

Non disgiunto da queste considerazioni è il ragionamento sull'impatto della revisione anagrafica post censuaria sul calcolo della popolazione straniera residente e della sua distribuzione per genere e Paese di cittadinanza, che viene brevemente esposto nel paragrafo seguente.

[16] Non è stato in effetti mai calcolato e diffuso un dato al 9 ottobre 2011 derivante dall'indagine sulla popolazione straniera residente, in quanto a quella data il dato di popolazione ufficiale è rappresentato per definizione dal dato censuario.

Tavola 2 - Popolazione straniera residente in Italia calcolata e anagrafica al 31 dicembre 2010 e popolazione censita al 9 ottobre 2011 *(primi venticinque Paesi di cittadinanza)*

Popolazione calcolata al 31dic 2010			Popolazione Anagrafica al 31dic 2010			Popolazione Censita al 9 ott 2011		
Paese	Stranieri residenti	%		Stranieri residenti	%	Paese	Stranieri residenti	%
Romania	968.576	21,2	Romania	974.215	21,0	Romania	823.100	20,4
Albania	482.627	10,6	Albania	483.664	10,4	Albania	451.437	11,2
Marocco	452.424	9,9	Marocco	454.977	9,8	Marocco	407.097	10,1
Cinese, Repubblica Popolare	209.934	4,6	Cinese, Repubblica Popolare	212.294	4,6	Cinese, Repubblica Popolare	194.510	4,8
Ucraina	200.730	4,4	Ucraina	200.936	4,3	Ucraina	178.534	4,4
Filippine	134.154	2,9	Filippine	141.302	3,0	Moldova	130.619	3,2
Moldova	130.948	2,9	Moldova	130.976	2,8	Filippine	129.015	3,2
India	121.036	2,6	India	122.649	2,6	India	116.797	2,9
Polonia	109.018	2,4	Polonia	111.244	2,4	Perù	93.905	2,3
Tunisia	106.291	2,3	Tunisia	108.360	2,3	Polonia	84.619	2,1
Perù	98.603	2,2	Perù	100.346	2,2	Tunisia	82.066	2,0
Ecuador	91.625	2,0	Egitto	93.621	2,0	Ecuador	80.645	2,0
Egitto	90.365	2,0	Ecuador	91.717	2,0	Bangladesh	80.639	2,0
Macedonia, Repubblica di	89.900	2,0	Macedonia, Repubblica di	90.031	1,9	Macedonia, Repubblica di	73.407	1,8
Bangladesh	82.451	1,8	Sri Lanka (ex Ceylon)	87.263	1,9	Senegal	72.458	1,8
Sri Lanka (ex Ceylon)	81.094	1,8	Bangladesh	85.206	1,8	Sri Lanka (ex Ceylon)	71.203	1,8
Senegal	80.989	1,8	Senegal	81.852	1,8	Pakistan	69.877	1,7
Pakistan	75.720	1,7	Pakistan	76.057	1,6	Egitto	65.985	1,6
Nigeria	53.613	1,2	Nigeria	54.812	1,2	Nigeria	47.338	1,2
Serbia, Repubblica di	52.954	1,2	Serbia, Repubblica di	53.283	1,1	Ghana	44.031	1,1
Bulgaria	51.134	1,1	Bulgaria	51.285	1,1	Serbia, Repubblica di	43.608	1,1
Ghana	46.890	1,0	Brasile	47.550	1,0	Kosovo	41.575	1,0
Brasile	46.690	1,0	Ghana	47.498	1,0	Bulgaria	40.982	1,0
Germania	42.531	0,9	Germania	44.518	1,0	Brasile	37.208	0,9
Francia	33.400	0,7	Francia	35.237	0,8	Germania	35.109	0,9
Totale 15 Paesi	**3.933.697**	**86,1**	**Totale 15 Paesi**	**3.980.893**	**85,7**	**Totale 15 Paesi**	**3.495.764**	**86,8**
Totale	**4.570.317**	**100,0**	**Totale**	**4.643.576**	**100,0**	**Totale**	**4.027.627**	**100,0**

Fonte: elaborazione su dati Istat (Censimento della popolazione 2011, Rilevazione sul Movimento e calcolo della popolazione straniera residente e struttura per cittadinanza)

2.6 La revisione anagrafica dei cittadini stranieri

Alla data del censimento, l'ammontare dei cittadini stranieri presenti nelle anagrafi comunali ma risultati non censiti è stato pari a 1.017.078 unità, (il 42,7% del totale degli individui irreperibili al censimento); mentre sono risultati censiti ma non residenti nel comune di censimento 192.279 stranieri (il 27% del totale dei censiti non iscritti in LAC).

In particolare, per quanto riguarda gli irreperibili al censimento, la documentazione della revisione, effettuata dai comuni tra il 2012 e il 2014, ha messo in luce che il 25,6% del totale è sfuggito alla rilevazione in quanto stava effettuando un movimento anagrafico a cavallo della

tornata censuaria, il 33,5% ha confermato la propria presenza nel comune di residenza, mentre il 38,7% è effettivamente stato cancellato dall'anagrafe per irreperibilità censuaria[17].

Le cancellazioni per irreperibilità censuarie, effettuate secondo le disposizioni del regolamento anagrafico (art.11, D.P.R. n. 223/1989), hanno consentito ai comuni di depennare dai registri anagrafici coloro che, a seguito di ripetuti accertamenti, sono risultati non rintracciabili all'indirizzo di residenza indicato. Tali cancellazioni, che non comportano conteggio ai fini del calcolo della popolazione, rappresentano uno strumento legislativo per migliorare la qualità dei registri anagrafici, rendendoli convergenti alla situazione *de facto*.

Al termine delle operazioni di revisione post-censuaria dell'anagrafe, i cittadini stranieri cancellati per irreperibilità censuaria sono risultati pari a 393.248 unità: di questi, in seguito a verifica, ben il 10% è stato cancellato per doppio mancato censimento in quanto non era stato censito neanche nel 2001.

L'analisi della distribuzione delle prime quindici cittadinanze per numero di cancellati per irreperibilità censuaria evidenzia differenze tra le diverse collettività a non cancellarsi dalle anagrafi in caso di trasferimento in altro comune o all'estero (Tavola 3).

Tavola 3 - Cancellazioni di cittadini stranieri dalle Anagrafi per revisione post-censuaria per motivo e cittadinanza. Anni 2012-2014 *(valori assoluti e percentuali)*

Paese di cittadinanza	Stranieri cancellati per irreperibilità censuaria		Paese di cittadinanza	Stranieri cancellati per altro motivo non comportante conteggio	
	N.	% su totale		N.	% su totale
Romania	93.518	23,8	Romania	4.098	28,9
Marocco	24.666	6,3	Marocco	1.211	8,5
Cinese, Repubblica Popolare	18.187	4,6	Albania	735	5,2
Albania	16.324	4,2	Cinese, Repubblica Popolare	636	4,5
Polonia	14.618	3,7	Ucraina	566	4,0
Egitto	14.402	3,7	Tunisia	550	3,9
Tunisia	13.738	3,5	Polonia	509	3,6
Ucraina	11.976	3,0	Brasile	472	3,3
Sri Lanka (ex Ceylon)	9.006	2,3	India	335	2,4
Francia	8.819	2,2	Egitto	260	1,8
India	7.318	1,9	Germania	231	1,6
Germania	7.072	1,8	Moldova	206	1,5
Regno Unito	7.008	1,8	Senegal	203	1,4
Brasile	6.741	1,7	Bangladesh	200	1,4
Bulgaria	6.223	1,6	Sri Lanka (ex Ceylon)	196	1,4
Altre cittadinanze	133.632	34,0	Altre cittadinanze	3.795	26,7
Totale	**393.248**	**100,0**	**Totale**	**14.203**	**100,0**

Fonte: Istat, Sistema di Revisione delle Anagrafi (SIREA)

Un'altra componente che ha effetti nulli sul calcolo dei bilanci comunali della popolazione è rappresentata dai cancellati per altro motivo non altrimenti classificabile. Si

[17] La percentuale residua degli stranieri irreperibili al censimento è stata revisionata come "errore di lista", modalità con cui venivano classificate in SIREA le persone che, pur in anagrafe e censite, sono state considerate erroneamente tra gli irreperibili.

tratta di casi residuali di individui non censiti che, essendo stati cancellati dai registri anagrafici senza comportare conteggio, si possono ritenere in qualche modo assimilabili ai cancellati per irreperibilità censuaria.

In particolare, i cittadini stranieri cancellati dalle anagrafi per altro motivo non comportante conteggio sono risultati pari a 14.203 unità: l'1,4% del totale dei non censiti iscritti in LAC.

3. Il calcolo della popolazione residente straniera a partire dalla popolazione censita e la stima delle distribuzioni per genere e Paese di cittadinanza

Si è già evidenziato come dopo ogni censimento il calcolo dello stock di popolazione straniera residente riparta dalla popolazione censita, applicando anno per anno i dati sui flussi (naturale, migratorio e per altri motivi) contenuti nei bilanci demografici rilevati con l'indagine sulla popolazione straniera residente. La procedura determina un'interruzione della serie storica della popolazione, in quanto il nuovo punto di partenza per il calcolo diviene la popolazione straniera censita con l'ultimo censimento e non più quella censita con il censimento precedente[18].

3.1 I bilanci demografici e il calcolo della popolazione

Come accennato (cfr. paragrafo 2.2), per ciascun anno t e per ciascun comune c, il calcolo della popolazione a fine anno $P_{t_2}^c$ si effettua sommando alla popolazione iniziale[19] $P_{t_1}^c$ il saldo naturale $Sn_{t_2-t_1}$, il saldo migratorio $Sm_{t_2-t_1}$, il saldo per altri motivi $Sa_{t_2-t_1}$, registrati nell'anno (equazione della popolazione) [20]:

$$P_{t_2}^c = P_{t_1}^c + Sn_{t_2-t_1} + Sm_{t_2-t_1} + Sa_{t_2-t_1} \quad (1)$$

Negli anni immediatamente successivi al censimento, tra gli altri motivi si deve tenere conto anche delle iscrizioni e cancellazioni anagrafiche per rettifiche post-censuarie al calcolo della popolazione. Si tratta delle iscrizioni di persone erroneamente non censite $I_{t_2-t_1}^{vc}$ e delle cancellazioni di persone erroneamente censite $C_{t_2-t_1}^{vc}$, effettuate dal comune c

[18] Al fine di recuperare la confrontabilità in serie storica con i dati passati, dopo ogni censimento viene effettuata una ricostruzione all'indietro della popolazione residente, riferita agli anni tra l'ultimo censimento e il precedente. La ricostruzione, agendo sui saldi migratorio e per altri motivi dei bilanci demografici annuali intercensuari, consente di riportare la popolazione calcolata alla data del censimento a partire dai dati del censimento precedente, già pubblicata, a coincidere con la popolazione censita con l'ultimo censimento disponibile. La ricostruzione per la popolazione straniera (complessiva e non distinta per Paese di cittadinanza) è stata effettuata per la prima volta per il decennio 2002-2011.

[19] La popolazione censita se si tratta di anno di censimento, la popolazione calcolata all'inizio dell'anno per gli anni successivi.

[20] Siano rispettivamente t_1 il primo gennaio dell'anno di riferimento t e t_2 il 31 dicembre dello stesso anno. Poiché convenzionalmente i dati di popolazione vengono riferiti alla mezzanotte, il dato al primo gennaio dell'anno t coincide con il dato al 31 dicembre dell'anno $t - 1$.

nell'anno t[21]:

$$Sa^{vc}_{t_2-t_1} = I^{vc}_{t_2-t_1} - C^{vc}_{t_2-t_1} \quad (2)$$

In occasione dell'ultimo censimento il saldo delle rettifiche post-censuarie $Sa^{vc}_{t_2-t_1}$ è stato calcolato a partire dagli esiti individuali della revisione anagrafica, documentati con il sistema SIREA (cfr. paragrafo 2.3). Al termine delle operazioni di revisione anagrafica teoricamente, per effetto del recupero nel calcolo degli errori di censimento e della "ripulitura" delle anagrafi degli errori in essa contenuti, la popolazione calcolata e la popolazione anagrafica dovrebbero coincidere. Di fatto, per ragioni che non è il caso di esaminare nel dettaglio in questa sede, per alcuni comuni delle differenze permangono, anche se il *gap* tra le due popolazioni risulta solitamente ridotto[22].

3.2 La stima per comune della popolazione straniera residente calcolata per genere e Stato estero di cittadinanza.

Per i comuni per i quali, nonostante la revisione anagrafica, per gli anni di rilevazione 2011 e successivi, a fine anno permanga una differenza tra popolazione calcolata e popolazione anagrafica si pone il problema di ricondurre i dati sulla popolazione straniera iscritta in anagrafe distribuita per genere e Paese di cittadinanza ai totali di popolazione calcolati attraverso i bilanci demografici (cfr. paragrafo 2.2). Comune per comune, lo stock (distinto per genere) derivante dal calcolo rappresenta l'obiettivo cui far convergere la distribuzione, al fine di produrre il dato ufficiale sugli stranieri residenti distribuiti per genere e Paese di cittadinanza alla fine di ciascun anno di calendario.

L'algoritmo per l'allineamento dei dati anagrafici sugli stranieri residenti nei comuni italiani per genere e cittadinanza ai totali per genere della popolazione calcolata sulla base dell'equazione della popolazione (1) si fonda sul principio che la medesima equazione, valida per i cittadini stranieri nel complesso, è valida anche a livello della singola cittadinanza.

Il punto di partenza è la popolazione censita (o quella calcolata al 31 dicembre dell'anno precedente), distribuita per comune, genere e Stato estero di cittadinanza. Dall'equazione (1) si evince che, per il calcolo della popolazione a fine anno distribuita per genere e Stato estero di cittadinanza occorrerebbe conoscere i dati sui flussi naturale, migratorio e per altri motivi, a livello comunale e distribuiti anch'essi per genere e cittadinanza. I dati tratti dai bilanci demografici (i soli ad oggi disponibili tempestivamente, ossia con cinque-sei mesi di ritardo-data) tuttavia forniscono solo i flussi distribuiti per genere, senza distinzione in base alla variabile cittadinanza[23]. Il metodo di stima si basa sulla proprietà che il saldo complessivo

[21] Gli erroneamente non censiti sono le persone non censite, ma verificate come effettivamente residenti nel comune alla data del censimento con la revisione anagrafica, gli erroneamente censiti sono le persone censite, ma verificate come effettivamente non residenti con la revisione anagrafica.

[22] Al 31 dicembre 2013, quindi praticamente quasi al termine delle operazioni di revisione anagrafica relative all'ultimo censimento della popolazione (che si sono chiuse a settembre 2014), degli 8.052 comuni con stranieri residenti (sugli 8.092 complessivamente esistenti a quella data) i comuni in cui la popolazione calcolata e la popolazione anagrafica ancora non coincidevano erano 3.057, quasi il 38%; tuttavia per 1.640 comuni, il 20,4%, la differenza tra le due popolazioni era inferiore o uguale a 3 unità. Complessivamente la popolazione calcolata risultava ancora inferiore rispetto alla popolazione anagrafica di circa 101 mila unità (-2%).

[23] Attualmente i flussi di bilancio comportanti conteggio distribuiti per genere e Stato estero di cittadinanza possono essere ricavati (direttamente o indirettamente) dalle rilevazioni di dati individuali sulle nascite e sui decessi (Modelli

(migratorio, naturale e per altri motivi), distinto per genere e singolo Paese di cittadinanza, può essere ricavato in modo indiretto, utilizzando i dati sulla distribuzione della popolazione straniera che risulta iscritta nelle anagrafi comunali alla fine di ciascun periodo.

3.2.1 I saldi anagrafici

I dati sullo stock di popolazione straniera anagrafica, distribuita per genere e cittadinanza possono essere utilizzati per valutare indirettamente il saldo da applicare alla popolazione calcolata alla fine del periodo precedente per ottenerne l'aggiornamento a fine anno. Per i comuni con popolazione anagrafica e calcolata non coincidenti occorre procedere all'allineamento della distribuzione anagrafica alla corrispondente popolazione calcolata[24]. Comune per comune, i dati sugli stranieri distribuiti per genere e cittadinanza devono essere portati a coincidere, se totalizzati rispetto alla variabile cittadinanza, con i totali per genere definiti con il calcolo della popolazione.

Detto $St_{t_2-t_1}^{p,g,c}$ il saldo totale relativo alla popolazione straniera di genere g (ove g=uomo/donna) e Paese di cittadinanza p (con p che assume una delle modalità di cui nella classificazione Istat degli Stati esteri – cfr. Appendice) che risulta iscritta nell'anagrafe del comune c, questo può essere scomposto nelle sue componenti (saldo naturale, migratorio, per altri motivi):

$$St_{t_2-t_1}^{p,g,c} = Sn_{t_2-t_1}^{p,g,c} + Sm_{t_2-t_1}^{p,g,c} + Sa_{t_2-t_1}^{p,g,c} \quad (3)$$

Ma il saldo (3) si può evidentemente anche esprimere - cfr. la (1) - come:

$$St_{t_2-t_1}^{p,g,c} = P_{t_2}^{p,g,c} - P_{t_1}^{p,g,c} \quad (4)$$

ossia come differenza tra la popolazione residente anagrafica a fine anno e la popolazione residente anagrafica alla fine dell'anno precedente. La differenza tra i due stock di popolazione è infatti proprio il risultato delle operazioni di iscrizione e cancellazione degli individui dall'anagrafe (per tutte le sue componenti dinamiche) intervenute nel periodo. Il saldo (4) registra le variazioni intervenute nei registri anagrafici. I saldi per genere e cittadinanza di cui nella (3) non sono noti, in quanto nei bilanci

Istat P.4 e Istat P.5, rispettivamente), dai dati raccolti con la rilevazione dell'Istat sui trasferimenti di residenza della popolazione (modello APR.4) o indirettamente dalle LAC (per le cancellazioni per acquisizione della cittadinanza italiana, derivabili attraverso il confronto tra LAC riferite a anni successivi). La rilevazione sui trasferimenti di residenza consente di distinguere la tipologia di iscrizione/cancellazione dall'anagrafe e può fornire la quantificazione per comune, genere e cittadinanza degli individui interessati da flussi comportanti conteggio. Tuttavia l'organizzazione delle suddette rilevazioni, estesa ad esempio nel caso del modello APR/4 a milioni di trasferimenti, al momento non permette di disporre di dati validati entro i tempi previsti dal Regolamento europeo sulle statistiche sulle migrazioni per la fornitura degli stock di popolazione straniera residente a fine anno.

[24] La distribuzione anagrafica per genere e cittadinanza degli stranieri residenti nel comune a fine anno, rilevata nella seconda sezione del modello Istat P.2&P.3, per essere correttamente utilizzata ai fini del calcolo indiretto dei saldi deve essere preliminarmente emendata da eventuali errori sistematici o casuali. Si tratta principalmente di eliminare possibili doppioni nelle serie delle cittadinanze comunali - cittadinanze ripetute, all'interno del modello di rilevazione, causa errata indicazione di uno o più Paesi - e di risolvere i casi di sbalzi anomali e ingiustificati nella serie storica dei dati comunali, dovuti anch'essi di solito a errata indicazione del Paese di cittadinanza. Questi errori vengono individuati e corretti in una prima fase del processo di validazione, ricorrendo a metodologie che prevedono il confronto dei dati in serie storica e con i corrispondenti dati contenuti nelle LAC.

demografici della popolazione straniera residente la variabile cittadinanza non è contemplata. E invece nota la distribuzione per genere e cittadinanza dello stock di popolazione anagrafica alla fine di ciascun anno.

La popolazione calcolata al tempo t_2, per il comune c, di genere g e Paese di cittadinanza p, $Pc_{t_2}^{p,g,c}$, può pertanto essere ricavata a partire da quella calcolata al tempo t_1 con la formula seguente:

$$Pc_{t_2}^{p,g,c} = Pc_{t_1}^{p,g,c} + St_{t_2-t_1}^{p,g,c} = Pc_{t_1}^{p,g,c} + \left(Pa_{t_2}^{p,g,c} - Pa_{t_1}^{p,g,c}\right) \quad (5)$$

in cui, con ovvio significato dei simboli, $Pa_{t_2}^{p,g,c}$ e $Pa_{t_1}^{p,g,c}$ sono rispettivamente la popolazione residente anagrafica al tempo t_2 e quella al tempo t_1.

Questa valutazione indiretta del saldo, calcolabile comune per comune distintamente per genere e cittadinanza, consente l'aggiornamento del calcolo della popolazione anche in assenza di dati sui flussi naturale, migratorio e per altri motivi distribuiti secondo le medesime variabili.

La formula per la stima indiretta dei saldi si complica leggermente per gli anni di svolgimento della revisione anagrafica.

3.2.2 Il fattore di correzione legato alle operazioni di revisione anagrafica

Nel periodo immediatamente successivo al censimento, sul saldo anagrafico (4) non incidono solo i saldi di cui nella (3), relativi ad operazioni comportanti conteggio rispetto alla popolazione censita. Gli archivi anagrafici vedono infatti aumentare o diminuire di numero gli individui in essi registrati anche per effetto di iscrizioni e cancellazioni non comportanti conteggio nel calcolo della popolazione a partire dalla popolazione censita[25]. Il saldo (4) può essere definito come un "saldo lordo". Per valutare correttamente l'effettivo incremento di popolazione rispetto al censimento, durante la revisione anagrafica occorre depurare il saldo lordo (4) dell'effetto dei movimenti non comportanti conteggio. Se non lo si facesse, si rischierebbe rispettivamente di sopravvalutare (nel caso di presenza di iscrizioni non comportanti conteggio) e di sottovalutare (nel caso di cancellazioni) la popolazione calcolata a fine periodo (5). Si tratta infatti di mere operazioni anagrafiche che tuttavia, in quanto tali, si trovano di fatto automaticamente incluse nel saldo lordo (4). Pertanto queste poste devono essere sottratte dal saldo lordo al fine di determinare l'effettivo incremento/decremento della popolazione.

Per l'anno t, il comune c, il Paese di cittadinanza p e il genere g si introduce pertanto il fattore di correzione $C_{t_2-t_1}^{p,g,c}$ da applicare in detrazione al saldo (4):

$$C_{t_2-t_1}^{p,g,c} = Incc_{t_2-t_1}^{p,g,c} - Cncc_{t_2-t_1}^{p,g,c} \quad (6)$$

In cui $Incc_{t_2-t_1}^{p,g,c}$ sono le iscrizioni non comportanti conteggio di individui di

[25] Si tratta in particolare delle iscrizioni di individui già censiti ma erroneamente non iscritti in anagrafe alla data del censimento e delle cancellazioni di individui già non censiti ma erroneamente risultanti come iscritti in anagrafe alla data del censimento. Che si tratti di errori dell'anagrafe viene verificato, con le previste procedure, in sede di revisione anagrafica, utilizzando le procedure previste dalla normativa. Le iscrizioni e le cancellazioni non comportanti conteggio non debbono essere riportate nei bilanci demografici ai fini del calcolo della popolazione.

cittadinanza p e genere g, effettuate nel comune c nel periodo considerato, mentre $Cncc_{t_2-t_1}^{p,g,c}$ sono le cancellazioni, con analogo significato dei simboli.

Il saldo lordo (4) con l'applicazione del fattore di correzione (6) diviene un saldo netto (corretto)[26]:

$$Stc_{t_2-t_1}^{p,g,c} = (P_{t_2}^{p,g,c} - P_{t_1}^{p,g,c}) - C_{t_2-t_1}^{p,g,c} \quad (7)$$

Il fattore di correzione (6) è derivabile dagli esiti della revisione anagrafica registrati in SIREA, e consente il calcolo corretto del saldo netto da applicare alla popolazione di partenza negli anni subito dopo il censimento, al fine di determinare correttamente la distribuzione della popolazione residente calcolata a fine anno, per genere e cittadinanza.

3.2.3 Le procedura per l'allineamento alla popolazione calcolata: la gestione dei resti, la verifica della quadratura e la validazione dei dati

Per la determinazione della distribuzione della popolazione straniera per genere e Stato estero di cittadinanza che riproduce i totali per genere calcolati sulla base dei bilanci della popolazione si utilizza una procedura che prevede diversi *step*[27].

Innanzitutto si individuano i comuni per i quali la popolazione calcolata a fine periodo e la popolazione anagrafica, per genere, coincidono. Per questi comuni non si rende necessaria alcuna operazione di allineamento della distribuzione della popolazione per genere e cittadinanza: viene considerata valida la distribuzione anagrafica. Essi vengono pertanto estromessi dalle operazioni successive e messi da parte. Vengono successivamente separati in due sottogruppi i restanti comuni, con popolazione calcolata e anagrafica non coincidenti, in funzione dell'entità della differenza tra le due popolazioni: sono trattati per primi i casi per i quali la differenza in valore assoluto, distintamente per genere, risulta minore o uguale a una soglia prefissata[28]. Per questi casi viene eseguita una semplice attribuzione della differenza alla cittadinanza più frequente. Per i casi con differenza in valore assoluto superiore alla soglia, invece, è previsto un algoritmo di allineamento più elaborato. Si applica la (5), nella quale per gli anni di svolgimento della revisione

[26] Nella (6) il minuendo (iscrizioni di individui già censiti) assume, per la popolazione straniera, normalmente un valore molto ridotto rispetto al sottraendo. Sono infatti molto più frequenti i casi di cancellazioni per irreperibilità censuaria e per altri motivi non comportanti conteggio che vengono effettuate nel periodo post censuario durante le operazioni di revisione anagrafica eseguite dai comuni. Anche per il censimento del 2011 ciò si è verificato. L'entità delle iscrizioni relative ad individui già censiti ma non residenti in anagrafe è stata così esigua, che si è deciso di non considerarla nel fattore di correzione.
Le cancellazioni per irreperibilità censuaria considerate ai fini della validazione della distribuzione dei cittadini stranieri residenti al 1° gennaio 2014 sono state complessivamente oltre 319 mila. Hanno interessato complessivamente oltre 4 mila comuni, hanno riguardato per oltre il 57% uomini, per il 22,6% cittadini rumeni, per il 5,7% marocchini, per il 4,5% cinesi, per il 4,1% albanesi, per il 4% egiziani. Le cancellazioni relative ad altre operazioni non comportanti conteggio sono state oltre 7 mila e hanno interessato oltre 400 comuni. La distribuzione per genere e secondo le prime cinque cittadinanze maggiormente interessate è analoga a quella già evidenziate per le cancellazioni per irreperibilità, ma al 4° e 5° posto nella graduatoria si trovano gli Afghani (4,1%) e i Polacchi (3,6%) rispettivamente.

[27] La procedura è realizzata in ambiente SAS.

[28] La soglia predefinita è pari a tre unità: normalmente garantisce una qualità sufficiente ma può essere, in caso di necessità, variata a piacimento o determinata in valore percentuale, anziché in valore assoluto. La procedura prevede in ogni caso la possibilità di verificare a posteriori squilibri nelle distribuzioni introdotti eventualmente dal metodo nei comuni con numero esiguo di stranieri.

anagrafica il saldo $St_{t_2-t_1}^{p,g,c}$ è il saldo netto $Stc_{t_2-t_1}^{p,g,c}$ calcolato come nella (7). Si ricalcolano quindi i totali per genere della popolazione calcolata $Pc_{t_2}^{p,g,c}$ così determinata e si verifica la presenza di un eventuale persistente scostamento. Lo scostamento residuo viene eliminato redistribuendo il resto proporzionalmente rispetto alla distribuzione determinata al passo precedente. Per questo sottogruppo di comuni, l'attribuzione dello scarto proporzionalmente a tutte le cittadinanze (non solo, quindi, sulla più frequente) rappresenta una tutela nei confronti della possibile introduzione di distorsioni nella distribuzione, nei casi in cui lo scarto stesso dovesse rivelarsi, a valle dell'applicazione della (7), ancora importante.

Può accadere che alla fine di questo processo le popolazioni ancora non convergano, per qualche comune. Ciò succede quando per ottenere l'allineamento l'algoritmo dovrebbe determinare delle frequenze negative, o in tutti gli altri casi in cui esso non riesce a risolvere automaticamente la quadratura. Nella procedura sono previsti dei vincoli volti ad evitare l'introduzione di frequenze negative o l'attribuzione automatica di resti eccessivamente elevati, in valore assoluto o in proporzione. Questi casi particolari, che si verificano normalmente in numero molto ridotto (nell'ordine delle unità), vengono segnalati dalla procedura e proposti per essere risolti con intervento ragionato manuale da parte dell'operatore esperto della materia.

Risolti i casi particolari residuali la procedura viene fatta rigirare e normalmente arriva a convergenza.

Al termine del processo descritto sopra finalmente si perviene, per ciascun comune interessato, a una distribuzione per genere e Stato estero di cittadinanza della popolazione residente a fine periodo che riproduce nei totali per genere la popolazione calcolata alla medesima data.

Si rende opportuno tuttavia un ultimo passaggio di verifica, volto a controllare che effettivamente tutti i totali delle distribuzioni comunali per cittadinanza riproducano esattamente i totali calcolati con i bilanci e che non siano state introdotte frequenze negative. Occorre prestare attenzione ai casi per i quali si fossero eventualmente determinati incrementi/decrementi sotto la soglia di sicurezza ma comunque elevati, o "sbilanciamenti sospetti" nel rapporto tra le frequenze dei due sessi. Questi casi devono essere esaminati manualmente dall'operatore per valutare se sono dovuti ad effettive modificazioni nella struttura della popolazione (dovute alle iscrizioni/cancellazioni anagrafiche intervenute nel periodo) o se piuttosto, per ottenere l'allineamento, è il caso di ricorrere a una valutazione ed imputazione manuale delle frequenze anziché alla procedura automatica, in quanto quest'ultima non riesce a fornire una soluzione soddisfacente.

Sistemati gli ultimi casi residui, si possono considerare definitivamente validati i dati. Non resta altro da fare che produrre il file finale, rimettendo insieme i record relativi ai diversi gruppi di comuni: quelli con popolazioni calcolata e anagrafica che coincidevano sin dall'inizio, quelli che presentavano differenze in valore assoluto minori della soglia, quelli che presentavano grandi differenze. Il file è pronto per essere completato con la generazione dei codici dei raggruppamenti dei Paesi di cittadinanza per area geopolitica e

continente di appartenenza. I dati possono finalmente essere diffusi[29].

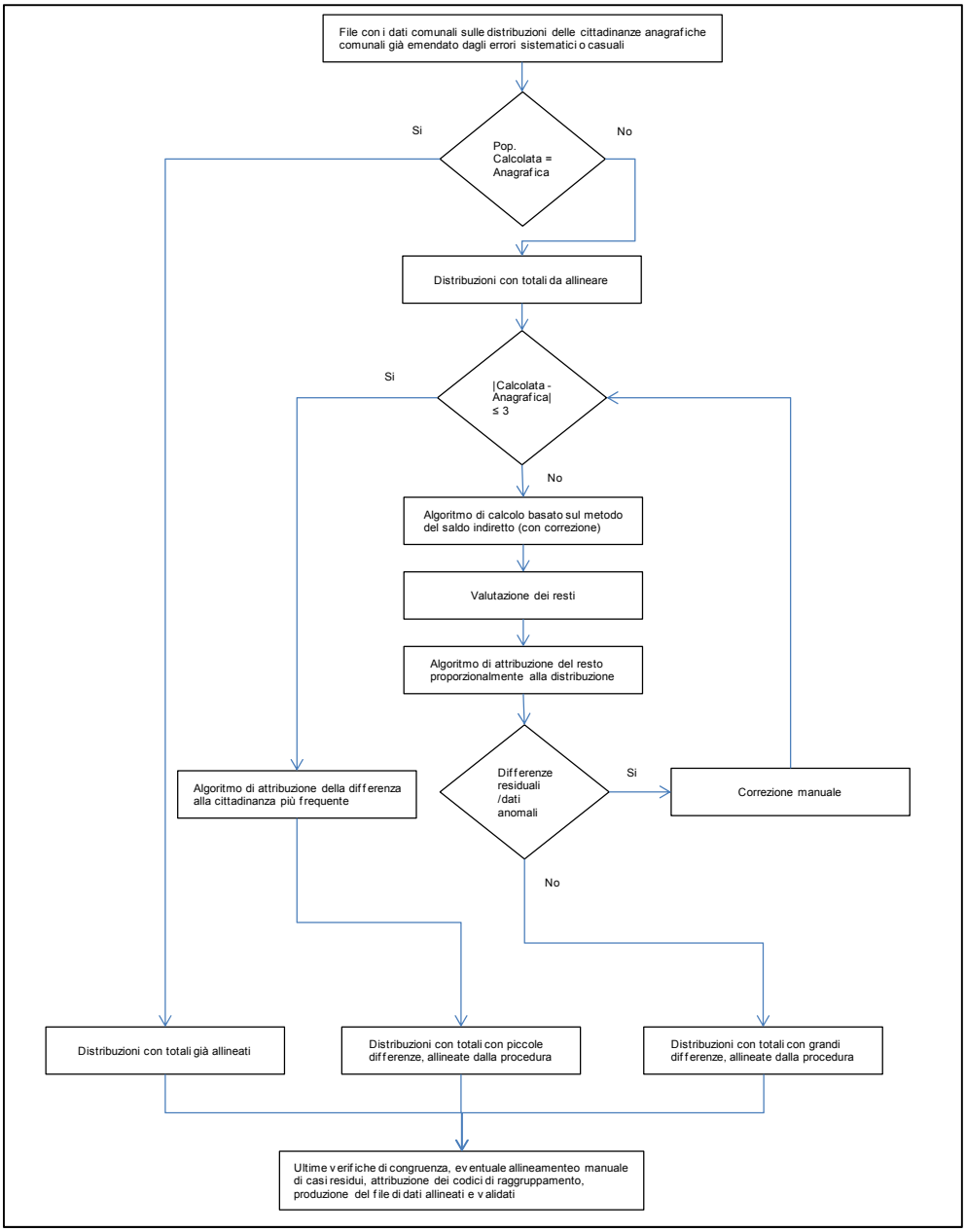
Nella Figura 1 si riporta la schematizzazione del processo appena descritto di allineamento ai totali per genere della popolazione calcolata dei dati delle distribuzioni per genere e Paese di cittadinanza della popolazione straniera iscritta nelle anagrafi comunali.

Occorre da ultimo precisare che, evidentemente, il metodo di stima proposto non è del tutto esente da possibili errori di sovra-sotto copertura dei contingenti di popolazione straniera residente distribuiti per Stato estero di cittadinanza, in particolare nei comuni più grandi. Considerate le informazioni a disposizione e i tempi ristretti per la pubblicazione dei risultati imposti dal Regolamento (CE) n. 862/2007 il metodo, negli anni in cui è stato applicato, è risultato tuttavia essere il più completo tra quelli applicabili: il più idoneo a garantire una transizione graduale dalla distribuzione della popolazione straniera per cittadinanza rilevata al censimento a quella risultante dal confronto censimento-anagrafe, aggiornata con i movimenti anagrafici registrati nei bilanci demografici annuali[30]. Il metodo adottato in produzione ha potuto tenere conto infatti, per definizione, tanto dei livelli di popolazione rilevati al censimento quanto di quelli registrati nelle anagrafi, nonché delle operazioni di revisione anagrafica conseguenti al confronto censimento/anagrafe (cfr. anche paragrafo 4).

[29] Si può ad esempio generare il file da fornire in input al data warehouse di diffusione dell'Istituto I.Stat (http://dati.istat.it/) o al sito tematico DEMO (http://demo.istat.it/), o preparare le tavole per le altre pubblicazioni correnti o per altri scopi di analisi o ricerca (Istat, 2011).

[30] Il Regolamento 862/2007 del Parlamento europeo e del Consiglio, dell'11 luglio 2007, relativo alle statistiche comunitarie in materia di migrazione e di protezione internazionale fissa norme comuni riguardo alla rilevazione di dati e alla compilazione di statistiche comunitarie in materia di immigrazione, emigrazione, protezione internazionale, residenza, immigrazione clandestina e rimpatri.

Figura 1 – La procedura per l'allineamento dei dati sulle cittadinanze della popolazione straniera anagrafica ai totali della popolazione calcolata



4. Valutazione del metodo dei saldi indiretti netti

In questo paragrafo verranno riassunte le principali caratteristiche del metodo dei saldi indiretti netti adottato in produzione. Si vuole mostrare che, con le fonti a disposizione, il metodo risulta il più accurato ai fini della stima delle distribuzioni comunali dei cittadini stranieri residenti per genere e cittadinanza negli anni immediatamente successivi al censimento.

Nel paragrafo verranno esposti anche i limiti del metodo, dovuti essenzialmente alla necessità di ricorrere ad una valutazione indiretta dei saldi, a causa della carenza di informazioni sui flussi di bilancio per genere e singola cittadinanza, stante l'attuale organizzazione del processo produttivo e considerate le fonti al momento disponibili.

A livello nazionale verrà infine effettuato il confronto dei risultati ottenuti in produzione con i risultati ottenibili applicando due possibili metodi alternativi: il metodo dei saldi indiretti lordi e il metodo del "riproporzionamento" semplice[31].

4.1 Punti di forza e punti di debolezza del metodo

Il metodo dei saldi indiretti netti ha consentito, negli anni di svolgimento della revisione anagrafica, la stima delle distribuzioni per genere e cittadinanza dei residenti stranieri a fine anno nei comuni italiani in assenza di informazioni sui flussi di bilancio comportanti conteggio distribuite secondo l'incrocio delle medesime variabili. Per la prima volta dopo il censimento del 2011 si è potuto procedere alla quantificazione del saldo dei movimenti non comportanti conteggio a livello di singolo comune e per genere e Paese di cittadinanza e quindi alla quantificazione del fattore di correzione (6). La quantificazione è stata ottenuta a partire dai dati individuali sulle iscrizioni e cancellazioni non comportanti conteggio rispetto al calcolo della popolazione operate dalle anagrafi con la revisione e registrate nel SIREA. Non è stato possibile fare altrettanto in occasione del censimento del 2001: il fattore di correzione non era disponibile secondo l'incrocio di variabili richiesto (comune, genere, cittadinanza), in quanto le informazioni sui movimenti non comportanti conteggio erano state rilevate solo a livello aggregato.

La possibilità di applicare il fattore di correzione (6) ha avuto un forte impatto sulla qualità dei risultati delle operazioni di stima in quanto, come si vedrà (cfr. anche paragrafo 2.6 e paragrafo 4.2), la distribuzione per genere e Paese di cittadinanza degli stranieri interessati da cancellazioni non comportanti conteggio è risultata diversa da quella dello stock di popolazione straniera residente distribuito secondo le medesime variabili.

Per garantire il rispetto delle tempistiche previste dal regolamento sulle statistiche comunitarie in materia di migrazione e di protezione internazionale, per la stima delle distribuzioni comunali della popolazione straniera residente distribuita per genere e cittadinanza è stato necessario ricorrere alle informazioni provenienti dai modelli di dati aggregati (che possono essere raccolti e lavorati con maggiore tempestività) e alla stima dei

[31] Si precisa che le distribuzioni a livello nazionale, derivanti dall'applicazione dei metodi a confronto, sono state elaborate per aggregazione di quelle stimate a livello comunale. Per tutte le procedure esposte nel seguito infatti l'allineamento al calcolo delle distribuzioni della popolazione straniera per genere e cittadinanza è stato originariamente effettuato a livello di singolo comune.

saldi indiretti netti.

Nel periodo successivo al XV Censimento della popolazione il metodo dei saldi indiretti netti ha garantito la transizione graduale dei dati statistici sulla popolazione straniera residente distribuita per genere e Stato estero di cittadinanza verso la reale distribuzione della presenza straniera regolare sul territorio risultante dal confronto tra la fonte censuaria e la fonte anagrafica, a valle delle operazioni di rettifica dei registri attuate dai comuni.

4.2 Confronto con due possibili metodi alternativi al metodo dei saldi indiretti netti

Date le fonti a disposizione, due possibili metodi alternativi a quello utilizzato in produzione sono il metodo dei saldi indiretti lordi e il metodo del riproporzionamento semplice. Il primo metodo alternativo ricalca in tutto e per tutto il metodo in produzione, a meno della correzione di cui nella (6). Si tratta quindi di applicare ai dati comunali la (5), non tenendo conto del fattore di correzione derivato dal SIREA. Lo definiremo nel prosieguo "metodo dei saldi indiretti lordi" o, più brevemente, "metodo dei saldi lordi".

Il secondo metodo alternativo consiste invece nel determinare, per ciascun comune, la distribuzione per cittadinanza degli stock per genere calcolati sulla base dei bilanci demografici ripartendo le differenze tra stock calcolati e anagrafici in base alla distribuzione per cittadinanza degli stock anagrafici. Cioè se, al tempo t_2 e nel comune c, $c_{t_2}^{p,g,c} = Pa_{t_2}^{p,g,c}/Pa_{t_2}^{g,c}$ è il rapporto tra il numero di stranieri iscritti in anagrafe di Paese di cittadinanza p e genere g ($Pa_{t_2}^{p,g,c}$) e il numero di tutti gli stranieri di genere g iscritti in anagrafe ($Pa_{t_2}^{g,c}$), il coefficiente $c_{t_2}^{p,g,c}$ può essere utilizzato per attribuire alla specifica cittadinanza p la quota di scarto tra popolazione calcolata e anagrafica che le compete, ossia:

$$Pc_{t_2}^{p,g,c} = Pa_{t_2}^{p,g,c} + c_{t_2}^{p,g,c} * (Pc_{t_2}^{g,c} - Pa_{t_2}^{g,c}) \quad (8)$$

In cui $Pc_{t_2}^{p,g,c}$ è la popolazione calcolata al tempo t_2 per il Paese di cittadinanza p e genere g nel comune c, mentre $Pa_{t_2}^{p,g,c}$, come già indicato, è la popolazione iscritta in anagrafe secondo le medesime caratteristiche. Il risultato del prodotto di cui nel secondo addendo della (8) viene arrotondato all'unità, per evitare frequenze decimali. Effettuato il calcolo (8) per le diverse cittadinanze presenti nel comune, si procede alla rideterminazione dei totali per genere e si attribuiscono eventuali differenze residue (positive o negative) alla cittadinanza più frequente. Per comodità di lettura il metodo appena descritto sarà denominato "metodo del riproporzionamento".

I due metodi alternativi sono stati applicati alla stima della popolazione residente straniera comunale, distribuita per genere e Paese di cittadinanza, al 31 dicembre 2012 e al 31 dicembre 2013. Si vuole mostrare che il metodo dei saldi netti, rispetto agli altri due considerati, è preferibile in quanto garantisce una transizione graduale dalla distribuzione rilevata al censimento verso la distribuzione risultante dal confronto tra fonte censuaria e anagrafica, al termine delle operazioni di revisione anagrafica. Ciò equivale a dire che il metodo si presta meglio di altri ad accompagnare il delicato passaggio tra il risultato del censimento e i dati di popolazione derivanti dal calcolo negli anni immediatamente successivi, tenuto conto delle operazioni di revisione delle anagrafi.

Per ragioni di semplicità, la valutazione è stata effettuata sulle distribuzioni nazionali: è probabile che a livello territoriale disaggregato le differenze messe in luce tra i tre metodi risultino in taluni casi anche più ampie.

Per comparare i tre metodi (metodo adottato in produzione, metodo dei saldi lordi e metodo del riproporzionamento) si è esaminato il grado di correlazione esistente tra le distribuzioni ottenute con l'applicazione di ciascuno di essi ai dati al 31 dicembre 2012 e 2013. Queste distribuzioni sono state confrontate tra loro, con la distribuzione al censimento, con quella risultante nelle anagrafi comunali alla data del censimento, al 31 dicembre 2012 e al 31 dicembre 2013. Per le $\binom{10}{2}$ = 45 diverse coppie di distribuzioni è possibile misurare la correlazione binaria con il coefficiente di correlazione di Bravais[32]. Nella Tavola 4 si riportano i valori del coefficiente, per le possibili abbinate. E' possibile osservare che le distribuzioni sono tutte molto positivamente e significativamente correlate fra loro (valori del coefficiente pari a uno indicano massima correlazione positiva)[33]. Si nota tuttavia che la distribuzione anagrafica è sin dall'inizio (alla data del censimento) relativamente poco correlata con quella censuaria e la correlazione, nel tempo (alle altre date di riferimento considerate – 31 dicembre 2012 e 2013), si va man mano riducendo.

Tavola 4 - Matrice di correlazione tra le serie di Paesi per numero di cittadini residenti in Italia

	Censimento	Popolazione anagrafica (9 ott 2011)	Metodo dei saldi netti (31 dic 2012)	Metodo dei saldi lordi (31 dic 2012)	Metodo del riproporzionamento (31 dic 2012)	Popolazione anagrafica (31 dic 2012)	Metodo dei saldi netti (31 dic 2013)	Metodo dei saldi lordi (31 dic 2013)	Metodo riproporzionamento (31 dic 2013)	Popolazione anagrafica (31 dic 2013)
Censimento	**1,000**	0,998	0,999	0,998	0,998	0,997	0,997	0,997	0,996	0,996
Popolazione anagrafica (9 ott 2011)		**1,000**	0,999	0,999	1,000	1,000	0,999	0,999	0,999	0,999
Metodo dei saldi netti (31 dic 2012)			**1,000**	1,000	0,999	0,999	0,999	0,999	0,999	0,999
Metodo dei saldi lordi (31 dic 2012)				**1,000**	0,999	0,999	1,000	1,000	0,999	0,999
Metodo del riproporzionamento (31 dic 2012)					**1,000**	1,000	0,999	0,999	1,000	0,999
Popolazione anagrafica (31 dic 2012)						**1,000**	1,000	0,999	1,000	1,000
Metodo dei saldi netti (31 dic 2013)							**1,000**	1,000	1,000	1,000
Metodo dei saldi lordi (31 dic 2013)								**1,000**	1,000	0,999
Metodo del riproporzionamento (31 dic 2013)									**1,000**	1,000
Popolazione anagrafica (31 dic 2013)										**1,000**

Fonte: elaborazione su dati Istat (Censimento della popolazione 2011, Rilevazione sul Movimento e calcolo della popolazione straniera residente e struttura per cittadinanza)

[32] Le combinazioni senza ripetizione di dieci elementi presi due a due sono 10!/8!(10-8)!=45. Si tratta di combinazioni senza ripetizione in quanto come è noto la correlazione è un concetto simmetrico.

[33] Per le serie che fanno riferimento a momenti diversi ovviamente il grado di correlazione è influenzato anche dalla data di riferimento, oltre che dal metodo/fonte cui si riferiscono i dati.

Al 9 ottobre 2011 i dati ricavati con il metodo in produzione (metodo dei saldi netti) presentano lo stesso grado di correlazione con i dati censuari e con la corrispondente popolazione anagrafica. Nel 2012 il metodo in produzione è quello che presenta la correlazione maggiore con il censimento e nel 2013 conserva una correlazione con il censimento superiore rispetto al metodo del riproporzionamento. Sempre nel 2012 il metodo in produzione presenta uguale livello di correlazione con il censimento e con l'anagrafe, mentre nel 2013 la correlazione con l'anagrafe diviene superiore rispetto a quella con il censimento. Già nel 2012 e anche nel 2013 sia il metodo dei saldi lordi che, in misura ancora più evidente, il metodo del riproporzionamento sono più correlati con il corrispondente dato anagrafico che con il censimento.

Ai fini del confronto tra le distribuzioni esaminate può essere interessante comparare le corrispondenti graduatorie dei Paesi per numero di cittadini residenti in Italia. Per la comparazione si può ricorrere al confronto tra i ranghi dei Paesi: una misura sintetica della differenza tra graduatorie è il numero complessivo di "salti" tra una graduatoria e l'altra, ovvero la somma delle differenze tra i ranghi che competono a ciascun Paese nelle diverse coppie di graduatorie. E' evidente che quanto più elevata è questa somma per la coppia di graduatorie presa in considerazione, tanto più discordanti sono le graduatorie che la compongono. Nella tavola 5 si riporta il valore delle somme delle differenze tra ranghi per le coppie di graduatorie esaminate. Dalla tavola si evince che, sia nel 2012 che nel 2013, dei tre metodi a confronto il metodo in produzione è quello che garantisce meno salti in graduatoria rispetto al censimento.

Con riferimento alla distribuzione anagrafica, sia nel 2012 che nel 2013 il metodo garantisce un numero di salti inferiore al metodo dei saldi lordi, ma superiore al metodo del semplice riproporzionamento. Del resto quest'ultimo è il metodo che, per come concepito, maggiormente privilegia la distribuzione anagrafica. Nel 2012 il metodo di produzione si avvicina di più al metodo dei saldi lordi che al metodo del riproporzionamento. Nel 2013 (a revisione anagrafica pressoché ultimata) esso si avvicina invece di più al metodo del riproporzionamento che al metodo dei saldi lordi. E' molto vicino anche alla distribuzione anagrafica: nel 2013 molti comuni avevano portato a termine o effettuato gran parte della revisione della propria anagrafe.

Tavola 5 - Somma dei "salti" *(cambiamenti di posizione)* **nelle graduatorie dei Paesi per numero di cittadini stranieri residenti in Italia**

	Censimento	Popolazione anagrafica (9 ott 2011)	Metodo dei saldi netti (31 dic 2012)	Metodo dei saldi lordi (31 dic 2012)	Metodo del riproporzionamento (31 dic 2012)	Popolazione anagrafica (31 dic 2012)	Metodo dei saldi netti (31 dic 2013)	Metodo dei saldi lordi (31 dic 2013)	Metodo riproporzionamento (31 dic 2013)	Popolazione anagrafica (31 dic 2013)
Censimento	**0**	903	704	786	853	907	863	1062	917	921
Popolazione anagrafica (9 ott 2011)		**0**	715	773	424	377	717	1047	615	602
Metodo dei saldi netti (31 dic 2012)			**0**	176	457	532	345	592	431	438
Metodo dei saldi lordi (31 dic 2012)				**0**	477	549	309	578	385	402
Metodo del riproporzionamento (31 dic 2012)					**0**	179	470	853	383	395
Popolazione anagrafica (31 dic 2012)						**0**	488	889	395	387
Metodo dei saldi netti (31 dic 2013)							**0**	549	238	236
Metodo dei saldi lordi (31 dic 2013)								**0**	680	677
Metodo del riproporzionamento (31 dic 2013)									**0**	37
Popolazione anagrafica (31 dic 2013)										**0**

Fonte: elaborazione su dati Istat (Censimento della popolazione 2011, Rilevazione sul Movimento e calcolo della popolazione straniera residente e struttura per cittadinanza)

Nella tavola 6 si riportano le posizioni in graduatoria dei primi venticinque Paesi per numero decrescente di cittadini stranieri residenti al 31 dicembre 2013, secondo i tre metodi di allineamento alla popolazione calcolata messi a confronto. Come si osserva, le graduatorie sono identiche fino all'ottava posizione. Dal punto di vista dell'ordine per importanza numerica i tre metodi non determinano cambiamenti per le prime otto cittadinanze, che secondo il metodo in produzione annoverano quasi il 70% del totale degli stranieri residenti nel Paese alla fine del 2013. Differenze nelle graduatorie si hanno invece a partire dalla nona posizione, anche se spesso (ma non sempre) si limitano allo scambio di posizioni adiacenti, almeno con riferimento ai primi venticinque posti.

Ad una certa stabilità nelle graduatorie per le cittadinanze maggiormente rappresentate si accompagnano tuttavia stime dei contingenti relativi a ciascuna cittadinanza, ottenute con i diversi metodi, in alcuni casi anche numericamente piuttosto diverse.

Se si considerano i primi otto Paesi, il totale degli stranieri residenti in Italia calcolato con il metodo adottato in produzione (4.268.966) risulta inferiore all'analogo totale calcolato con il metodo dei saldi lordi (4.304.356) e superiore a quello ottenuto con il metodo del riproporzionamento semplice (4.263.342). Poiché non tiene conto delle

viene ripartita proporzionalmente in base all'importanza numerica della cittadinanza: questa ripartizione tende a privilegiare le nazionalità più numerose, che non sempre coincidono con le nazionalità che sono risultate maggiormente soggette (in termini relativi, ossia rispetto alla loro numerosità) a cancellazioni per irreperibilità al censimento e alle altre cancellazioni non comportanti conteggio. Questa è la ragione per cui collettività molto numerose con il metodo dei saldi lordi possono risultare sovrastimate. Si spiega in questo modo anche la differenza nei totali parziali di cui sopra.

Si consideri ad esempio il caso della comunità Filippina che con poco più di 5.500 posizioni cancellate dall'anagrafe si piazza solo al 15° posto per numero di cancellazioni per irreperibilità al censimento effettuate dai comuni italiani nel 2013 (anno in cui da parte di molti comuni è stata effettuata gran parte della revisione anagrafica): nella graduatoria per numerosità di residenti in Italia all'inizio dello stesso anno essa è al 4° posto, con oltre 140 mila residenti[34]. Analoga la condizione dei cittadini indiani (rispettivamente al 14° e all'8° posto nelle due graduatorie). Similmente, la Moldova conta al 1° gennaio 2013 quasi 140 mila residenti, piazzandosi al 7° posto nella graduatoria delle cittadinanze, mentre risulta solo al 24° posto per numero di cancellazioni per irreperibilità al censimento nel 2013 (meno di 3.700).

[34] Non è l'oggetto di questo lavoro, ma le motivazioni possono essere molteplici. Si può ipotizzare una diversa propensione, da parte delle diverse collettività, a farsi censire e/o a cancellarsi dall'anagrafe in caso di emigrazione all'estero. E' plausibile ad esempio che la collettività filippina, più di altre, sia caratterizzata da una certa propensione a farsi censire in quanto i suoi componenti, spesso regolarmente impiegati in attività come i servizi alle famiglie, tendono ad essere facilmente reperibili. Se si considera una misura grezza della propensione, quale il tasso di cancellazione non comportante conteggio per mille residenti, questo assume valori a volte anche molto differenti a seconda della cittadinanza.

Tavola 6 - Popolazione straniera residente in Italia secondo i tre metodi a confronto, al 31 dicembre 2013 *(primi venticinque Paesi di cittadinanza con il metodo in produzione)*

Paese di cittadinanza	Metodo dei saldi netti		Metodo dei saldi lordi		Metodo del riproporzionamento	
	Stranieri residenti	Rango	Stranieri residenti	Rango	Stranieri residenti	Rango
Romania	1.081.400	1	1.082.875	1	1.091.348	1
Albania	495.709	2	505.602	2	489.846	2
Marocco	454.773	3	457.604	3	458.550	3
Cinese, Repubblica Popolare	256.846	4	262.081	4	247.531	4
Ucraina	219.050	5	221.939	5	219.643	5
Filippine	162.655	6	174.497	6	156.000	6
Moldova	149.434	7	152.688	7	146.644	7
India	142.453	8	142.149	8	140.665	8
Bangladesh	111.223	9	118.450	9	104.265	10
Perù	109.851	10	115.414	10	107.267	9
Polonia	97.566	11	92.262	15	101.461	12
Tunisia	97.317	12	93.411	13	101.226	13
Egitto	96.008	13	92.479	14	102.159	11
Sri Lanka (ex Ceylon)	95.007	14	95.950	11	95.860	14
Ecuador	91.861	15	93.682	12	91.430	16
Senegal	90.863	16	91.097	17	92.137	15
Pakistan	90.615	17	91.354	16	90.905	17
Macedonia, Repubblica di	78.424	18	79.395	18	80.232	18
Nigeria	66.833	19	66.811	19	67.333	19
Bulgaria	54.932	20	52.330	20	56.450	20
Ghana	51.602	21	52.090	21	51.039	21
Serbia, Repubblica di	46.958	22	48.367	22	46.040	22
Kosovo	46.248	23	47.585	23	43.341	24
Brasile	43.202	24	40.175	24	43.878	23
Germania	38.136	25	34.069	26	38.092	25
Totale 25 Paesi	**4.268.966**		**4.304.356**		**4.263.342**	
Totale	**4.922.085**		**4.922.085**		**4.922.085**	

Fonte: elaborazione su dati Istat (Rilevazione sul Movimento e calcolo della popolazione straniera residente e struttura per cittadinanza)

La distribuzione dei cancellati per irreperibilità al censimento secondo la cittadinanza non ricalca fedelmente la distribuzione della popolazione residente: utilizzare il metodo dei saldi indiretti lordi che non tiene conto delle suddette cancellazioni nel colmare gli scarti tra popolazione anagrafica e calcolata, comporta il rischio di distorsioni nelle distribuzioni finali.

Un'ulteriore conferma di quanto appena detto si ha se si considera l'altro metodo messo a confronto: il metodo del riproporzionamento semplice. Questo metodo in effetti è quello che produce per le prime otto cittadinanze il totale più basso. Il metodo, non tenendo conto di alcun tipo di saldo, neppure lordo, tende a colmare l'intera differenza tra calcolo e anagrafe strettamente in base alle proporzioni tra cittadinanze negli stock di popolazione comunali: rispetto ai precedenti due metodi tende quindi a "spalmare" maggiormente la differenza sull'intero ventaglio delle cittadinanze (anche quelle meno numerose).

In definitiva, il metodo adottato in produzione, tenendo conto dei reali incrementi

netti di numerosità per le differenti nazionalità registrati nel periodo, negli anni di svolgimento della revisione anagrafica è risultato quello che ha consentito l'allineamento al calcolo della distribuzione anagrafica della popolazione straniera residente in grado di fornire la stima più precisa della reale consistenza delle diverse collettività residenti sul territorio a fine anno.

5. Conclusioni

Negli anni immediatamente successivi al XV Censimento generale della popolazione e delle abitazioni il metodo dei saldi indiretti netti è stato applicato ai fini dell'allineamento dei dati anagrafici comunali alla popolazione calcolata con i bilanci demografici (a partire dalla popolazione censita) e quindi ai fini della stima della popolazione straniera residente in ciascun comune italiano, distribuita per genere e cittadinanza. Date le fonti e le tempistiche per la diffusione dei dati dettate dal Regolamento (CE) 862/2007, al momento in cui si è operato (rilevazioni degli anni 2011, 2012 e 2013) vi erano altri due metodi alternativi possibili. In questo lavoro si è mostrato come la stima sia risultata non indifferente all'applicazione dell'uno o dell'altro metodo, in ragione del fatto che la distribuzione delle cancellazioni per irreperibilità al censimento è risultata molto diversa dalla distribuzione degli stock di popolazione straniera residente a fine anno, per genere e cittadinanza.

In questo quadro, il metodo dei saldi netti ha garantito una transizione graduale dalla distribuzione degli stranieri censiti verso la distribuzione della popolazione straniera risultante dal confronto tra la fonte censuaria e quella anagrafica. In seguito alla revisione delle anagrafi tale popolazione dovrebbe "fotografare" al meglio la reale presenza straniera regolare sul territorio del Paese.

Naturalmente, qualora si fosse potuto disporre in tempo utile dei flussi validati individuali sulle iscrizioni e cancellazioni dalle anagrafi comportanti conteggio sarebbe stato possibile effettuare un calcolo diretto. In concomitanza con la prossima implementazione dell'Anagrafe Nazionale della Popolazione Residente[35] sono in fase di progettazione sistemi di rilevazione dei flussi anagrafici individuali, organizzati in una base statistica di dati individuali relativi al movimento e allo stock di popolazione, l'Anagrafe Virtuale Statistica ANVIS (Gazzelloni, 2013). Grandi trasformazioni sono in cantiere anche per quanto riguarda l'organizzazione dei futuri censimenti della popolazione, indirizzata verso l'attuazione di un "censimento permanente" (Istat, 2014). Il quadro complessivo delle statistiche demografiche risulta quindi in questa fase in rapido e profondo mutamento.

La validità del metodo di stima adottato in produzione per la determinazione della popolazione straniera residente comunale, distribuita per genere e Paese di cittadinanza, negli anni immediatamente seguenti l'anno di svolgimento del XV Censimento generale

[35] Decreto-legge 18 ottobre 2012, n. 179 "Ulteriori misure urgenti per la crescita del Paese", convertito, con modificazioni, dalla legge 17 dicembre 2012, n. 221, che nella Sezione I "Agenda e identità digitale", al comma 1 dell'articolo 2, sostituisce integralmente l'articolo 62 del Codice dell'Amministrazione Digitale (CAD), di cui al decreto legislativo 7 marzo 2005, n. 82. La nuova formulazione del citato articolo 62, composta di sei commi, prevede e disciplina l'istituzione, presso il Ministero dell'Interno, di una nuova base di dati, denominata "Anagrafe nazionale della popolazione residente" (ANPR), compresa tra quelle di interesse nazionale, individuate nell'articolo 60 del CAD.

della popolazione e delle abitazioni è del resto strettamente legata al contesto in cui si sono svolti il censimento e le successive operazioni di revisione anagrafica. Nel momento in cui saranno del tutto delineati gli scenari futuri delle statistiche demografiche, attualmente ancora in fase di sperimentazione, sarà possibile eventualmente definire nei dettagli in quale veste rinnovata il metodo potrà essere applicato anche in futuro[36].

[36] Ad esempio, sulla base delle informazioni al momento disponibili (Istat, 2014) il calcolo della popolazione ottenuto applicando i dati di flusso demografico (naturale, migratorio e per altri motivi) alla popolazione legale (censita nel 2011) dovrebbe proseguire nel quinquennio 2016-2020, primo quinquennio di esecuzione dell'indagine C-sample, l'indagine campionaria per il calcolo della popolazione nel quadro del censimento permanente. La popolazione calcolata continuerebbe a costituire la popolazione ufficiale dei comuni. La strategia del censimento permanente ai fini del calcolo della popolazione prevede in effetti a partire dal 2016 l'effettuazione di un test statistico (test d'ipotesi) basato sui risultati della C-Sample per la validazione della popolazione anagrafica comunale. Il test per ciascun comune dovrebbe verificare l'ipotesi nulla che la differenza tra popolazione anagrafica e popolazione calcolata con la C-sample sia pari a zero, fissata una certa probabilità di errore di prima specie *alpha*. Per i comuni per i quali la statistica test porterà ad accettare l'ipotesi nulla (popolazione anagrafica "validata" dal test statistico) si considererà come popolazione ufficiale la popolazione anagrafica. Per i comuni per i quali il valore calcolato per la statistica test porterà a rigettare l'ipotesi nulla (popolazione anagrafica "non validata" dal test statistico), l'Istat dovrebbe determinare con il contributo di altre fonti amministrative – essenzialmente il Sistema Integrato di Micro dati SIM (Istat, 2014) - un fattore correttivo per la popolazione anagrafica distribuita per sesso, età e cittadinanza italiana/straniera. Come accennato sopra, tuttavia, la correzione dovrebbe essere applicata alle anagrafi solo "a regime", a partire dal 2021. A far capo da tale data infatti si prevede che siano stati sondati almeno una volta tutti i comuni sotto i 50 mila abitanti. Pur non dando luogo a correzione, i risultati del test verrebbero notificati ai comuni sin dal 2016, per indurre le anagrafi ad effettuare le necessarie verifiche sui propri archivi, con l'obiettivo di migliorarne la qualità. A partire dal 2021 la popolazione censuaria dei comuni che non ottenessero la validazione del test statistico verrebbe invece corretta per un fattore di stima basato sui risultati della C-Sample e sulle informazioni (segnali di presenza/assenza) desumibili dal SIM. Alle anagrafi verrebbe in ogni caso comunicata la correzione, per i necessari controlli anagrafici, da effettuarsi nell'anno successivo all'effettuazione della C-Sample.
In definitiva il metodo proposto nell'articolo per la determinazione delle distribuzioni comunali della popolazione straniera per genere e Stato estero di cittadinanza potrebbe restare operativo tanto nella fase di transizione, in cui il calcolo della popolazione continuerebbe ad essere effettuato con il metodo tradizionale, quanto "a regime", a partire dal 2021. Resta a questo punto aperto il problema della stima della popolazione straniera per genere e singolo Paese di cittadinanza per i comuni con differenza significativa tra popolazione anagrafica e conteggio derivante dal calcolo (per la prima fase) o da C-Sample/SIM (a regime, a partire dal 2021). Stima che a livello aggregato potrebbe essere effettuata utilizzando il metodo descritto in questo lavoro, tenuto conto che le rettifiche anagrafiche potrebbero assumere anch'esse, come il censimento, natura permanente. Se disponibili in tempo utile, si potrebbero in alternativa utilizzare per l'aggiornamento i saldi diretti netti derivanti dai flussi individuali rilevati per ANVIS. Questo secondo approccio risentirebbe tuttavia dei tempi di validazione dei flussi di ANVIS.
Occorre in ogni caso tenere presente che non è escluso che il sopra citato quadro di riferimento si modifichi nel prossimo futuro ulteriormente, in modo più o meno sostanziale in quanto la strategia del censimento permanente, ad oggi, non è ancora completamente definita.

Appendice

La codifica della variabile cittadinanza

Come già accennato, la variabile cittadinanza dello straniero è una variabile fondamentale per l'analisi della presenza straniera in Italia. L'analisi per cittadinanza di origine fornisce informazioni importanti sul fenomeno migratorio: informazioni indispensabili per una migliore comprensione e gestione dello stesso. La codifica/decodifica della variabile, ai fini della rilevazione, del trattamento statistico, della diffusione e dell'interpretazione dell'informazione con essa espressa, rappresenta un passaggio molto delicato e importante nel processo di produzione dei dati sulla popolazione straniera residente.

La classificazione Istat degli Stati esteri di cittadinanza

L'Istat aggiorna annualmente la classificazione degli Stati esteri, utilizzabile per la codifica della variabile cittadinanza[1]. La classificazione attribuisce un codice a ciascuno degli Stati totalmente indipendenti e sovrani, riconosciuti a livello internazionale e/o dall'Italia. Non include invece i territori che, possedendo titolo e grado di autonomia o di sovranità nullo o al massimo parziale, non siano riconosciuti come totalmente indipendenti da un altro Stato sovrano. Ad esempio non comprende tutti gli ex possedimenti di Paesi con storia coloniale quali la Francia, il Regno Unito, i Paesi Bassi; non comprende i territori sotto l'influenza degli USA, della Finlandia, della Danimarca, ecc.[2].

Complessivamente ad oggi la classificazione Istat degli Stati esteri conta circa duecento voci relative ad altrettanti Stati (con le eccezioni evidenziate), raggruppate in aree

[1] La versione aggiornata della classificazione e le variazioni in essa intervenute vengono diffuse annualmente sul sito dell'Istat, all'indirizzo http://www.istat.it/it/archivio/6747, oltre che attraverso il Sistema delle Classificazioni dell'Istat (http://www.istat.it/it/strumenti/definizioni-e-classificazioni). L'ultimo aggiornamento risale al 31 dicembre 2014.

[2] La classificazione, nata ai fini della codifica della variabile cittadinanza, include la modalità "Apolide". Gli apolidi sono coloro che non sono riconosciuti cittadini di alcuno Stato. E' possibile essere apolidi per origine (non avendo mai posseduto una cittadinanza) o per derivazione (in seguito alla perdita di una pregressa cittadinanza e alla mancanza della contestuale acquisizione di una nuova cittadinanza). La classificazione include poi alcune casistiche particolari come quella dei "Riconosciuti non cittadini lettoni", un territorio storicamente oggetto di contestazione a livello geografico-politico (i Territori dell'Autonomia Palestinese), altri territori non del tutto autonomi (Isola di Man, Isole Antille Olandesi e Isole Jersey) ai fini dell'adeguamento della classificazione alle direttive di Eurostat. La classificazione è utilizzata nel modello elettronico Istat P.2&P.3 per la raccolta delle informazioni sulla cittadinanza degli stranieri residenti. Accedendo al modello elettronico, nella seconda sezione i rispondenti (gli operatori comunali) possono selezionare i singoli Stati di cittadinanza da un apposito menù a tendina indicando contestualmente, accanto alle voci selezionate, le corrispondenti frequenze assolute (distinte per genere) dei cittadini stranieri residenti nel comune. I dati possono anche essere caricati da file esterno. In questo caso il controllo sulla codifica delle cittadinanze si attiva in modalità batch al termine del caricamento. Il caricamento viene bloccato in caso di codici non riconosciuti per i quali viene prodotta una segnalazione di errore.
E' allo studio una classificazione storica degli Stati esteri esistiti dal 1918 ad oggi e una classificazione organica dei principali territori di interesse per fini statistici che, pur non trovando applicazione per la codifica delle cittadinanze, possono rivelarsi utili per la codifica di altre variabili quali ad esempio il luogo di nascita. Il luogo di nascita non viene rilevato con l'indagine sul "Movimento e calcolo della popolazione residente straniera e struttura per cittadinanza", ma l'informazione sul luogo di nascita viene invece raccolta da altre rilevazioni di fonte amministrativa sulla popolazione straniera, come ad esempio le LAC.

geopolitiche di appartenenza (ad esempio “Unione europea”, o “Africa settentrionale”, ecc.) e per continente.

La classificazione dei Paesi di cittadinanza del XV Censimento della popolazione e delle abitazioni

La classificazione adottata per la codifica della variabile cittadinanza in occasione del XV Censimento generale della popolazione del 2011 è molto simile alla classificazione Istat degli Stati esteri valida alla fine dello stesso anno. Le voci differenti sono pochissime e inoltre per ciascuna di esse è stata garantita la riconducibilità con le voci dell’altra classificazione[3]. La classificazione adottata per il censimento della popolazione è stata stabilita da un’apposita direttiva di Eurostat, la Direttiva 1201 del 2009[4]. Essa è stata utilizzata per la codifica delle risposte relative al quesito “Paese di cittadinanza” contenuto nelle schede individuali del Foglio di famiglia (Modello Istat CP.1) e per la pubblicazione dei relativi risultati (Albani et al., 2013).

La corrispondenza e la riconducibilità delle voci eventualmente non corrispondenti tra le classificazioni si è rivelata una caratteristica essenziale ai fini della validazione dei dati sugli stranieri residenti nei comuni italiani derivanti dal modello Istat P.2&P.3 negli anni successivi al censimento.

[3] Un esempio delle differenze è rappresentato dal Kosovo, che per l’Italia dal 17 febbraio 2008 costituisce uno stato indipendente. Essendo lo Stato riconosciuto, al 29 ottobre 2013, solo da 108 dei 193 Stati membri dell’ONU tuttavia, la classificazione adottata per il XV Censimento della popolazione non ha previsto questa voce nella classificazione per la codifica della variabile cittadinanza. I cittadini kosovari sono stati considerati dal censimento come cittadini Serbi.

[4] Commission Regulation (EC) No 1201/2009 of 30 November 2009 implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses as regards the technical specifications of the topics and of their breakdowns

Riferimenti bibliografici

Albani M. *Dalle classificazioni ai dizionari implementati: Stato estero* in Il trattamento delle variabili testuali nel 15° Censimento generale della popolazione *(*a cura di Macchia S. e Mastroluca S.). Roma- Istat, Working Papers n.3/2013.

Albani M. *La popolazione residente straniera. Nazionalità, territorio e bilancio demografico* in Dossier Statistico Immigrazione 2014, a cura del Centro Studi e Ricerche IDOS/Immigrazione Dossier Statistico. Roma, ottobre 2014

Albani M., Brandimarti P. *La ristrutturazione dell'indagine sulla popolazione straniera residente*. Roma: Istat (Working Papers n.8/2012)

Albani M., Conti C., Guarneri A. *Cittadinanza e territorio: un'analisi spaziale della presenza straniera in Italia*, Rivista Italiana di Economia, Demografia e Statistica, Volume LXIV – N.4, Ottobre-Dicembre 2010.

Albani M., Gualtieri M., Guarneri A. *Foreign resident population in Italy: a local labour market areas approach* Paper presentato alla European Population Conference 2008 – Barcellona 9-12 luglio 2008.

Calzola L. *La valutazione della copertura del censimento della popolazione sulla base delle rettifiche anagrafiche*. Paper presentato alle Giornate di studio sulla popolazione, Padova, 16-18 febbraio 2005.

Commissione per la Garanzia dell'Informazione Statistica. *Il confronto tra censimento e anagrafe: per un maggior grado di coerenza tra le due fonti*, Rapporto di ricerca CGIS, n. 99.10, Roma Luglio 1999.

Cortese A. *Censimento ed archivi amministrativi: un rapporto da ridefinire*, Relazione presentata alla Prima Sessione "Censimento della popolazione: il contesto internazionale e l'esperienza italiana" della Conferenza Istat "Censimenti generali 2010-2011. Criticità e innovazioni", Roma 21-22 novembre, 2007.

Cortese A., Gallo G. e Paluzzi E. *Il censimento della popolazione straniera: opinioni a confronto sul principale aspetto definitorio*. Roma – Istat, Contributi n.1/2010)

Cortese A. *Il concetto di "dimora abituale" e l'accertamento statistico della popolazione residente*, I Servizi Demografici, N. 5, 2008a.

Cortese A. *La conta degli stranieri: una bella sfida per il censimento demografico del 2011*, Contributo N. 5 alla voce "Pubblicazioni scientifiche" sul sito ufficiale dell'Istat, 2008b.

Dardanelli S., Mastroluca S., Sasso A. e Verrascina M. *La progettazione dei censimenti generali 2010 – 2011: Novità di regolamentazione internazionale per il 15° Censimento generale della popolazione e delle abitazioni*. Roma - Istat, Documenti n.1/2009

Crescenzi F., Fortini M., Gallo G. e Mancini A. *La progettazione dei censimenti generali 2010 – 2011: Linee generali di impostazione metodologica, tecnica e organizzativa del 15° Censimento generale della popolazione.* Roma – Istat, Documenti n.6/2009

Data warehouse DEMO (http://demo.istat.it/), Cittadini stranieri/Bilancio demografico – Anno 2014

Data warehouse I.Stat (http://dati.istat.it/), Popolazione e famiglie/Stranieri e immigrati/Popolazione straniera residente al 1° gennaio – focus sulla cittadinanza – Anno 2015

Fortini M., Gallo G., Paluzzi E., Reale A. e Silvestrini A. *La progettazione dei censimenti generali 2010–201: Criticità di processo e di prodotto nel 14° Censimento generale della popolazione e delle abitazioni: aspetti rilevanti per la progettazione del 15° Censimento*. Roma – Istat, Documenti n. 10/2007.

Gallo G., Paluzzi E., Silvestrini A. e Cortese P.F. *Il confronto tra anagrafe e censimento 2001 nel Comune di Roma*. Roma – Istat, Documenti n.6/2010.

Gazzelloni S., *L'Anagrafe Virtuale Statistica - Banca dati di interesse nazionale.* Relazione presentata al Convegno USCI, Messina - 27 settembre 2013

Gesano G., F. Heins, F. Paganelli. *Differenze anagrafe-censimento: verifica di alcune motivazioni politicoamministrative*, relazione presentata alle Giornate di Studio sulla Popolazione. Bologna, 6-7 dicembre 1993.

Istat. *Anagrafe della popolazione*. Metodi e norme, serie B – n. 29. Edizione 1992. Roma, 1992.

Istat. *Bilancio demografico nazionale - Popolazione residente in totale e straniera, natalità, mortalità, migrazioni, famiglie e convivenze - Anno 2013.* Comunicato stampa Istat - Roma, 16 giugno 2014 (http://www.istat.it/it/archivio/125731)

Istat. *Classificazione degli Stati esteri al 31 dicembre 2013* (http://www.istat.it/it/archivio/6747). Roma, 2014

Istat. *Conoscere il censimento: i documenti, 14° Censimento generale della popolazione e delle abitazioni*. Roma, 2006.

Istat. *Disposizioni per gli Organi periferici e Istruzioni per il rilevatore, 14° Censimento generale della popolazione e 8° Censimento generale dell'industria e dei servizi.* Roma, 2001.

Istat. *Il Piano di rilevazione e il Sistema di produzione. 14° Censimento generale della popolazione e delle abitazioni, Conoscere il censimento*. Roma, 2006.

Istat. *Indagine di copertura del 15° Censimento generale della popolazione e delle abitazioni.* Istat Nota informativa – Roma, 14 gennaio 2015

Istat. *La popolazione straniera residente in Italia al 1° gennaio 2011*. Roma, 22 settembre 2011 (http://www.istat.it/it/archivio/39726)

Istat. *Linee strategiche del censimento permanente della popolazione e delle abitazioni – Metodi, tecniche e organizzazione*. Roma, Istat Metodi – Letture statistiche, 2014

Istat. *Primi risultati del 15° Censimento della popolazione e delle abitazioni* – Periodo di riferimento: 9 ottobre 2011. Comunicato stampa Istat – Roma, 19 dicembre 2012 (http://www.istat.it/it/archivio/77877)

Livi Bacci M.. *Introduzione alla demografia.* Loescher Editore - Torino, 1999

Leti G.. *Statistica descrittiva.* Il Mulino - Bologna, 1983.

Manzelli S. *Anagrafe, iscritto anche se assente*. Italia Oggi, 6 dicembre 2005.

Simone M. *La revisione delle anagrafi e SIREA*, in L'Italia del Censimento. Struttura demografica e processo di rilevazione (a cura di Stassi G. e Valentini A.). Istat, ottobre 2012, pp. 47-49

United Nations Economic Commission for Europe Conference of European Statisticians. *Conference of european statisticians Recommendations for the 2010 censuses of population and housing*, ECE/CES/GE.41/2006, United Nations, Geneva, 2006.

Reliability of causes-of-death statistics: the Italian experience from the ICD-10 training course[1]

Francesco Grippo[2], Enrico Grande, Silvia Simeoni, Simona Pennazza, Simona Cinque, Tania Bracci, Luisa Frova

Abstract

Cause of death (CoD) statistics are a major health indicator. One of the most important instrument for improving their reliability is the appropriate use of the ICD-10 as instrument of harmonization and quality. Six research assistants recruited by Istat followed an in-depth coding course and a 8 weeks mentoring period in which they coded 4.050 cases previously coded by experts. The CoD attributed by the trainees was compared with the one attributed by experts. The overall agreement increased during the mentoring reaching the value of 78.4% which is comparable with the literature findings. From the study it emerges the relevance of having accurate and continuous training in order to achieve the best quality for official CoD statistics.

Keywords: ICD-10, mortality coding, cause-of-death statistics, ICD-10 training.

1. Introduction

Cause of death (CoD) statistics are used to monitor the health of populations and are important for health planning and setting priorities for disease prevention. The production of these data is based on harmonized tools and methodologies which allow high comparability of data in time and space. Nevertheless such statistics are exposed to many sources of variability as the completion of the death certificate, the multiple cause coding and the selection of the underlying cause of death. The reliability of CoD coding is an important factor for improving comparability of data at international level and, in order to increase it, many instruments have been developed. Among these the most important are the internationally agreed Classifications including coder's instructions and the automated coding systems (ACS). In this paper, we focus on the variability of CoD coding and how it can be reduced by an appropriate coding training. In particular we describe the experience of the training course provided to recently recruited personnel and the results achieved in terms of coding performance, quality and comparability with official statistics. In this paper, before introducing the methodology and the results, a description of the ICD-10 coding tool and of some international studies for the measure of coding reliability are provided.

[1] The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat.

[2] Istat, *e-mail*: frgrippo@istat.

1.1 The ICD-10: a tool for classifying mortality and morbidity data

The International Statistical Classification of Diseases and Related Health Problems - 10th Revision (ICD-10) is the standard diagnostic tool for epidemiology, health management and clinical purposes (WHO, http://www.who.int/classifications/icd/en/). It belongs to the International Classification Family, edited by the World Health Organization (WHO) and used for the description of health related topics (WHO, 2009). ICD-10 was adopted by the World Health Assembly in 1990. The ICD was historically developed for coding causes of death but, since 1948 (the sixth revision) it was also used for coding morbidity data. The ICD provides a system of organized categories representing different morbid entities. These categories are identified by an alphanumeric code which allows the standardization of terminology, the organized capture, memorization and the systematic analysis of morbidity and mortality data.

The ICD-10 is a dynamic tool, with an annual updating process that allows to keep up with the continuous advances in medical sciences and to ensure the best use of it. The ICD-10 is published in three different volumes: volume 1 contains the organized list of categories and other tools such as lists of tabulation; volume 2 encloses definitions and application rules; volume 3 represents the alphabetical index with the medical terminology and the related ICD-10 code. In Italy, the ICD-10-version 2009 is currently adopted and it is available for online browsing at Istat website (Istat, 2014).

1.2 Cause of death coding: complexity of rules

The starting point of CoD statistics is the death certificate where the causes of death are notified. In Italy, the certificate is filled out by a medical doctor who generally knows the medical history of the deceased (family doctor or attending physician at hospital), but can also be filled by necropsy physician when attending is not available. The Italian death certificate follows the structure of the international one, provided in the ICD-10, which consists in two parts: in the first part, the physician should report the complete sequence of conditions directly leading to death; in the second part, the other conditions contributing to death.

In the death certificate many conditions can be reported, but to allow comparison of cause of death statistics within and between countries, only one cause is selected: the underlying cause of death (UCD). The concept of UCD was introduced with the sixth revision of the ICD and corresponds to "(a) the disease or injury which initiated the train of morbid events leading directly to death, or (b) the circumstances of the accident or violence which produced the fatal injury".

The Classification provides a set of mortality coding rules that allows the standardization of the underlying cause of death selection and the choice of ICD code that better fulfills the definition of underlying cause, using all the information reported in the death certificate.

The ICD rules can be divided into two groups: the selection and the modification rules. The selection rules are *General principle*, *Rule 1*, *Rule 2* and *Rule 3* and they are used to identify the originating antecedent cause that initiates the sequence of conditions leading to deaths. The modification rules, rules A trough F, are used to modify the first selected condition in order to get a more relevant and informative code. The rules are described in the ICD-10 volume 2, where examples, applicability and comments for the application are also provided (WHO, 2009).

In practice, the application of rules presents various problems. When applying General Principle, Rule 1 and Rule 2, the coder should analyze the sequence of conditions reported by the certifier and decide if each condition is in a correct causal relationship with the others. On the other hand, when applying Rule 3, the coder should decide if the condition temporarily selected as underlying cause could be considered an obvious consequence of others conditions reported. Moreover modification rules are provided for giving a more detailed, specific and relevant information. In this case, the coder should know which condition is the most relevant in order to select it.

The ICD volume 2 provides guidelines for the application of coding rules. Despite this, some instructions could give rice to personal interpretation. An international tool for limiting this problem is represented by a set of decision tables used also by the automated coding systems. These tables were developed by the US National Institute for Health Statistics (NCHS, 2009) as part of the ACME software (CDC, 2007), a tool for the automated selection of the underlying cause of death. Successively, also in Europe, the tables have been maintained as part of Iris software (Iris, 2014), a new coding tool which integrates all phases of automated coding (text recognition, coding of each single diagnostic terms and the selection of the underlying cause) and allows an interactive handling of the rejects.

The prerequisite for the correct application of decision tables is the coding of each condition present in the death certificates. This is a complex task because different ICD-10 codes can be attributed to the same condition depending on different variables, such as the age and gender of the deceased, the duration of each condition (interval between onset of diseases and death) and the presence of other conditions on the certificate.

The complexity of coding is increased by the fact that, besides the described coding rules, other special instructions are provided for specific cases such as: perinatal and infant mortality, congenital conditions, external causes, complications of surgery.

1.3 Problems and measures of cause-of-death coding reliability: the international experience

The complexity of coding rules requires that coders should be deeply trained in the use of ICD-10. Intensive training coding course are necessary to increase the competence of coders and consequently the accuracy of CoD coding. To evaluate the effectiveness of training course and generally the reliability of CoD coding, different methods and statistical indicators are proposed by many authors. In this paragraph the strengths and weaknesses of different methods are reported.

The coding process is expected to be independent of the coding person, coding time and space. Nevertheless, despite the detailed and specific rules provided in the ICD, even coding experts show different opinion in selecting the underlying cause of death (Buchalla C. et al. 2013). This fact leads some authors to define the use of ICD for coding a “matter of chance” (Stausberg J. et al., 2008). The same Authors refer that some coding errors are due to the intrinsic limitations of ICD-10 which actually includes some ambiguities and inconsistencies.

Errors in ICD-10 coding can derive from different sources such as: (1) the incorrect and/or incomplete reporting of the causes on the death certificate by the physician, (2) the complexity of medical nomenclature and national language, (3) the interpretation of coding or selection rules, (4) individual deliberation of coders.

Actually the information available on the death certificate and how the physicians report it, is crucial for a proper coding. According to our experience, certifiers often report more than one underlying cause, despite the recommendation; moreover the reported condition often corresponds to the immediate cause or complications of the actual UCD (Grippo F et al. 2013). All these factors make arduous to properly apply the ICD-10 coding rules and to avoid the personal interpretation.

An important source of UCD variability is related to the complexity of medical nomenclature and its interpretation by coders, who normally are not medical doctors. Moreover, the use of medical terms in national language can be different, leading to national or even regional differences.

The interpretation of ICD rules and guidelines is not unequivocal ant it leaves room for individual choices (Stausberg J. et al. 2008). For most cases, the above discussed ACME decision tables provide the correct way for rules application. Nevertheless, tables do not cover all textual instruction. This is especially the case of special instructions applied when the death certificate reports complications of surgical intervention. For these instructions the ACME tables cannot be used as a reference guide. Errors that can affect the selection of the underlying cause of death can be labeled as miscoding and misspecification (O'Malley K.J. et al., 2005). Miscoding occurs when the underlying cause code is misaligned with the evidence found in the death certificate. Misspecification includes assignment of generic codes when information exists for assigning more specific codes.

The reliability of cause of death coding can be evaluated by different methods divided mainly into two groups: the ones that use gold standard (GS) and the others that don't use gold standard (NGS). In the first group, the underlying cause attributed by each coder is compared with GS, generally the UCD coded by a reviewer (Lu T.H. et al., 2000). In the second group, many coders code the same certificates and the UCDs are compared with each other (Harteloh P. et al., 2010).

The indicators used are: the percentage of agreement P (i.e. the percentage of death certificates for which all coders (or between each coder and the reviewer) give the same UCD) and the K statistic (Cohen J., 1960) generally thought to be a more robust measure than the simple percent agreement calculation since it takes into account the agreement occurring by chance. Indicating with P the relative observed agreement among raters, and with Pe the hypothetical probability of chance agreement, the K statistics is calculated as $P\text{-}Pe/1\text{-}Pe$. When the raters are in full agreement, then K=1. If there is complete disagreement, then K=-1; if there is independence among the raters K=0. Besides the K statistic, other indicators are used as false positive and false negative rates.

In Lu's article (2000) the underlying cause attributed by each coder is compared with GS: 5,621 death certificates were re-coded by an expert reviewer. The UCD selected by the expert was treated as GS and used to calculate the agreement rate and the K value. The overall agreement rates between the reviewer and coders according to the 3 digit and 2 digit categories of ICD-9 were 80.9% and 83.9%. The percentage of agreement decreases with the number of conditions per certificate and the age of deceased but not significant differences were observed by sex. Higher agreement was found for malignant neoplasms (K=0.94) and injuries and poisoning (K=0.97), but there was poor agreement for nephrotic diseases (K=0.74), hypertension-related diseases (K=0.74), and cerebral infarction (K=0.77).

In Harteloh's article (2010), the authors study the reliability of cause of death statistics in the Netherlands, calculating the percentage of agreement among coders (method NGS).

The percentage of agreement is measured as the percentage of death certificates for which all coders (four) give the same UCD. They calculated the inter-coder agreement, by comparing the UCD of each death certificate attributed by different coders and the intra-coder agreement, by comparing the UCD attributed by the same coder in different periods. 10,833 death certificates, already coded, were manually re-coded by four coders. The intra-coder agreement was 88–90% at a 4 or 3 digit level and 95–96% at chapter level. It was the same in magnitude as the inter-coder agreement for pairs of coders (87% at a 4 digit, 89% at a 3 digit and 94% at chapter level) and the authors concluded that "the coding process in itself has limited reproducibility and is not bound by individual preferences of coders". The agreement of coding process was associated with the level of detail of the ICD-10 code (chapter, 3 digit, 4 digit), the age of the deceased, the number of coders and the number of diseases reported on the death certificate. The reliability of cause-of-death statistics turned out to be high (90%) for major causes of death such as cancers and acute myocardial infarction. For chronic diseases, such as diabetes and renal insufficiency, reliability was low (70%). These conditions are associated to higher number of diseases per certificate and older age of deceased and this factors can contribute to a major variability.

It is difficult to compare different studies of reliability because of the variety of protocols and measures (different number of coders, different statistical indicators, etc.). Nevertheless, it is possible to draw some general conclusions: what emerges is that the coding reliability is lower when certificates report chronic diseases such as diabetes, hypertension related diseases, chronic liver diseases, etc., because they are long term diseases associated with old age of deceased and they are part of very complex morbid patterns. Value of coding reliability indicators also decreases with the number of codes per deceased (that increases with the age). It is necessary to understand better the coding process weaknesses and to increase the reliability of coding by an adequate training course of the coders and clearer instructions provided by the ICD, especially for some cases.

1.4 The cause of death coding process adopted in Italian National Institute of Statistics (Istat)

In Italy Istat is in charge of the cause-of-death coding. Each year, around 600 thousands death certificates are collected by Istat and electronically recorded. The certificates are processed through an automated coding system (ACS) and the rejected ones are manually coded by expert coders.

ACS process can be divided into three different steps.

Step1. The death certificates are analyzed by ACTR, a software for text recognition (Wenzowski, 1988) that transforms each recognized entry (diagnostic term) into a standardized code (Entity reference number - *ERN*) .

Step 2. The second software, MICAR (Mortality Medical Indexing, Classification, and Retrieval) converts *ERNs* in the correct ICD-10 code.

Step3. The third software, ACME, automatically applies the international rules of the ICD-10 and selects the underlying cause of death.

Rejects can be produced in each step of this process. When a certificate is rejected, manual coding is necessary. If the reject occurs for failure in step 1 or 2 the manual coder can either correct the ICD codes attributed to each rejected diagnostic term (multiple cause coding) and then submit the certificate to ACME, or can manually select the UC. Rejected in the step 3 are only handled in this second modality.

About 80% of death certificates is fully automatically coded, the remaining 20% are rejected and manually coded. The rejects are more complex than the other certificates. In fact, certificates with many diagnostic terms have more probability to be rejected and the automated coding cannot handle some complex cases such as surgery deaths, external causes, deaths mentioning drug therapies.

2 Methods

2.1 The training

Six research assistants were recruited by the Italian National Institute of Statistics for the cause of death unit. They have a university education in statistics, mathematics or biological sciences.

They followed a coding training course divided into ICD-10 lectures structured in three modules:

Module 1) Cause of death statistics: the data workflow and the use of cause of death data, one day duration: six hours;

Module 2) Coding and selecting causes of death using ICD-10 version 2009; 11 days: 71.5 hours;

a. part 1: selection and modification rule
b. training on the job on selected cases (one week),
c. special cases (external causes, complication of surgery and medical therapy, rheumatic heart diseases, infant deaths, drug poisoning, interpretation of death certificate)

Module 3) Software tools for coding: two days: 10 hours.

The lectures period lasted from January to March 2013. The reference manual for the course was prepared by Istat (2010) as an extensive integration of ICD-10 volume 2, based on NCHS manual 2a (2007) and referring to ICD updates until 2009 (Istat, 2010). The WHO training tool (2012) was also consulted.

The teachers of the course were the senior coders of the Istat cause of death unit with a long experience in ICD-10 mortality coding.

After the training course, the six research assistants (i.e. trainees) underwent also a period of mentoring lasting 17 weeks (from March to June 2013). During this period, each trainee coded real cases and was supervised by senior coders. Periodic meetings were organized in which coding doubts were clarified and some cases were revised. Coding results were evaluated and monitored to individuate possible errors of application or misunderstanding of international coding rules.

2.2 Evaluation of learning process

During the mentoring period, each trainee coded the same set of 4,050 death certificates rejected by the automated coding system. These certificates, referring to deaths occurred in Italy during the month of December 2010, had been previously coded by senior coders of Istat during the routine data processing. As discussed in the introduction, these certificates can generally be considered more complex than those fully automatically coded.

The UCD attributed by senior coders was taken as gold standard. Certificates were the same for all the trainees and, at the end, a total of 24,300 deaths certificates were available

for the analysis. The coding was computer assisted and requested the completion of multiple causes (MC) i.e. the complete coding of each condition reported on the death certificate. For certificates with complete MC, ACME software was used to select the UCD. Manual selection was performed on certificates with incomplete MC, certificates containing complications of surgery or external causes.

As previously reported in literature, many indicators of coding reliability have been used for different settings and purposes. For our objective the best indicator had to provide a direct and summary measure of misalignment between the coding performed by the trainees and the standard coding practices adopted by the Istat senior coders (gold standard). We did not use the *K* as it is designed to measure the degree to which the different coding choices agree with each other (precision) rather than the accuracy of the choice (closeness to the gold standard) (Viera A.J., 2005; Kwiecien R., 2011). Moreover, as reported in the literature, the definition of chance agreement is highly controversial (Brennan, R. L., 1981) and often not applicable in practice. In our case the probability of attribution of the same code due to chance is very low (close to 0, as the number of ICD-10 attributable codes is about 10,000). This makes the values of the raw proportion of agreement *P* very close to the *K* values, so we chose the first indicator as it enables more immediate interpretation of the results.

Therefore we used the indicator P_i defined, for each trainee *i*, as the proportion of certificates for which there was an agreement on the UCD with the senior coders.

The basic formula for the agreement P_i was:

$$P_i = \frac{n_i}{N_i}$$

where n_i was the number of certificates coded by the trainee *i* with UCD that agreed with the one attributed by the senior coder, and N_i was the total number of certificates coded by the trainee *i*.

The 95% confidence intervals for the agreement P_i was calculated as follows:

$$P_i \pm 1.96 \times \sqrt{\frac{P_i \times (1-P_i)}{N_i}}$$

The overall agreement *P* for the all six trainees combined was

$$P = \frac{\sum_i n_i}{\sum_i N_i} \qquad \text{for } i=1,\ldots,6.$$

The overall agreement was calculated at different level of detail of the ICD-10 classification: at 4 digit level; at 3 digit level and at group level.

A time-trend evaluation of the agreement was performed by calculating the indicator weekly.

The cause of death agreement was calculated by grouping the certificates according to the UCD coded by senior coders. Conforming to this approach, the proportion described above was calculated for a specific set of certificates with cause of death *c* selected as UCD (N^c):

$$P_i^c = \frac{n_i^c}{N^c}$$

In our study *c* indicated a broad category of causes of death, such as ICD-10 chapters, or specific coding topic, i.e. sequelae or rheumatic heart diseases.

To make the interpretation of results easier, the agreement was expressed as percentage.

The daily number average of coded certificates was used to monitor the increase of work rhythm during the mentoring period.

An additional analysis was carried out in order to investigate the agreement of certificates containing medical procedures. The total set of death certificates coded during the mentoring was divided into two groups: certificates containing mention of surgery and other medical procedures (781 deaths) and the certificates not containing it.

3 Results

The average number of certificates coded per day by each trainee increases during the mentoring period from 18 to 96 (table 1). Actually, this is an expected result as the ability of coding increases with the practice. Of all the 4,050 death certificates coded by each trainee, the overall agreement with the senior coders is 78.0% at 4 digit level and 82.3% at 3 digit. Both these values increment significantly over time: at 4 digit level it passes from 70.8% to more than 78.4%. The maximum value of agreement is reached in about 7 weeks (80.1%).

Table 1 – Overall agreement by mentoring week, between trainees and senior coders, at 4 and 3 digit level

Week	Number of certificates	Person – day	Average certificates coded by each trainee per day	Overall agreement P at 4 digit					Overall agreement P at 3 digit		
		(N)	(N)	%	IC95%		maximum	minimum	%	IC95%	
					inf	sup				inf	sup
1-2	1.177	65	18,0	70,8	68,2	73,4	64,9	78,7	77	74,1	79,8
3-4	1.299	56	23,0	75,1	72,7	77,5	69,1	79,8	79,9	77,4	82,4
5-6	1.542	43	36,0	77,2	75,1	79,3	66,9	88,9	82,2	79,9	84,4
7-8	2.894	57	51,0	80,1	78,6	81,6	76,2	88,2	84,2	82,7	85,8
9-10	3.143	37	85,0	78,6	77,2	80,0	75,3	84,0	83,3	81,8	84,8
11-12	4.595	55	84,0	77,6	76,4	78,8	71,5	82,5	81,7	80,4	83,0
13-14	3.308	51	65,0	79,1	77,7	80,5	74,5	83,5	84,0	82,5	85,5
15-17	6.342	66	96,1	78,4	77,4	79,4	74,2	85,4	82,1	81,2	83,1
Total	24.300	430	57,0	78,0	77,5	78,5	64,9	88,9	82,3	81,8	82,9

Thereafter the level of agreement is quite stable until the end of the mentoring period. At the same time, the variability of the agreement by trainee results to decrease: the maximum and minimum values of the agreement are observed to converge on high agreement levels (table1).

In table 2 the agreement for each trainee is presented: it has a range of variation of about 4 (from 75.8% to 79.7%). In the first four weeks this range is greater (about 6, ranging from 69.7% to 75.9%) and reaches the value of 5 in the last four weeks. The trainees which started with a lower rate of agreement compared to the others, show a greater improvement (difference between first and last four weeks). The agreement variability among the trainees decreases from 6 to 4 percent points.

Table 2 – Agreement by trainee during the first and last four mentoring weeks, at 4 digit

Trainee	Overall agreement			First four weeks			Last four weeks			Difference between first and last four weeks
	%	IC95%		%	IC95%		%	IC95%		
		inf	sup		inf	sup		inf	sup	
1	79.7	78.4	80.9	74.1	69.6	78.5	80.8	78.9	82.7	6.8
2	75.8	74.5	77.1	69.7	64.9	74.5	76.7	74.6	78.9	7.0
3	78.4	77.2	79.7	72.6	68.0	77.1	78.7	76.7	80.8	6.2
4	75.9	74.6	77.2	72.9	68.8	77.0	76.7	74.8	78.6	3.8
5	78.2	77.0	79.5	72.4	68.2	76.6	79.5	77.5	81.5	7.1
6	79.7	78.5	80.9	75.9	72.1	79.6	79.8	77.7	81.8	3.9
Total	78.0	77.5	78.5	73.1	71.3	74.8	78.7	77.8	79.,5	5.6

Another interesting result is the agreement at 4 digit by groups of UCD which has significantly different values (table 3): it varies from a value of 36.8% for medical procedures and therapies to 87.9% for congenital malformations and chromosomal abnormalities. Besides this latter group, a significantly higher value of the agreement is observed for type II diabetes mellitus, dementia and endocrine, nutritional and metabolic diseases (except diabetes) for which the value of this indicator is higher than 85.5%. For other groups of causes such as symptoms and signs; malignant neoplasm; chronic liver diseases; diseases of the nervous system and circulatory diseases (except rheumatic) the value ranges between 79.6% and 84.3%.

On the other hand, alongside mentioned medical procedures and therapies, the group of mental and behavioral disorders (excluded dementia); sequelae codes; infectious diseases (other); diseases of blood and blood forming organs; transport accidents and diseases of the digestive system (other) show low values of the agreement between 57.9% and 70.0%. For rheumatic heart diseases; other valvular diseases; diseases of the genitourinary system and external causes the agreement has a value between 70.4% and 76.1%.

The overall agreement at group level has a value of 90% (IC95% 89.6-90.3). The analysis of this indicator by cause of death confirms the results discussed at 4 digit level. Nevertheless there is a difference for external causes and transport accidents: the agreement at 4 digit shows lower values compared to the average, while the agreement at group level is significantly higher. This finding indicates that for these causes of death there are difficulties in attributing the appropriate 4 digit, but the cause of death remains classified in the same group.

A more detailed analysis of the agreement between trainees and senior coders can be drawn from table 4, where the UCD attributed by trainees, at group level, is cross-tabulated against the gold standard UCD. On the diagonal of the table, there is the percentage of certificates that are coded in the same group by trainees and senior coders (agreement at group level already presented in table 3). Figures outside the diagonal represent the percent of certificates which are attributed to another cause by the trainees compared to the gold standard. The additional information provided by this table is the possibility to evaluate the direction of the different classification between trainees and senior coders, i.e. the percent of cases allocated to a different group by the trainees. For example, from the table, it is evident the misclassification between viral hepatitis and chronic liver diseases; non-malignant neoplasm and malignant neoplasm; rheumatic heart diseases and other valvular diseases; medical procedures and sequelae.

The inter-coder variability of the agreement has a different behavior according the UCD, as shown in figure 1, where the agreement by trainee (continuos line) is compared with average one (dotted line). Distinct scenarios are presented. In correspondence of causes of death with high rate of agreement, low variability is

observed for neoplasms, while there is a certain degree of variability for congenital malformations (with two trainees out of six reaching the 100% level of agreement). At intermediate level of agreement, it is possible to observe low variability pattern (respiratory diseases) or high degree of variability (viral hepatitis). At last, the results for rheumatic heart disease and especially for sequelae represent two examples of how the inter-coder variability varies in correspondence of lower level of agreement.

Table 3 – Overall agreement at 4 digit level and at group level by underlying cause of death groups

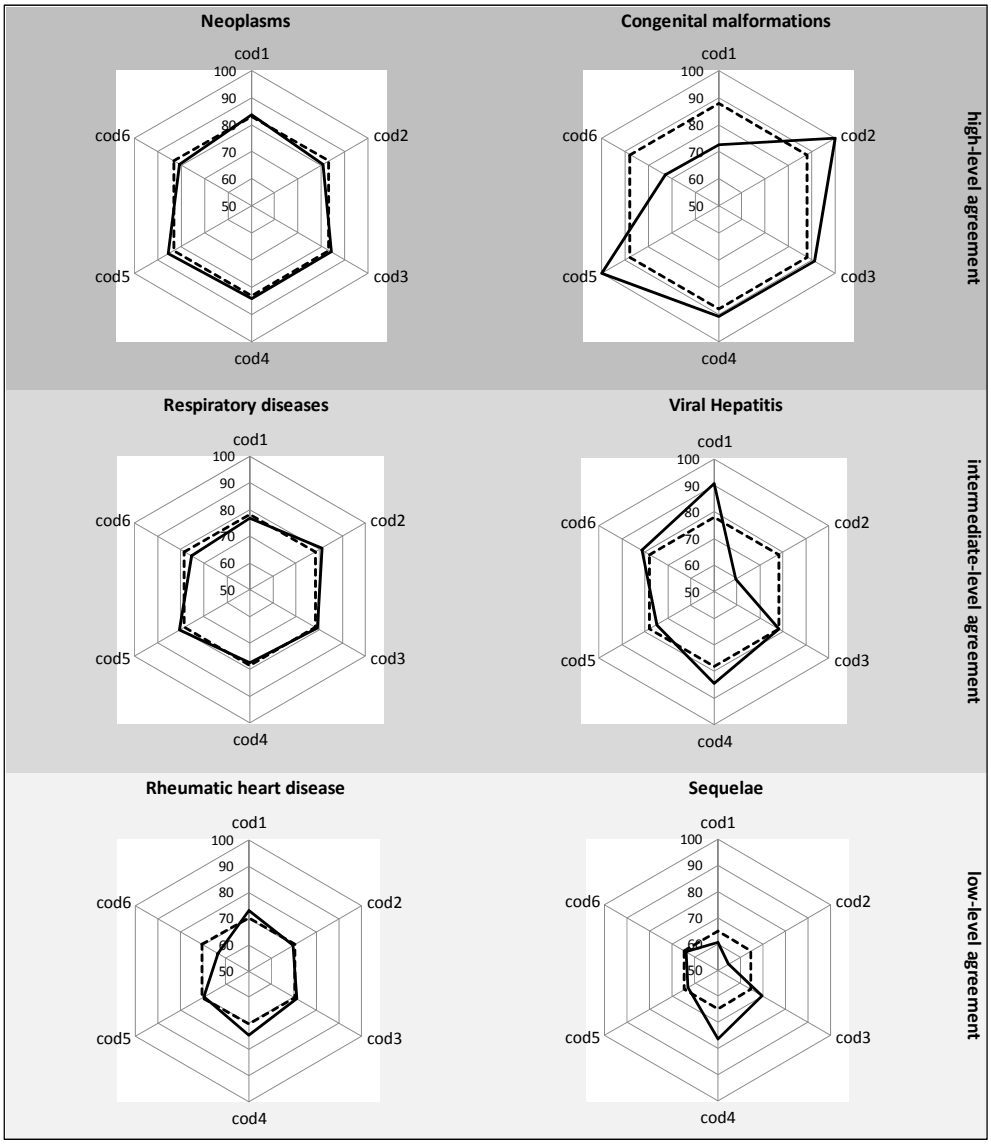
Cause of death groups	Number of certificates	Agreement at 4 digit			Agreement at group level		
			IC95%			IC95%	
		%	Inf	Sup	%	Inf	Sup
Viral hepatitis	192	78.1	72.3	83.9	83.9	78.7	89.1
Infectious diseases (other)	306	67.6	62.4	72.8	80.7	76.3	85.1
Malignant neoplasm	5,880	83.4	82.4	84.4	95.5	95.0	96.0
Other neoplasms	438	79.7	75.9	83.5	87.0	83.9	90.1
Diseases of blood and blood forming organs	162	67.9	60.7	75.1	73.5	66.7	80.3
Type II diabetes	138	86.2	80.4	92.0	93.5	89.4	97.6
Diabetes (Other)	480	80.0	76.4	83.6	88.5	85.6	91.4
Endocrine, nutritional and metabolic diseases (other)	372	85.5	81.9	89.1	92.5	89.8	95.2
Dementia	342	86.0	82.3	89.7	87.4	83.9	90.9
Mental and behavioural disorders (other)	114	57.9	48.8	67.0	72.8	64.6	81.0
Diseases of the nervous system	870	79.9	77.2	82.6	86.8	84.6	89.0
Rheumatic heart diseases	378	70.4	65.8	75.0	87.0	83.6	90.4
Other valvular diseases	444	71,2	67.0	75.4	77.7	73.8	81.6
Circulatory diseases (other)	6.45	79.6	78.6	80.6	91.1	90.4	91.8
Respiratory diseases (other)	1,140	78.3	75.9	80.7	88.7	86.9	90.5
Chronic liver diseases	306	82.4	78.1	86.7	85.9	82.0	89.8
Diseases of the digestive system (other)	924	69.9	66.9	72.9	81.0	78.5	83.5
Diseases of the genitourinary system	300	72.0	66.9	77.1	83.7	79.5	87.9
Congenital malformations, deformations and chromosomal abnormalities	66	87.9	80.0	95.8	92.4	86.0	98.8
Symptoms, signs and abnormal clinical and laboratory finding	108	84.3	77.4	91.2	89.8	84.1	95.5
Transport accidents	1014	69.6	66.8	72.4	96.7	95.6	97.8
Medical procedures and therapies	144	36.8	28.9	44.7	58.3	50.2	66.4
External causes (other)	2,670	76.1	74.5	77.7	93.8	92.9	94.7
Sequelae	732	64.8	61.3	68.3	70.2	66.9	73.5
Other	330	63.9	58.7	69.1	74.8	70.1	79.5
Total	24,300	78.0	77.5	78.5	90.0	89.6	90.3

Table 4 –Cross tabulation of underlying cause of death selected by senior coders and trainees*

Cause of death attributed by trainees	Cause of death attributed bySenior coders (gold standard)																								
	Viral hepatitis	Infectious diseases (other)	Malignant neoplasm	Other neoplasms	Diseases of blood and blood forming organs	Type II diabetes	Diabetes (Other)	Endocrine, nutritional and metabolic diseases (other)	Dementia	Mental and behavioural disorders (other)	Diseases of the nervous system	Rheumatic heart diseases	Other valvular diseases	Circulatory diseases (other)	Respiratory diseases (altre)	Chronic liver diseases	Diseases of the digestive system (other)	Diseases of the genitourinary system	Congenital malformations, deformations and chromosomal abnormalities	Symptoms, signs and abnormal clinical and laboratory finding	Trasport accidents	medical procedures and therapies	External causes (other)	sequele	Other
Viral hepatitis	83,9		0,0		3,7					1,8						8,2									
Infectious diseases (other)	0,5	80,7	0,0	0,5			0,4				0,3			0,2	0,1		4,0					0,7		0,8	2,7
Malignant neoplasm		2,0	95,5	5,3	3,1		0,4	0,3	0,3		1,7	0,3	0,2	0,2	1,2	2,6	0,9	0,3				3,5	0,1	0,4	
Other neoplasms			1,3	87,0	0,6			0,3			0,2		0,2	0,1	0,3		0,2	1,7					0,0		
Diseases of blood and blood forming organs			0,2	0,7	73,5					0,9	0,3			0,1								3,5	0,1		
Type II diabetes						93,5			0,3		0,1			0,0			0,3								
Diabetes (Other)	1,6						88,5		0,6		0,5			0,4	0,1								0,2	0,5	
Endocrine, nutritional and metabolic diseases (other)	1,0		0,2	0,2	1,9		0,6	92,5		2,6	0,1			0,1	0,2	0,7				0,9				0,4	1,2
Dementia		0,3							87,4					0,1	0,4		0,8								
Mental and behavioural disorders (other)				0,2			0,2		0,6	72,8	0,1			0,0		0,3	0,4						0,2	0,3	
Diseases of the nervous system		2,0	0,0	0,5	1,2		0,6	1,3	2,3	2,6	86,8	0,3		0,8	0,7		0,1					3,5	0,8	0,8	3,0
Rheumatic heart diseases	0,5	0,7			3,7			0,3				87,0	10,8	0,6	0,1		0,1					1,4			0,9
Other valvular diseases			0,0		1,9		0,4					4,8	77,7	0,5			0,1							0,1	
Circulatory diseases (other)	2,1	1,6	0,6	2,1	1,9	2,2	4,0	1,9	4,1	2,6	2,9	5,6	7,2	91,1	3,4		4,1	3,0	3,0	1,9	0,2	9,7	0,7	10,1	5,2
Respiratory diseases (altre)		3,3	0,4	0,2							0,6		0,9	0,5	88,7				1,5			0,7	0,4	1,4	0,9
Chronic liver diseases	7,3		0,0					2,2				0,3		0,2		85,9									
Diseases of the digestive system (other)			0,4	0,5			1,3		0,3		0,5	0,3		0,4	0,4	1,3	81,0	1,0				3,5	1,1	0,1	0,3
Diseases of the genitourinary system		0,3	0,1			0,7					1,4			0,3	0,4		0,5	83,7				4,2		0,7	0,9
Congenital malformations, deformations and chromosomal abnormalities														0,1			0,1	0,3	92,4						
Symptoms, signs and abnormal clinical and laboratory finding								0,3		0,9	0,1			0,1			0,1	0,3		89,8	0,1	2,1			
Trasport accidents										5,3				0,0	0,1						96,7		1,3	0,8	
Medical procedures and therapies		6,2	0,8	2,1	4,9	2,2	1,0	0,5	1,5	1,8	1,0	1,3	2,5	1,4	1,2	0,3	5,1	6,7	1,5	4,6	0,1	58,3	0,7	2,6	2,4
External causes (other)	3,1	0,3	0,2		3,7	0,7	0,6	0,3	0,3	3,5	1,0			1,0	1,3	0,7	0,8	2,0		0,9	2,9	2,1	93,8	10,1	7,0
sequele		2,3	0,0			0,7	0,4		1,2		1,1	0,3	0,5	1,6	0,7		1,0	1,0				0,7	0,2	70,2	0,6
Other		0,3	0,1	0,9			1,5	0,3	1,2	5,3	1,1			0,2	0,9		0,4		1,5	1,9		6,3	0,3	0,5	74,8

* The reported values are column percent. On the diagonal there is the overall agreement at group level

Figure 1 – Agreement at 4 digit by trainee, for specific underlying causes of death*



* Continuos line represents the agreement by trainee (cod1-cod6) for each cause of death, dotted line represents the overall agreement by cause

3.1 Medical procedures

A special analysis is carried out on certificates containing medical procedures (table 5). For this group of certificates the level of agreement is 74.6% at 4 digit, a lower value compared to the agreement found for the other certificates (78.8%). This confirms the major complexity of surgical certificates and the difficulty in applying the coding rules. In fact, for these cases the trainees are subjected to a greater interpretation and subjectivity.

While the total certificates show an increase of the agreement over time, the medical procedures certificates have an agreement that raises gradually until the 13th week of mentoring period, reaching 77.8%, but then it decreases until 72.3%, just one point percent more than the first weeks. Moreover the presence of medical procedures on the death certificate increases the inter-coder variability: the range of variation of the agreement (67%-78%) is wider than the range observed for the entire set of coded certificates (76%-80%).

Table 5 – Agreement in certificates with mention of medical procedures and comparison with other certificates

	Number of certificates	Agreement %	IC95%	
			inf	sup
Certificates mentioning medical procedures	4,686	74.6	73.3	75.8
Other certificates	19,614	78.8	78.2	79.3

Table 6 – Agreement in certificates with mention of medical procedures by mentoring week and by trainee

	Number of certificates	Agreement %	IC95%	
			inf	sup
Week of mentoring		Overall		
1-4	508	71.3	67.3	75.2
5-8	868	75.7	72.8	78.5
9-13	1,488	77.8	75.7	79.9
13-17	1,822	72.3	70.3	74.4
Trainee		By trainee		
Trainee 1	781	78.1	75.2	81.0
Trainee 2	781	73.2	70.1	76.3
Trainee 3	781	76.3	73.3	79.3
Trainee 4	781	66.8	63.5	70.1
Trainee 5	781	76.3	73.3	79.3
Trainee 6	781	76.7	73.7	79.7

4. Discussion

In this study, an agreement of UCD selection equal to 78.0% at ICD-10 4 digit level and 82.3% at 3 digit was found. Despite the difficulties to compare the studies of reliability because of different applied methodologies, our results fit in with other works of coding reliability and in comparison with the other countries, we perform on average. Nevertheless, comparisons with other studies are impaired by different coding practices among countries. Some studies refer to settings where the coding is performed manually for all deaths (Harteloh et. al 2010). In this situation an agreement of 88-90% was found. Although this figure appears higher than what observed for Italy, it is necessary to take into account that the present study is based on the cases rejected by automated coding, i.e. on the most complex cases. Studies conducted in settings comparable with the Italian show an agreement of 80.9% at 3 digit level (Lu T.H. et al., 2000). On the other hand the objective of the study was to evaluate the importance of a deep training course for better mortality statistics and not to evaluate the reliability of cause-of-death data.

Actually it is not possible to reach a complete agreement between coders due to many factors such as inappropriate completion of certificate by the certifying physician, personal interpretation of medical terms, complexity of ICD and different interpretation of coding rules.

In this work the degree of learning ICD-10 rules was evaluated in order to assure the agreement of the trainees with the UCD attributed by senior coders. Prior to introduce the trainees in the routine of national cause of death coding, we wanted to verify the comparability of coding with the senior coders in order to avoid discontinuity in data series. The need of reaching comparability with previous figures is the reason for choosing the coding performed by senior coders as gold standard.

During the mentoring period, we observed an increment of certificates coded per day (from 18 to 96), an increase of the overall agreement (from 70.8% to 78.4% at 4 digit) and a decrease of variability among trainees (from 6% to 4%). The improvement of trainee performance was due to a major coder experience, but even to regular didactic interventions in order to clarify coding doubts.

During the first 8 weeks the overall agreement between trainees and senior coders increased until a maximum of 80%. Then, from the 9th to the 17th week there was a slight decrease of the agreement (78.4%). This can be explained because some of the most problematic certificates were left stand-by and discussed at the end of the mentoring period. These certificates usually correspond to those not properly completed by the certifying physicians. Especially when the conditions are misspelled or not properly reported, the coder subjectivity plays an important role in coding. Hence for these certificates, a greater value of the variability among coders is expected. The variability of these cases is not related to the coder's training but to the poor completion of some certificates.

Similarly to other studies, the source of major discussion during the didactic interventions was on the choice of different code for equivocal terms, inappropriate judgment of casual relationship and incorrect interpretation of selection rule 3 and modification rules.

According to our results, some UCD categories need a particular attention and a continuous monitoring. The causes most subjected to variability are rheumatic heart diseases, sequelae, infectious diseases (especially viral hepatitis) and chronic liver conditions.

An additional study was carried out on the certificates mentioning medical procedures and therapies. All these certificates are coded manually, hence they significantly contribute to the quota of certificates manually coded: these certificates account for 3.6% of all deaths and represent approximately the 19% of the total rejected certificates. Coding these certificates is

not easy especially because they are often not correctly completed (e.g. the reason for surgery is omitted) and consequently coders have difficulties in attributing the correct UCD.

The overall agreement for the certificates mentioning medical procedures was lower compared to those not mentioning them (74.6% vs 78.8%). This clearly reflects the higher difficulty in UCD selection for these cases. This was also confirmed by the inter-coder variability that was greater than the one observed for the entire set of coded certificates. Moreover we observed a decrease in the agreement over time (from 77% to 72%). This was due to the pileup of the most complicated cases in the last weeks, as discussed above, but also to some erroneous coding practices that had affected the gold standard, identified during the didactic interventions. An important feed-back we had from this experience was the revision of some coding practices for medical procedures resulting in an improved specificity and quality of UCD coding.

Finally, the small number of subject participating to the study might be considered a limitation to the study, nevertheless all the measures are provided with confidence intervals and show robust results.

Other confounding factors, such as demographic and/or social characteristics of the trainees may have had an impact to the results. However we did not find any differences by gender (3 out of 6 where males) and all of the students had an university attainment.

Moreover our training course has been provided only to those that were afterwards enrolled in the official cause-of-death coding for Italy. By our point of view this is the strength of the study because it reflects a real case and not a theoretical investigation.

5 Conclusion

The reliability of cause of death coding is a hot topic in health statistics. The coding process has to be independent of the coding person, time and space. Nevertheless in practice the coding has an intrinsic variability and may influence trends of mortality statistics, not always allowing proper statistical comparison among countries or different periods. In Italy, 80% of death certificates are coded by the automated system that avoids the variability of coding; the remain 20% is coded manually by senior coders.

This work highlights the complexity of the coding process and how the coding variability can be reduced by appropriate training courses. To avoid bias in mortality official statistics, it is necessary that an in-depth know how on mortality cause coding is achieved by a single coder before he/she contributes to the statistical data processing.

After the training period, the percentage of coding agreement is 78% at 4 digit level. This agreement achieved is the same observed for other countries that release mortality data. Assuming that automatically coded certificates are not affected by variability (100% agreement), the final agreement estimated for Italy is 96%, calculated as the weighted average of automated and manual coding agreement.

Moreover, the study points out that a percentage of coding variability persists even after the coding course. This is due to an incorrect completion of death certificates by physicians and to ambiguities and weaknesses of classifications and coding rules that leaves room to the coders' personal interpretation.

We can assert that the CoD statistics are reliable in Italy with regard to the coding process as it is centrally managed and revised. Although further improvement on the reliability of coding can be achieved by clearer instructions on the ICD-10 coding rules and by improving the completion of death certificates by the physicians.

Acknowledgment

The training course was organized with the assistance of Advanced School for statistics and socio-economic analyses (SAES) especially with the collaboration of Tiziana Carrino and Antonio Ottaiano. The Authors would like to thank very much the teachers of the training course: Gennaro di Fraia, Stefano Marchetti, Marilena Pappagallo, Paola Rocchi; and the trainees which contributed to the coding: Gianfranco Alicandro, Annarita Mayer, Alessandro Mistretta, Simone Navarra, Chiara Orsi.

References

Brennan R. L. 1981. *Coefficient Kappa: Some Uses, Misuses, and Alternatives.* Educational and Psychological Measurement, 41: 687-699.

Buchalla C., D. Hoyert, L.A. Johansson, M.T. Cravo, P. Wood, S. Walker, T. Cawford, 2013. *The underlying cause of death coding exam: Developing New questions*, Poster presented at the WHO-FIC Annual meeting, Beijing, China, 12-18 October.

CDC (Centers for Disease Control and Prevention) 2007, National Center for Health Statistics: *About the Mortality Medical Data System.* U.S Department of Health and Human Services.

Cohen, J., 1960. *A coefficient of agreement for nominal scales*, Educational and Psychological Measurement 20: 37-46.

Grippo F., E. Grande, M. Pace, 2013. *Cause-of-death reporting: how to measure inaccurate completion.* Genus, 69: 25-45.

Harteloh P., K. De Bruin, J. Kardaun, 2010. *The realiability of cause-of-death coding in the Netherlands.* European Journal of Epidemiology, 25: 531-538.

Iris Institute,2014. http://www.dimdi.de/static/en/klassi/koop/irisinstitute/

Istat. 2014. Sistema Informativo delle Classificazioni Ufficiali. http://sistemaclassificazioni.istat.it/class/sistemaclassificazioni/

Istat, 2010. *Supplementary Instructions for Using ICD-10 to codify Causes of Death – Second Edition with WHO Updates up to 2009.* Metodi e Norme, 43.

Kwiecien R, Kopp-Schneider A, Blettner M, 2011. *Concordance Analysis*. Dtsch Arztebl Int, 108(30):515-21.

Lu T.H., M.C. Lee, M.C. Chou, 2000. *Accuracy of cause-of-death coding in Taiwan: types of miscoding and effects on mortality statistics*, International Journal of Epidemiology, 29: 336-43.

NCHS, 2007. *Instruction manual part 2a - Instructions for classifying the underlying cause of death*, National Centre for Health Statistics, Hyattsville, MD.

NCHS 2009. *Instruction Manual part 2c, ACME Decision Tables for Classifying the Underlying Cause of Death*,. US Department of Health and Human Services. http://www.cdc.gov/nchs/nvss/instruction_manuals.htm

O'Malley K.J., K.F. Cook, M.D. Price, K.R. Wildes, J.F. Hurdle, C.M. Ashton, 2005. *Measuring Diagnoses: ICD Code Accuracy*, Health Services Research, 40: 1620-1639.

Stausberg J., N. Lehmann, D. Kaczmarek, M. Stein, 2008. *Reliability of diagnoses coding with ICD-10.* International Journal of Medical Informatics, 77(1):50-57.

WENZOWSKI M.J., 1988. *ACTR - A Generalized Automated Coding System.* Survey Methodology, 14(2): 299-308.

WHO, 2012. *ICD-10 Interactive Self Learning Tool - Cause of death certificate version, for persons that fill in causes of death on a death certificate*.

http://apps.who.int/classification/apps/ICD/ICD10training/ICD-10%20Death%20Certificate/html/index.html

WHO, 2009. *International Statistical Classification of Diseases and Related Health problems – 10th revision. 2008 Edition.* World Health Organization.

Viera A. J., Garrett J. M.,2005. *Understanding Interobserver Agreement: The Kappa Statistic.* Family Medicine, 37(5):360-363.

Norme redazionali

La Rivista di statistica ufficiale pubblica contributi originali nella sezione "Temi trattati" ed eventuali discussioni a largo spettro nella sezione "Interventi". Possono essere pubblicati articoli oggetto di comunicazioni a convegni, riportandone il riferimento specifico. Gli articoli devono essere fatti pervenire al Comitato di redazione delle pubblicazioni scientifiche corredati da una nota informativa dell'autore contenente attività, qualifica, indirizzo, recapiti e autorizzazione alla pubblicazione. Ogni articolo prima della pubblicazione dovrà ricevere il parere favorevole di due referenti scelti tra gli esperti dei diversi temi affrontati.

Per l'impaginazione dei lavori gli autori sono tenuti a conformarsi rigorosamente agli standard editoriali fissati dal Comitato di redazione e contenuti nel file RSU stili o nella classe LaTex, entrambi disponibili on line. La lunghezza dei contributi originali per entrambe le sezioni dovrà essere limitata entro le 35 pagine. Una volta che il lavoro abbia superato il vaglio per la pubblicazione, gli autori sono tenuti ad allegare in formato originale tavole e grafici presenti nel contributo, al fine di facilitare l'iter di impaginazione e stampa. Per gli standard da adottare nella stesura della bibliografia si rimanda alle indicazioni presenti nel file on line.

Tutti i lavori devono essere corredati di un sommario nella lingua in cui sono redatti (non più di 120 parole); quelli in italiano dovranno prevedere anche un abstract in inglese.

Nel testo dovrà essere di norma utilizzato il corsivo per quei termini o locuzioni che si vogliano porre in particolare evidenza (non vanno adoperati, per tali scopi, il maiuscolo, la sottolineatura o altro).

Gli articoli pubblicati impegnano esclusivamente gli autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

Per contattare la redazione o per inviare lavori: rivista@istat.it. Oppure scrivere a:
Segreteria del Comitato di redazione delle pubblicazioni scientifiche
all'attenzione di Gilda Sonetti
Istat
Via Cesare Balbo, 16
00184 Roma

La Rivista di Statistica Ufficiale accoglie lavori che hanno come oggetto la misurazione e la comprensione dei fenomeni sociali, demografici, economici ed ambientali, la costruzione di sistemi informativi e di indicatori come supporto per le decisioni pubbliche e private, nonché le questioni di natura metodologica, tecnologica e istituzionale connesse ai processi di produzione delle informazioni statistiche e rilevanti ai fini del perseguimento dei fini della statistica ufficiale.

La Rivista di Statistica Ufficiale si propone di promuovere la collaborazione tra il mondo della ricerca scientifica, gli utilizzatori dell'informazione statistica e la statistica ufficiale, al fine di migliorare la qualità e l'analisi dei dati.

La pubblicazione nasce nel 1992 come collana di monografie "Quaderni di Ricerca ISTAT". Nel 1999 la collana viene affidata ad un editore esterno e diviene quadrimestrale con la denominazione "Quaderni di Ricerca - Rivista di Statistica Ufficiale". L'attuale denominazione, "Rivista di Statistica Ufficiale", viene assunta a partire dal n. 1/2006 e l'Istat torna ad essere editore in proprio della pubblicazione.

1B012015003000000

ISSN 1828-1982

€ 15,00