GIORNATE DELLA RICERCA IN ISTAT | 10-11 NOVEMBRE 2014

SESSIONE III
INTEGRAZIONE E USO DI DATI AMMINISTRATIVI PER FINI STATISTICI

I codici identificativi univoci all'interno del SIM (Sistema Integrato di Microdati)

Simone Ambroselli

Istat

I codici identificativi univoci all'interno del SIM (Sistema Integrato di Microdati)

Simone Ambroselli

Istat

ambrosel@istat.it

Sommario

L'uso crescente dei dati amministrativi a fini statistici riguarda la maggior parte degli Istituti Nazionali di Statistica (INS). Per massimizzare il vantaggio che deriva dall'avere a disposizione un'enorme quantità di informazioni è necessario costruire un sistema interconnesso di fonti amministrative acquisite dall'INS. Considerando che i dati amministrativi sono raccolti per altri fini, il loro uso statistico necessita di una definizione chiara delle unità. Ogni unità nel sistema dovrebbe essere sempre univocamente identificata anche nel corso del tempo. Tale risultato si ottiene meglio se realizzato utilizzando un sistema di identificazione e integrazione delle unità.

L'Istat si sta muovendo in questa direzione con l'accentramento di alcune funzioni per l'acquisizione, l'archiviazione, l'integrazione e la valutazione della qualità dei dati amministrativi. Nel SIM (Sistema Integrato di Microdati) si realizza l'integrazione dei microdati acquisiti da fonti amministrative e l'attribuzione dei codici identificativi univoci per: individui e unità economiche; luoghi; relazioni tra individui e unità economiche.

Parole chiave: codici identificativi, dati amministrativi, integrazione microdati.

Abstract

The increasing use of administrative data for statistical purposes is relevant for the majority of the National Statistical Institutes (NSI). To maximize the benefit that comes from having an enormous amount of information available is necessary to build an inter-linked system of the administrative sources acquired by the NSI. Considering that the Administrative Data are primarily used for other purposes, their statistical use requires a clear definition of the units. Each unit in the system should always be uniquely identified also over time. The result is best achieved by using a system of identification and integration of the units.

Istat is moving in this direction by centralizing some functions for the acquisition, storage, integration and administrative data quality evaluation. In the new system SIM (Integrated System of Microdata) is realized the microdata integration and the attribution of the unique identification codes for: individuals and economic units; places; relationships among individuals and units.

Key words: identification codes, administrative data, microdata integration.

Introduzione

Il Sistema Integrato di Microdati (SIM) su individui, famiglie e unità economiche è una struttura informativa di base realizzata mediante l'integrazione concettuale e fisica dei microdati acquisiti da fonti amministrative e statistiche di carattere censuario, organizzato con lo scopo di supportare i processi di produzione statistica dell'Istat. Gli obiettivi del processo di integrazione del SIM sono:

- identificare ogni oggetto (individui; unità economiche; loro relazioni) in fonti diverse con un codice univoco e stabile nel tempo;
- definire, per ogni oggetto, le relazioni logiche e fisiche, nel tempo e nello spazio, tra le informazioni disponibili da fonti diverse.

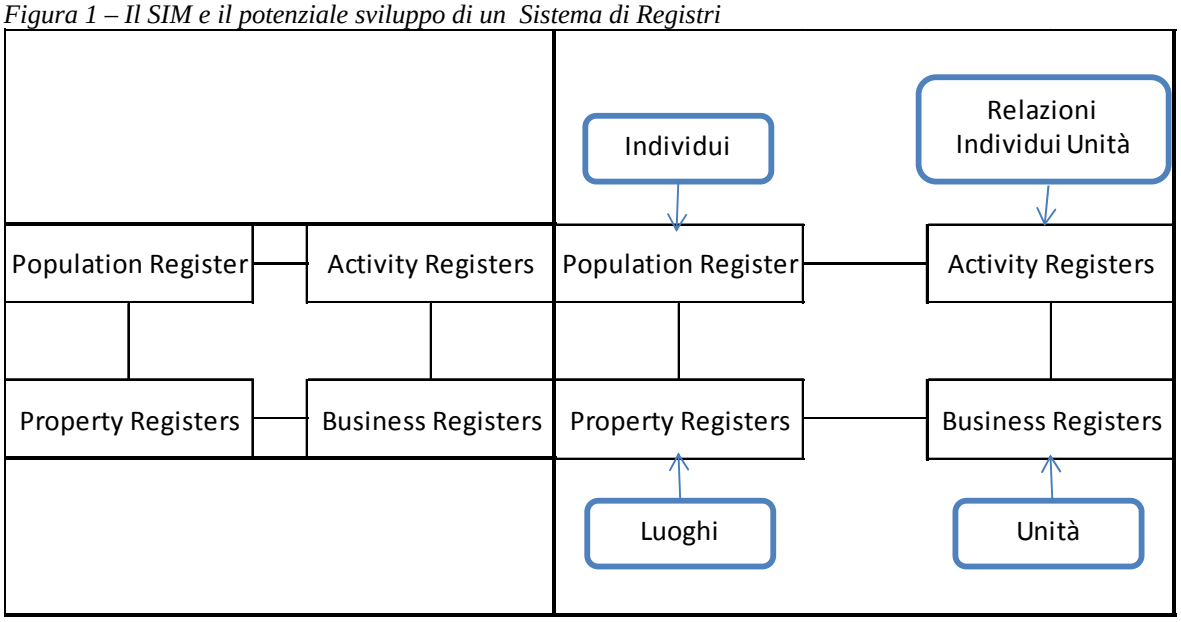
Il risultato finale è costituito dalla realizzazione di strutture di dati con unità elementari appartenenti a popolazioni statistiche utili per la realizzazione di Registri e sottosistemi informativi. Allo stesso tempo anche gli utilizzatori di specifici archivi amministrativi, ad esempio trattati in una o più fasi del processo di lavoro di un'indagine statistica, possono beneficiare dei risultati delle attività di identificazione e integrazione svolte nel SIM. I microdati di ciascuna fonte amministrativa inserita nel sistema, infatti, sono identificati in maniera univoca e risultano essere automaticamente integrati o, comunque, integrabili non solo con le altre fonti in cui sono presenti gli stessi oggetti di analisi ma anche con tutti i possibili *output* statistici che usano il SIM come punto di partenza per i loro processi.
Nel presente documento, sono dapprima sintetizzate le scelte di base che hanno portato alla progettazione del SIM, rapportate anche ad esperienze internazionali in parte comparabili.
Sono poi descritti i macro-processi alla base della realizzazione del SIM. L'integrazione delle fonti, principalmente amministrative, è sviluppata valutando il contesto generale, in relazione a problematiche trasversali, in modo da rendere la base informativa funzionale a più processi di produzione statistica.
Successivamente sono descritti i singoli sottosistemi di integrazione, in particolare, riguardo l'uso dei codici identificativi univoci assegnati alle unità di base del sistema, utili per collegare tutte le informazioni del SIM.

SIM come infrastruttura di base per i processi statistici

L'attribuzione di codici identificativi univoci validi nel tempo e per tutte le fonti a disposizione è certamente una condizione essenziale per favorire l'utilizzo dei dati amministrativi per fini statistici. Prima dello sviluppo delle procedure di assegnazione dei codici, però, è necessario definire le unità di base del sistema. Nei paesi del Nord Europa[1] i sistemi statistici sono da anni sviluppati a partire dall'identificazione di alcune unità di base con la creazione dei relativi Registri. Generalmente, le unità di base individuate sono tre: individui, proprietà immobiliari (in diverse sfaccettature interconnesse quali, terreni, immobili, case, indirizzi) e unità economiche. In alcuni contesti anche le "attività" svolte dagli individui (non solo lavorative ma, ad esempio, relative allo studio) sono considerate delle unità di base del sistema. Il risultato finale è costituito da un insieme di Registri interconnessi come mostrato nella parte sinistra della Figura 1.

[1] *Register-based statistics in the Nordic countries - Review of best practices with focus on population and social statistics* – UNECE.
To understand the Possibilities of Administrative Data you must change your Statistical Paradigm! - A. Wallgren, B. Wallgren.

Figura 1 – Il SIM e il potenziale sviluppo di un Sistema di Registri



Per la realizzazione del SIM sono state individuate due unità di base, gli individui e le unità economiche. L'integrazione delle fonti a disposizione in cui sono contenute, rispettivamente, informazioni su individui e unità economiche determina la realizzazione dei due sottosistemi di base del SIM, "SIM Unità" e "SIM Individui". I luoghi, in termini di indirizzi, associati alle unità e agli individui costituiscono un altro oggetto di analisi e il relativo processo di integrazione porta alla realizzazione dei sottosistemi "SIM Luoghi Individui" e "SIM Luoghi Unità Economiche"[2]. Infine, le relazioni tra le unità di base del sistema costituiscono ulteriori sottosistemi di integrazione: "SIM relazioni tra individui"; "SIM relazioni tra unità"; "SIM relazioni individui/unità".
Considerando lo schema proposto in precedenza, nella parte destra della Figura 1 sono illustrate le modalità con cui il SIM potrebbe integrarsi all'interno di un sistema basato sui Registri e, di conseguenza, sull'uso preponderante dei dati amministrativi.
Lo schema proposto chiarisce, dunque, come i sistemi di microdati del SIM non siano dei Registri statistici ma delle infrastrutture di supporto alla produzione statistica. L'integrazione dei dati amministrativi amplia sicuramente la possibilità di fare ricerca sia perché ci sono più dati a disposizione sia perché i processi statistici possono sfruttare la combinazione di diverse fonti per gestire eventuali errori non individuabili analizzando singolarmente i *data set* amministrativi.
Parte delle operazioni che tradizionalmente erano svolte dalle unità preposte alla realizzazione dei Registri o, nelle situazioni peggiori, da parte di tutti i singoli utilizzatori di specifiche fonti amministrative sono state accentrate e collocate a monte dei processi produttivi statistici. Gli archivi amministrativi utilizzati sono parte di un unico sistema; il loro trattamento individuale costituisce solo la prima fase di un lavoro in cui gli *output* ottenuti saranno di supporto ai processi statistici. Ad oggi, il SIM integra più di 60 *data set* amministrativi e i risultati ottenuti costituiscono certamente degli *output* autonomi su cui costruire

[2] Si deve sottolineare come la realizzazione dell'Archivio Nazionale dei Numeri Civici delle Strade Urbane (ANNCSU) permetterà all'Istat di disporre di un ulteriore sistema informativo di base necessario per il trattamento delle informazioni territoriali. E' previsto che i sottosistemi SIM dei luoghi contengano le variabili chiave per collegare SIM e ANNCSU.

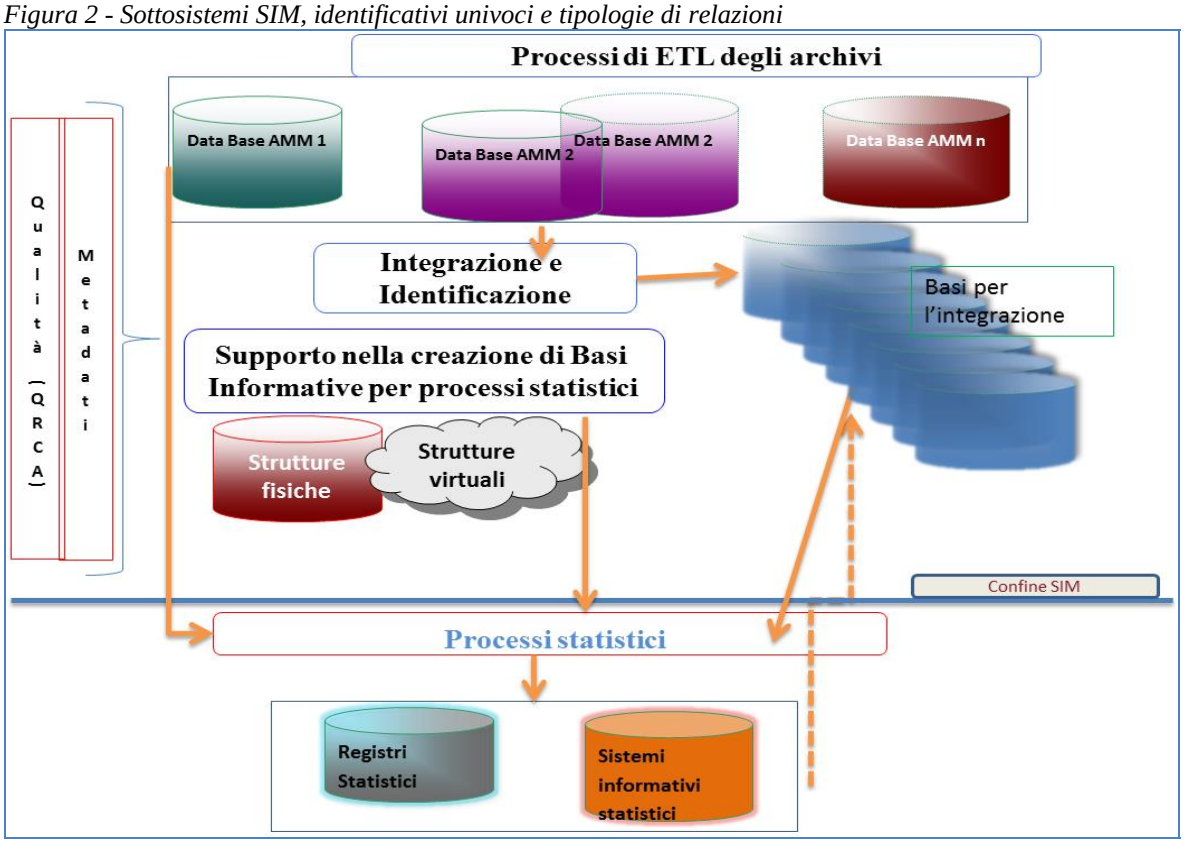
specifiche metodologie non solo di integrazione ma anche di analisi della qualità[3].

Il processo

Nel SIM integrare non vuol dire tendere ad un'unica struttura in cui convogliare tutte le informazioni presenti nelle varie fonti ma, creare delle tabelle, le basi per l'integrazione, con cui, una volta identificate le unità o le relazioni, poter collegare tutte le altre variabili disponibili nel sistema stesso.
Le macro fasi per la realizzazione del sistema sono mostrate in Figura 2:

- gestione delle singole fonti potenzialmente utili per il sistema (*ETL - extract, transform and load*);
- sviluppo delle procedure di identificazione e integrazione degli oggetti;
- supporto nella realizzazione di Basi Informative per processi statistici specifici.

Figura 2 - Sottosistemi SIM, identificativi univoci e tipologie di relazioni



Il primo passo del processo è costituito dal popolamento del Data Base con il contenuto dei *file* per ciascuna fonte di dati amministrativi inclusa nel sistema. Per ogni *data set* amministrativo viene definita la struttura concettuale per individuare i diversi oggetti rappresentati e i tipi di relazione esistenti.
Nella fase successiva avviene il vero e proprio processo di collegamento e integrazione fisica dei microdati registrati in fonti diverse secondo una specifica strategia di integrazione. L'identificazione delle unità e la conseguente integrazione è resa possibile dall'attribuzione

[3] *Towards a more efficient system of administrative data management and quality evaluation to support statistics production in Istat – G. Di Bella, S. Ambroselli* - Q2014

di un codice identificativo univoco e stabile nel tempo che alimenta lo sviluppo delle basi per l'integrazione di ciascun sottosistema. Queste ultime costituiscono dei *repository* di microdati utili per fornire una visione completa delle unità elementari e delle fonti in cui è presente. L'identificativo è riversato anche nelle singole fonti che fanno parte del relativo sottosistema. Il risultato è quello di disporre di una struttura di relazioni in cui ogni fonte è collegata, contestualmente, con la specifica base per l'integrazione e con tutte le altre facenti parte dello stesso sottosistema. Operativamente i codici identificativi del SIM sono di tipo numerico, non riconducibili a caratteristiche specifiche di identificazione delle unità.
Il supporto nello sviluppo delle basi dati informative tematiche per specifici processi statistici è l'ultima fase svolta all'interno delle attività del SIM. L'uso dei codici identificativi univoci consente di costruire strutture di microdati che costituiscono il punto di partenza per la realizzazione dei sistemi informativi statistici.
I momenti in cui gli utilizzatori statistici di dati amministrativi possono interfacciarsi con il SIM sono diversi. Infatti, è possibile utilizzare i singoli archivi amministrativi, uno o più sottosistemi o specifiche basi tematiche. In tutti i casi, però, il processo di identificazione e di integrazione delle unità elementari svolto nel SIM garantirà la produzione di *output* statistici automaticamente integrati (o integrabili).
In letteratura, generalmente, si assume che all'interno del processo di lavoro siano svolte tutte le fasi, anche quelle relative al trattamento dei dati di fonte amministrativa. Di conseguenza, le metodologie e la misura della qualità dei prodotti statistici sono sempre focalizzate su indagini o Registri che basano parte dei propri processi sull'utilizzo di *data set* amministrativi. In realtà, la realizzazione di Registri statistici permette agli utilizzatori di sfruttare il lavoro svolto a monte e di avere un riferimento comune anche per fasi particolari quali campionamenti e stime. La realizzazione del SIM costituisce un'esperienza ancora diversa in cui anche i Registri statistici sono inclusi tra gli utilizzatori delle basi integrate.

Le unità di base del sistema

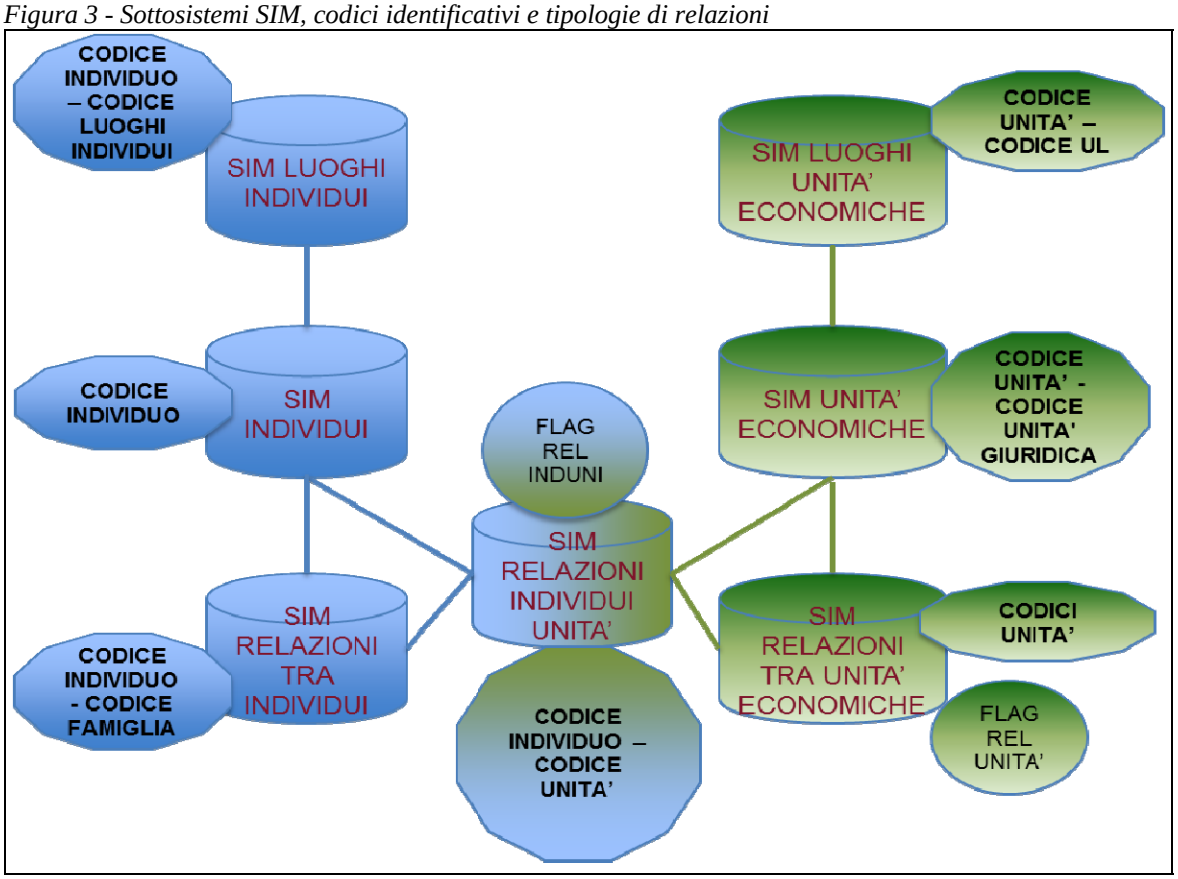
I sette sottosistemi individuati e in fase di sviluppo sono mostrati nella figura 3. Sono raggruppabili in: i) sottosistemi delle unità; ii) sottosistemi dei luoghi; iii) sottosistemi delle relazioni. Per ognuno di essi è adottata una diversa strategia di integrazione contestualmente allo sviluppo di specifici algoritmi di riconoscimento e di valutazione dei collegamenti creati. Il tradizionale dominio socio-demografico è visualizzato nella parte sinistra della figura 3 mentre quello economico è sulla destra. Al centro è posto il sottosistema in grado di collegare le informazioni sia dal lato individui sia da quello delle unità economiche.
Come detto in precedenza, i sottosistemi di base per l'intero sistema sono "SIM individui" e "SIM unità economiche". Sia da un punto di vista logico sia tecnico, questi due sottosistemi devono essere sviluppati prima degli altri perché è necessario assegnare le due chiavi primarie, rispettivamente il "codice individuo" e il "codice unità", a tutti gli oggetti di base del sistema. L'identificazione delle unità necessita di processi di riorganizzazione dei dati che vanno ad incidere solamente sulle basi per l'integrazione e non sulle singole fonti. Il processo individua in tutte le fonti le potenziali unità statistiche indipendentemente dal fatto che queste ultime siano gli oggetti di riferimento dell'archivio o un attributo di altri oggetti. Questo significa riorganizzare i dati di *input* del processo di integrazione sulla base delle unità da trattare.
Le procedure sviluppate per individui e unità economiche sono diverse. Tutte le fonti in cui sono contenuti dati sufficienti per l'identificazione degli individui alimentano la relativa base per l'integrazione. In questo caso la scelta che è stata fatta è quella di consentire agli utilizzatori di disporre del più ampio *set* possibile di unità elementari su cui implementare le metodologie specifiche dei processi statistici. Per le unità economiche, invece, l'identificazione delle unità statistiche deve seguire delle regole precise dettate dal Rego-

lamento Europeo n. 696 del 1993 che istituisce otto unità statistiche. In tale regolamento l'impresa è definita come "la più piccola combinazione di unità giuridiche che costituisce un'unità organizzativa per la produzione di beni e servizi". L'impresa è, dunque, un'unità "composita"[4] formata da una o più unità giuridiche. Queste unità esercitano totalmente o parzialmente un'attività e possono essere:

- "sia persone giuridiche la cui esistenza è riconosciuta dalla legge indipendentemente dalle persone o dalle istituzioni che le possiedono o che ne sono membri;
- sia persone fisiche che esercitano un'attività economica come indipendenti".

La base integrata per le unità economiche è, dunque, realizzata tenendo conto della necessità di individuare le unità giuridiche utili per la costruzione dell'unità statistica impresa.
Nel SIM è dapprima attribuito il codice identificativo delle unità giuridiche. In una seconda fase è possibile attribuire il "codice unità"[5] tenendo conto, però, della provenienza dell'informazione. Solo il trattamento di fonti amministrative che permettono di intercettare i legami tra partite IVA e codici fiscali determina l'attribuzione del codice unità. In definitiva, in questo sottosistema integrato sono individuate le unità economiche sia dal lato del riconoscimento formale sia da quello più propriamente statistico, in cui gli attori economici possono essere persone fisiche o giuridiche che possiedono una partita IVA.

Figura 3 - Sottosistemi SIM, codici identificativi e tipologie di relazioni



[4] *Topics of statistical theory for register-based statistics and data integration.* L. C. Zhang.
[5] Per l'identificativo delle unità economiche si è scelto di usare la denominazione "codice unità" perché il processo di attribuzione dell'identificativo non è limitato alle sole "imprese" generalmente oggetto di analisi nel settore delle *Business Statistics.*

I dati integrati sui luoghi delle unità economiche e degli individui confluiscono rispettivamente in "SIM luoghi unità economiche" e "SIM luoghi individui". Il primo contiene le localizzazioni delle unità economiche presenti nelle fonti amministrative. Nel secondo, sono presenti i luoghi che in qualche modo possono interessare le persone fisiche riconosciute dai *data set* amministrativi inseriti nel sistema: residenza anagrafica, domicilio fiscale, indirizzi delle utenze domestiche e così via. Nei SIM luoghi confluiscono le sole fonti in cui sono presenti gli indirizzi completi. A seguito di un processo di integrazione, ad ogni indirizzo completo è associato un proprio codice identificativo e la presenza delle chiavi primarie (codice individuo e codice unita) permette di collegare i microdati tra i sottosistemi di base e quelli dei luoghi.
I sottosistemi in cui sono evidenziate delle relazioni tra diversi oggetti sono tre: “SIM relazioni tra unità economiche”; “SIM relazioni tra individui”; “SIM relazioni tra individui e unità economiche”. Nel primo caso, l’obiettivo del sottosistema è quello di cogliere alcune possibili relazioni tra unità economiche quali, ad esempio, eventi di trasformazione e legami societari. Nel secondo caso lo scopo principale è identificare le relazioni tra gli individui: ad oggi, nel sottosistema, è disponibile una fonte di riferimento per individuare la famiglia anagrafica e altre, di fonte Agenzia delle Entrate, per quella di tipo “fiscale”. Nell'ultimo caso, il sottosistema integra le informazioni sulle relazioni tra gli individui e le unità economiche. Tale sottosistema si basa sulla presenza contestuale dei due identificativi necessari per legare i domini “socio-demografico” e “economico”: codice individuo e codice unità. Il sottosistema comprende, dunque, gli archivi di tipo LEED (*Linked Employer Employee Database*) in cui sono presenti le informazioni sulle imprese (datore di lavoro) e gli individui visti come lavoratori. Nei sottosistemi relazionali è presente anche una variabile che permette di distinguere il tipo di relazione individuata, affinando le attività di ricerca (*flag* relazioni). Nelle relazioni individui-unità, ad esempio, sono presenti al momento dieci diverse fattispecie riconducibili alle macro tipologie "Lavoro", "Ruolo societario" e "Studio". L’uso contestuale di codice individuo, codice unità e tipologia determina l’unità di base del sistema.
Lo sviluppo dei sottosistemi relazionali si basa sulle seguenti regole operative: i) le relazioni sono solo quelle ricavabili dai dati presenti nella singola fonte; ii) le relazioni sono inserite solo quando gli oggetti da relazionare sono identificati (cioè dotati di “codice individuo” e “codice unità”).
Sulla base delle metodologie proposte recentemente[6] riguardo la valutazione dell’errore nell’identificazione delle unità all’interno dei Registri statistici, seppur con formulazioni leggermente diverse, si deve sottolineare come lo spostamento di alcune funzioni nel SIM, prima dei singoli processi, possa determinare, in alcuni casi, uno sdoppiamento delle fasi di identificazione dell’errore stesso.
Pur considerando l’insieme degli archivi a disposizione, almeno in linea teorica, completo riguardo le popolazioni di riferimento (individui e unità economiche), la differenza con la base integrata ottenuta e, cioè, con le unità effettivamente identificate, determina comunque una parte di errore di copertura. Nel caso delle unità economiche, inoltre, la valutazione della differenza tra le unità statistiche attese e quelle ottenute riguarda sia il SIM sia chi realizza il Registro delle Unità Economiche (ASIA). In questo caso, il processo di “allineamento”, che consiste nel collegare le unità di base con quelle composite, e quello successivo di creazione o validazione delle unità statistiche determinano un ulteriore impatto

[6] B. F. M. Bakker; L. C. Zhang; ESSnet on Data Integration.

sulla popolazione effettivamente ottenuta. La teorica copertura completa del SIM su individui e unità economiche non equivale, comunque, ad avere sottopopolazioni automaticamente non affette, ad esempio, da problemi di sottocopertura. Infatti, non è detto che gli attributi in grado di caratterizzare determinate sottopopolazioni siano registrati nelle fonti o non siano contrastanti. Ad esempio, per la creazione di un sottosistema informativo sugli studenti, la presenza di eventuali *missing* nei *file* di base in cui si osserva concretamente l'oggetto di analisi, determinerebbe una sottocopertura anche nel caso in cui tutta la popolazione fosse identificata nel SIM Individui. In questo caso, infatti, mancherebbe l'informazione che qualifica l'individuo, identificato correttamente in altre fonti, come studente.

Per i sottosistemi relativi ai luoghi, invece, non si può assumere automaticamente che ci sia una completa copertura. Non tutte le fonti, infatti, dispongono di dati territoriali con lo stesso grado di dettaglio e, in questi casi, la differenza tra popolazioni target e popolazioni integrate ha un peso maggiore.

Riguardo le relazioni, per l'unità "famiglia", nel momento in cui si disporrà di più fonti da integrare, la combinazione degli effetti prodotti da errati *linkage* e errati allineamenti delle unità di base determinerà delle conseguenze nell'individuazione delle unità statistiche sia nel SIM sia nei processi a valle, soprattutto in un'ottica longitudinale.

Per le relazioni individuo-unità, infine, si deve rilevare come, a cascata, il relativo sottosistema risenta delle procedure di identificazione degli individui e delle unità economiche. Pur avendo a disposizione un legame, ad esempio una posizione lavorativa, uno degli oggetti di base potrebbe non essere stato identificato nel sistema. Nei processi statistici in cui l'obiettivo è il microdato (Registri, ad esempio) l'unità oggetto di analisi costituita dalla coppia individuo-unità potrebbe non essere utilizzata perché priva di almeno uno dei codici identificativi. O, in alternativa, potrebbero essere individuate delle opportune forme di imputazione e correzione. In casi come questi, il necessario completamento delle attività del SIM, visto come infrastruttura di base, è la possibilità di acquisire e trattare le modifiche dei Registri o di altri specifici processi statistici in modo da aggiornare contestualmente le basi integrate e le fonti d'origine attribuendo, ove necessario, nuovi codici identificativi univoci a beneficio di tutto il sistema.

Bibliografia

Bakker B. F. M. Micro-integration: State of the Art. 2010.

Council Regulation No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community

Di Bella G., Ambroselli S. Towards a more efficient system of administrative data management and quality evaluation to support statistics production in Istat. Q2014. 2014.

ESSnet on Data Integration. State of the art on statistical methodologies for data integration.

Unece. Register-based statistics in the Nordic countries - Review of best practices with focus on population and social statistics. 2007.

Wallgren A., Wallgren B. To understand the Possibilities of Administrative Data you must change your Statistical Paradigm! - Section on Survey Research Methods – JSM. 2011.

Zhang, L.-C.. Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica 66. 2012.