Istat
Istituto Nazionale
di Statistica

LINEE STRATEGICHE DEL CENSIMENTO PERMANENTE DELLA POPOLAZIONE E DELLE ABITAZIONI

METODI, TECNICHE E ORGANIZZAZIONE

METODI E LETTURE STATISTICHE

LINEE STRATEGICHE DEL CENSIMENTO PERMANENTE DELLA POPOLAZIONE E DELLE ABITAZIONI
METODI, TECNICHE E ORGANIZZAZIONE

ISBN 978-88-458-1780-9

DISTRIBUITO DA
STEALTH
BY SIMPLICISSIMUS BOOK FARM

INDICE

	Pag.
Premessa	5
Introduzione	7
1 - Caratteristiche generali del Censimento permanente	9
2 - Il conteggio della popolazione	15
2.1 - Controllo di copertura e correzione del Registro di popolazione	15
2.2 - Il disegno campionario della C-sample	18
2.3 - Tecniche di rilevazione e organizzazione della C-sample	20
3 - Il completamento dell'offerta informativa del Censimento permanente	23
3.1 - Le opportunità dell'integrazione tra dati amministrativi e statistici	23
3.2 - La rilevazione D-sample	24
3.2.1 - Il modello D-sample	24
3.2.2 - Il disegno campionario della D-sample	25
3.2.3 - Tecniche di rilevazione e organizzaznione della D-sample	27
Conclusioni	29
Allegato 1 - Il sistema di integrazione di microdati (sim) demografici, economici e sociali	31
Allegato 2 - La qualità delle fonti anagrafiche	37
Allegato 3 - L'architettura informatica per la gestione delle rilevazioni C-sample e D-sample	43
Allegato 4 - Variabili e dati da produrre con il Censimento permanente	55

PREMESSA

Il volume contiene il risultato degli studi che sono stati condotti negli anni 2012 e 2013 dal Dipartimento per i censimenti generali egli archivi amministrativi e statistici al fine di elaborare le linee strategiche sui metodi, le tecniche e l'organizzazione del censimento permanente della popolazione e delle abitazioni come stabilito dall'art. 3, del decreto legge 18 ottobre 2012, n. 179. Il percorso indicato parte dall'esperienza delle importanti innovazioni introdotte con il 15° Censimento generale della popolazione e delle abitazioni del 2011 e tiene conto delle esigenze di:

- garantire la conformità a quanto previsto dai regolamenti europei e dalle raccomandazioni internazionali mantenendo le caratteristiche di universalità, simultaneità, periodicità predefinita, delle informazioni raccolte e dei dati prodotti dal censimento permanente;
- accrescere la produzione di dati territoriali utili al Paese, nel rispetto della tradizione censuaria nazionale, ma assicurando una frequenza annuale anziché decennale delle informazioni;
- sfruttare il potenziale informativo derivante dall'utilizzo integrato di una pluralità di fonti e di tecniche di acquisizioni dei dati, garantendo una riduzione dei costi e del carico statistico sui rispondenti.

In particolare la strategia proposta delinea un censimento fortemente innovativo, basato su una pluralità di fonti amministrative e su rilevazioni campionarie a rotazione. L'impianto generale tiene conto dell'esigenza di continuare ad assicurare i principali risultati attesi dal censimento: la fornitura della popolazione legale di tutti i comuni italiani, la produzione di dati di carattere socio-demografico per rispondere alle principali esigenze informative a livello locale, regionale, nazionale; il confronto con l'anagrafe della popolazione residente al fine di una sua successiva revisione in via amministrativa.

Ulteriori approfondimenti e sperimentazioni programmate dall'Istat consentiranno la formulazione definitiva della metodologia e dell'organizzazione del censimento permanente.

INTRODUZIONE

In Italia il censimento della popolazione e delle abitazioni è stato sempre condotto, fino al 2001, mediante il metodo della rilevazione con distribuzione e raccolta diretta dei questionari di famiglia da parte dei rilevatori comunali, senza ausilio di dati desunti da fonti amministrative e in particolare di quelli delle anagrafi comunali.

In questo modo il censimento ha sempre assicurato almeno tre importanti risultati:

1. fornire il calcolo della popolazione legale di tutti i comuni italiani;
2. produrre dati di carattere socio-demografico per rispondere alle principali esigenze informative a livello locale, regionale, nazionale;
3. consentire il confronto con l'anagrafe della popolazione residente al fine di una sua successiva revisione in via amministrativa.

In occasione del 15° Censimento generale della popolazione e delle abitazioni del 2011 l'Istat ha introdotto numerose innovazioni di metodi e tecniche con gli obiettivi di semplificare l'impatto organizzativo sulle amministrazioni pubbliche, e in particolare sui comuni, di ampliare l'uso dei dati amministrativi, di recuperare tempestività nella diffusione dei dati definitivi, di ridurre il fastidio statistico sulle unità di rilevazione e di contenere i costi.

Nonostante le innovazioni del 2011, però, restano alcuni inconvenienti che contraddistinguono il censimento a cadenza decennale. Innanzitutto la disponibilità di dati si dimostra assai inadeguata in ragione delle rapide modificazioni che interessano la popolazione e la configurazione degli insediamenti. In secondo luogo, sebbene nel 2011 si sia fatto per la prima volta uso censuario dei dati contenuti nelle anagrafi comunali e negli archivi del Ministero dell'interno relativi ai permessi di soggiorno, i dati amministrativi restano in generale poco sfruttati rispetto alle potenzialità che essi offrono in Italia. Di fatto, nel censimento del 2011, i dati delle Liste Anagrafiche Comunali (LAC) non sono stati utilizzati in modo anche solo parzialmente sostitutivo dei quesiti posti nei questionari di famiglia, di convivenza e di edificio. In terzo luogo, sebbene nel 2011 ci siano stati recuperi consistenti di tempestività nel rilascio dei dati rispetto alle precedenti tornate censuarie, comunque la loro diffusione non riesce ad essere conclusa se non dopo due anni e mezzo dalla data di riferimento del censimento e ciò proprio a causa della grande scala dell'operazione sul campo e della enorme mole di dati da raccogliere ed elaborare; peraltro, con il censimento decennale l'offerta statistica di dati demografici e sociali resta limitata ad un momento raro nel tempo. Questo aspetto è in conflitto con la domanda di dati riferiti ad aree territoriali di livello comunale e sub-comunale che proviene dagli utenti e in primo luogo dagli enti locali, soprattutto se considerata a fronte degli oneri organizzativi da loro sopportati per l'operazione censuaria, delle energie spese e delle aspettative manifestate. Infine, alla concentrazione temporale delle operazioni di rilevazione censuaria corrisponde uno sforzo finanziario elevato da parte dello Stato e uno sforzo organizzativo altrettanto problematico da parte di tutti i comuni che solo in parte molto modesta rappresentano

La presente pubblicazione è stata curata da Manlio Calzaroni, Fabio Crescenzi, Marco Fortini, Andrea Mancini, Giuseppe Sindoni. Hanno partecipato alla elaborazione delle linee strategiche e alla redazione del documento: G. Gallo, P. Grossi, S. Mastroluca, A. Pezone, M. Picci, M. Venturi, R. Vivio, D. Zindato.

costi recuperabili a dieci anni di distanza dal precedente censimento.

A fronte di queste perduranti criticità le innovazioni introdotte in occasione del censimento del 2011 hanno mostrato le ulteriori opportunità che è possibile cogliere ampliando l'uso statistico delle fonti amministrative locali e centrali e intensificando l'integrazione tra i dati in esse contenuti con quelli raccolti mediante opportune rilevazioni campionarie riferite a domini territoriali di livello comunale e sub-comunale. L'aspettativa degli interlocutori politici e amministrativi e delle forze sociali ed economiche è di disporre, in tempi brevi e con continuità, di dati significativi sui principali mutamenti strutturali di tipo demografico e socio-economico. Questo fabbisogno può essere soddisfatto a costi contenuti integrando con metodi statistici pertinenti e tecnologie avanzate i numerosi e preziosi microdati già disponibili negli archivi di amministrazioni pubbliche, utilizzandoli nel rispetto del segreto statistico e delle norme poste a protezione dei dati personali.

L'obiettivo auspicato è perciò quello di una radicale revisione dell'impianto complessivo del censimento demografico, pur nel rispetto delle modalità e dei criteri stabiliti dal Regolamento (CE) n. 763/08 del Parlamento europeo e del Consiglio del 9 luglio 2008 e delle Raccomandazioni delle Nazioni Unite.[1]

Questa strategia è stata già recepita dall'Istat nell'ambito del programma Stat. 2015 e del Programma Statistico Nazionale 2014-2016 nel quale sono stati inclusi importanti lavori, quali: la realizzazione del Sistema Integrato di Microdati (SIM) da fonti amministrative plurime; l'acquisizione annuale delle LAC; la rilevazione sperimentale campionaria a rotazione del censimento permanente per il controllo del conteggio della popolazione abitualmente dimorante, desunto dai registri di popolazione residente (C-Sample); la rilevazione sperimentale campionaria a rotazione del censimento permanente per la produzione di dati socio-economici territoriali (D-Sample).

La prospettiva del censimento permanente è stata accolta anche dal decreto legge n. 179 del 18 ottobre 2012, convertito con modifiche in legge n. 221 del 17 dicembre 2012, che all'articolo 3, comma 1 prevede l'effettuazione annuale del censimento e che al successivo comma 3 autorizza l'Istat ad utilizzare i residui degli stanziamenti censuari del 2011 per realizzare le attività preparatorie all'introduzione del censimento permanente mediante indagini statistiche a cadenza annuale.

Va infine sottolineato che a livello internazionale la prospettiva del censimento permanente è stata ripresa sia nell'ambito dei lavori preparatori condotti dall'UNECE per la redazione delle nuove raccomandazioni per la tornata censuaria del 2021, sia da Eurostat in un recente documento programmatico della Direzione delle statistiche sociali[2] nel quale si prefigura di adottare per i futuri censimenti metodi di stima statistica della popolazione "usualmente residente" a partire dai dati dei registri amministrativi di popolazione residente, nonché di fondare i censimenti su basi di maggiore continuità anche mediante l'ausilio di rilevazioni focalizzate su dati territorialmente dettagliati.

In questo documento il Dipartimento per i censimenti e gli archivi amministrativi e statistici (DICA) presenta al Presidente e al Consiglio dell'Istituto Nazionale di Statistica le linee strategiche sui metodi, le tecniche e l'organizzazione del censimento permanente della popolazione e delle abitazioni, in adempimento a quanto previsto dagli obiettivi strategici assegnati per il 2013.

---

[1] Si veda "Recommendations for the 2010 Censuses of Population and Housing" ECE/CES/STAT/NONE/2006/4.

[2] Eurostat, Directorate F: Social Statistics "Issue paper for the DSS Board and DSS Discussion", Doc. Eurtostat/F/13/DSS Borad/01/EN., scheda 13, pagine 63-67.

1. CARATTERISTICHE GENERALI DEL CENSIMENTO PERMANENTE

Obiettivo generale del censimento permanente è di produrre annualmente i tradizionali dati censuari a livello comunale e sub-comunale attraverso il massimo uso dell'informazione reperibile dalle fonti amministrative e l'impiego di indagini campionarie a rotazione. Più in particolare, il censimento permanente si propone di rendere più frequente l'offerta di statistiche territoriali sulla struttura demografica di individui e famiglie e sulle loro principali caratteristiche sociali, economiche e abitative utilizzando metodi e tecniche atte a contenere i costi nel decennio di almeno il 40 per cento rispetto allo stanziamento erariale per il Censimento della Popolazione e delle Abitazioni del 2011, di ridurre considerevolmente il disturbo statistico sulle famiglie e l'impatto organizzativo sulla rete di rilevazione dei comuni.

Le informazioni di fonte amministrativa di base saranno rappresentate dalle stesse presenti nelle liste anagrafiche comunali (LAC) acquisite annualmente dall'Istat presso i servizi demografici dei comuni[1] e da quelle presenti negli archivi sui permessi di soggiorno contenenti i dati individuali relativi a stranieri con cittadinanza extra Unione Europea. Peraltro, la fonte delle LAC potrà essere sostituita dall'Anagrafe Nazionale della Popolazione Residente (ANPR), non appena questa diverrà operativa secondo quanto stabilito dall'articolo 2 del decreto legge n. 179 del 18 ottobre 2012.

Più in generale il censimento permanente si dovrà avvalere di tutte le fonti amministrative acquisite dall'Istituto all'interno del Sistema Integrato dei Microdati (SIM), già in funzione presso il DICA, entro il quale vengono operate tutte le attività di integrazione concettuale e fisica delle informazioni a livello micro. Allo stato attuale il SIM è alimentato da 58 fonti amministrative o statistiche tra loro integrate, in modo da rendere disponibili informazioni territoriali a livello di singola unità, sia essa persona fisica o giuridica.[2] Nel prossimo triennio i contenuti informativi di SIM[3] saranno regolarmente aggiornati e ampliati tenendo conto delle indicazioni fornite dalla Commissione degli Utenti dell'Informazione Statistica (CUIS), dal Comitato Consultivo per il "Sistema integrato di registri e censimenti continui per la produzione di dati strutturali e territoriali" e dal Comitato per l'utilizzazione delle fonti amministrative.

In particolare il SIM può contribuire alla realizzazione del censimento permanente fornendo apporti informativi specifici a livello di microdato individuale sia per il controllo del conteggio di popolazione (C-sample), sia per la stima delle variabili di interesse da produrre annualmente (D-sample), sia per la revisione anagrafica a cura degli enti titolari della tenuta dei registri di popolazione residente .

Rispetto al conteggio di popolazione il SIM può fornire segnali amministrativi circa la presenza sul territorio di individui, dando supporto alla stima aggregata della potenziale sovra/sottocopertura delle anagrafi a livello locale. In pratica, dal SIM è

[1] Fin quando necessario saranno riutilizzati il sistema di acquisizione STARLAC e il sistema di standardizzazione e controllo TRASLAC con geocodifica automatica degli indirizzi alle sezioni di censimento, predisposti per la conduzione del Censimento del 2011.

[2] La presenza in SIM di dati amministrativi riferiti alle unità economiche (imprese e istituzioni pubbliche e private) consente già di integrare le informazioni sugli individui contenute negli archivi anagrafici con quelle contenute negli archivi di fonte fiscale e di previdenza sociale.

[3] Le caratteristiche strutturali e i principali contenuti informativi attuali sono descritti nell'Allegato 1.

possibile individuare ex-ante informazioni a supporto dell'indagine di campo per il conteggio della popolazione.

Rispetto alla stima delle variabili di interesse censuario il SIM può fornire micro-dati utili a sostituire quesiti del questionario della D- Sample, ad integrarlo con ulteriori informazioni richieste da *stakeholder* nazionali e internazionali, a dare supporto informativo specifico alle attività di *editing*, a fornire variabili ausiliarie atte a rendere più efficiente la strategia campionaria.

Rispetto alla revisione delle anagrafi il SIM può fornire informazioni sulle probabilità di presenza/assenza di singoli individui e relative famiglie in un dato ambito territoriale. Infatti le informazioni presenti in SIM consentono di segnalare le persone che non presentano alcun legame con il territorio di competenza e quindi da candidare alla cancellazione dalla anagrafe e le persone che viceversa, anche se non sono iscritte nell'anagrafe del Comune, presentano legami con il territorio di competenza e sono quindi da candidare all'iscrizione. Il censimento permanente si potrà avvalere anche delle informazioni contenute nell'Archivio Nazionale dei Numeri Civici delle Strade Urbane (ANNCSU), realizzato ed aggiornato dall'ISTAT e dall'Agenzia delle entrate ai sensi dell'articolo 3, comma 2 del decreto legge n. 179 del 18 ottobre 2012. L'Archivio contiene le informazioni necessarie al corretto georiferimento dei dati statistici e di quelli contenuti negli archivi amministrativi del Paese sulle aree di circolazione urbane (elementi dello stradario) e sui numeri civici ad essa appartenenti, ovvero le informazioni relative alla denominazione, codifica, georiferimento e caratteristiche insediative di ciascuna area di circolazione urbana, nonché relative alla lista, codifica, georiferimento e caratteristiche insediative dei numeri civici ad essa appartenenti.[4]

Nonostante le loro notevoli potenzialità le informazioni presenti nelle fonti amministrative non sono sufficienti a garantire la qualità e la completezza dei risultati censuari. A tali fini resta necessario l'impiego di indagini statistiche dedicate da un lato alla certificazione dei conteggi relativi alle "popolazioni" di riferimento e, dall'altro lato, al completamento informativo. In questo senso la strategia generale del censimento permanente assegna specifiche funzioni alle rilevazioni campionarie annuali denominate C-sample e D–sample. La prima è finalizzata a misurare gli errori di copertura presenti nelle fonti amministrative ed eventualmente a produrre fattori di loro correzione statistica; la seconda è finalizzata ad integrare i contenuti informativi delle stesse fonti per prefissati domini territoriali, anche a livello sub comunale[5] nel caso dei comuni di dimensioni superiori ad una predeterminata soglia. Entrambe le indagini hanno come unità di rilevazione le famiglie, ma mentre la D-sample usa le LAC/ANPR come lista per la selezione delle famiglie campione, la C-sample si configura come un'indagine areale per campioni di sezioni di censimento[6] desunti dalle basi territoriali aggiornate dall'Istat con periodicità da stabilire. La cura delle geocodifiche mediante ANNCSU, e quindi del riferimento dei dati agli ambiti territoriali subcomunali, è un ulteriore strumento di gestione della qualità che si intende far valere nel censimento permanente.

Il censimento permanente basato su conteggi anagrafici, è perciò "temperato"

[4] L'elenco dettagliato delle variabili e delle loro definizioni sono stabilite da apposite direttive tecniche emanate dall'Istat d'intesa con l'Agenzia delle Entrate.

[5] Si tratta delle "aree di censimento" già utilizzate nei comuni con almeno 20.000 abitanti durante il censimento del 2011.

[6] Resta un'opzione ancora da valutare in via definitiva quella di utilizzare campioni di numeri civici tratti da ANNCSU in luogo delle sezioni di censimento.

dalla rilevazione campionaria C-sample, che misura gli errori per eccesso e per difetto dell'anagrafe (a meno dell'errore di campionamento dovuto all'indagine di controllo). Essa utilizza il metodo cattura-ricattura su un campione di sezioni di censimento e/o di numeri civici e a regime fornirà stime a livello comunale e sub-comunale (per aree di censimento nei comuni con più di 20.000 abitanti), interessando a rotazione tutti i comuni in un ciclo al massimo quinquennale. In particolare i comuni con almeno 50.000 abitanti saranno sondati annualmente, mentre ogni anno verrà sondato un diverso quinto dei comuni sotto i 50.000 abitanti, in modo che nell'arco di cinque anni tutti questi comuni siano sottoposti alla verifica del conteggio di popolazione. I comuni selezionati non saranno sondati per intero, ma solo su una loro porzione territoriale. Gli errori anagrafici verranno quindi misurati ed eventualmente corretti ammettendo un errore campionario residuo. L'errore censuario complessivo sarà mantenuto comunque inferiore a quello di partenza, garantendo un miglioramento nel risultato fornito dalla sola anagrafe o anche dal censimento tradizionale svolto con il sistema della rilevazione "porta a porta".

Un'ulteriore innovazione di rilievo è costituita dalla rilevazione campionaria a rotazione pluriennale "D" (Dati socio-economici) con questionario *long form* simile a quello adottato per il censimento del 2011. Ogni anno il campione di famiglie sarà estratto dalle liste contenute nei registri anagrafici della popolazione residente (LAC o ANPR) con riferimento a ciascun comune. Essa è disegnata per produrre stime di ipercubi di dati socio economici su famiglie, individui e abitazioni a livello comunale e sub-comunale che soddisfino le esigenze informative nazionali ed internazionali.

Entrambe le rilevazioni campionarie saranno completamente *paperless*, adotteranno tecniche di acquisizione dei dati di tipo CAWI (Computer Assisted Web Interviewing), direttamente *on line*, o CAPI (Computer Assisted Personal Interviewing) e saranno supportate da un sistema *web* di gestione della rilevazione derivato direttamente da quello utilizzato per il censimento del 2011. Nelle prime quantificazioni effettuate dal DICA il complesso delle famiglie da intervistare ogni anno potrebbe equivalere a circa 650 mila per la C-sample e a circa 1,5 milioni per la D-sample. In un decennio, quindi, l'ammontare di famiglie coinvolte sarà di circa 21,5 milioni, non superiore a quello nazionale (circa 25 milioni) producendo, al termine dei dieci anni, una diminuzione dei costi di un censimento tradizionale. Vantaggi organizzativi ed economici sono attesi dall'implementazione di strutture di rilevazione efficienti e in continuo lavoro nell'arco del tempo.

In sintesi sono cinque le componenti principali che caratterizzano il censimento permanente:

1. L'uso delle fonti amministrative, in particolare l'acquisizione e il trattamento dei dati contenuti nei registri anagrafici della popolazione residente e negli archivi dei permessi di soggiorno, per la produzione annuale di dati censuari sulla struttura demografica della popolazione con riferimento a individui e famiglie;
2. L'uso del SIM su individui, famiglie e unità economiche, realizzato mediante integrazione concettuale e fisica dei microdati acquisiti da fonti amministrative e statistiche;
3. La cura delle geocodifiche degli indirizzi e del riferimento territoriale dei dati mediante ANNCSU;
4. La misura degli errori di sovra e sottocopertura dei registri anagrafici, con o senza correzione del conteggio della popolazione abitualmente dimorante, attraverso

l'indagine campionaria a rotazione (C-sample), condotta in modo indipendente dalle liste anagrafiche;

5. Il completamento dell'informazione sulle variabili di interesse censuario per prefissati domini territoriali anche a livello sub-comunale attraverso l'indagine campionaria a rotazione (D-sample).

Secondo questa impostazione strategica il censimento permanente italiano sarebbe del tipo che le Raccomandazioni internazionali dell'UNECE e il Regolamento (CE) 9 luglio 2008, n. 763/2008 del Parlamento europeo e del Consiglio definiscono come "censimento fondato su dati amministrativi e assistito da indagini campionarie". Nei Principi e raccomandazioni dell'UNSD di New York per i censimenti del 2020 (che saranno varate nel 2015) è stato deciso di inserire esplicitamente il "Combined Census" nelle sue molteplici declinazioni.[7] D'altra parte operazioni paragonabili, per importanza, al censimento permanente sono già praticate negli Stati Uniti, dove l'American Community Survey completa le informazioni della rilevazione censuaria decennale mediante *short form* con la rilevazione ogni anno di dati socio demografici da circa 3 milioni di indirizzi ripetuta lungo un ciclo quinquennale, e in Francia, dove il *rolling census* di durata quinquennale ricorre ogni anno ad un campione di circa 3,8 milioni di famiglie (il 14 per cento della popolazione francese). Anche la Germania, la Spagna e probabilmente il Regno Unito[8] muoveranno verso un censimento che integra in modo strategico input di archivi amministrativi e registri con indagini mirate. In Europa il censimento tradizionale ad enumerazione esaustiva tenderà perciò a scomparire; infatti, mentre nei paesi del nord Europa si confermerà la preponderanza dei censimenti basati solo su registri e archivi amministrativi senza alcuna correzione dei conteggi, nei paesi europei di maggiore dimensione demografica diverse forme di *combined census*, prevalentemente con correzione o stima statistica dei conteggi[9] sostituiranno progressivamente il censimento tradizionale.

Secondo il disegno complessivo sopra sintetizzato le caratteristiche essenziali di un censimento (rilevazione individuale, simultaneità, universalità, disponibilità di informazioni per piccole aree territoriali e periodicità ben definita) sarebbero garantite dall'esaustività dell'archivio anagrafico di riferimento (LAC/ANPR), mentre la qualità della misurazione verrebbe certificata per mezzo dell'indagine campionaria di controllo. Il risultato complessivo sarebbe un conteggio della popolazione comunale mediamente più preciso sia di quello ottenibile con il censimento tradizionale, sia di quello fornito esclusivamente dall'anagrafe. L'errore di campionamento rimarrebbe comunque misurabile, mentre i costi risulterebbero contenuti e diluiti rispetto a quelli censuari. Contemporaneamente la disponibilità di dati diverrebbe annuale invece che decennale. Secondo questa impostazione i conteggi di popolazione ottenuti mediante l'anagrafe sarebbero considerati ufficiali nel caso in cui la stima dell'errore si mantenesse sotto un prefissato livello, che, una volta superato, comporterebbe invece

[7] Per "Combined Census" si intende un censimento che può impiegare dati di fonti diverse dalla raccolta esaustiva (dati di registri di popolazione/indirizzi, dati di registri e ulteriori fonti amministrative integrate) e può essere supportato da diverse tipologie di indagini (1. Campionaria esistente, 2. Campionaria ad hoc, 3. Campionaria a rotazione).

[8] Il Regno Unito sceglierà nel 2014 fra due opzioni: quella di un censimento continuo basato su archivi amministrativi e indagini ad hoc e quella di un censimento ancora decennale ma con raccolta di dati esclusivamente via web.

[9] Il Regno Unito con il "One number Census" corregge i conteggi censuari già dal 2001 e continuerà a farlo anche nel futuro censimento. La Germania applicherà nel suo futuro censimento una correzione statistica dei dati di conteggio dei registri di popolazione nei comuni oltre i 10.000 abitanti. La Francia produce i conteggi censuari mediante le stime statistiche del rolling census.

una correzione del dato anagrafico tenendo conto dei risultati della C-sample.

In Italia il censimento permanente dovrebbe essere avviato nel 2016, con data di riferimento al 1° gennaio di ciascun anno e con rilascio dei risultati entro il 31 dicembre dello stesso anno. Con questa tempistica sarebbe possibile completare il primo ciclo quinquennale delle rilevazioni campionarie in tempo utile a produrre, con riferimento al 2021, gli ipercubi di dati richiesti dal citato Regolamento dell'Unione europea. A tal fine è necessario effettuare le progettate rilevazioni campionarie sperimentali nel biennio 2014-2015, in modo da anticipare le possibili criticità e valutare compiutamente le possibili alternative per pervenire a una definizione conclusiva delle metodologie, delle tecnologie e dell'organizzazione del censimento permanente.

2. IL CONTEGGIO DI POPOLAZIONE

In generale il primo scopo del censimento demografico è sempre stato quello di enumerare gli individui e le relative famiglie abitualmente dimoranti su territori definiti a vari livelli di dettaglio. Questo obiettivo resta fondamentale anche per il censimento permanente in Italia, le cui principali unità di analisi continueranno ad essere:

- individui,
- famiglie,
- nuclei familiari,
- convivenze,[1]
- abitazioni.

Nell'ambito del censimento permanente il conteggio di popolazione sarà fondato sulle unità contenute nel Registro di popolazione.[2] La base dati di fonte anagrafica potrà essere acquisita dall'Istat, con riferimento al 1 gennaio di ciascun anno, attraverso una delle seguenti due fonti:

- le LAC, la cui rilevazione è prevista nel PSN fino al 2015, per acquisire elettronicamente, da ogni comune e secondo standard quantitativi e qualitativi predefiniti, gli archivi relativi agli individui residenti in famiglia e in convivenza;[3]
- l'Anagrafe Nazionale della Popolazione Residente (ANPR), istituita presso il Ministero dell'interno con il d.p.c.m. n. 109 del 23 agosto 2013. L'ANPR, costruita sulla base dell'Indice Nazionale delle Anagrafi (INA) e dall'Anagrafe degli Italiani residenti all'Estero (AIRE), secondo i piani attuali subentrerà alle anagrafi comunali entro il mese di maggio del 2015 e dovrebbe , quindi, essere disponibile per la data di avvio del censimento permanente.
- Nondimeno è noto che le basi dati amministrative sono generalmente affette da errori di copertura che il censimento permanente deve misurare e possibilmente correggere utilizzando le informazioni di controllo raccolte mediante la rilevazione campionaria denominata C-Sample.

2.1 Controllo di copertura e correzione del Registro di popolazione

Le indagini di controllo della copertura censuaria possono adattarsi a rispondere a diverse finalità: a) in un caso, peraltro molto noto, ci si può limitare a misurare gli errori, secondo il modello della Post Enumeration Survey già applicata al controllo delle risultanze censuarie da molti paesi, Italia compresa; b) in alternativa, oltre a misurare gli errori, i dati possono essere corretti applicando fattori generati dalla stessa indagine di controllo (modello adottato principalmente dal Regno Unito, dalla Spagna e in futuro anche dalla Germania). L'indagine campionaria C-sample rappresenta un'evo

[1] La rilevazione delle convivenze sarà oggetto di un ulteriore approfondimento *ad hoc,* data la dimensione piuttosto contenuta di questo aggregato di popolazione (poco più di 300 mila residenti al censimento del 2011).

[2] Il conteggio delle abitazioni occupate sarà basato anch'esso sulle informazioni del registro di popolazione, mentre il conteggio delle abitazioni non occupate dovrà essere effettuato, secondo modalità allo studio, a partire da altre fonti amministrative con indirizzi geocodificati alle sezioni di censimento.

[3] Nell'allegato 2 sono riportati informazioni e dati utili a valutare la qualità delle LAC e i miglioramenti attesi dalla realizzazione di ANPR.

luzione delle tradizionali indagini di controllo adattata ad un contesto di censimento permanente che fonda il conteggio di popolazione sui dati del registro anagrafico.

L'obiettivo primario dell'indagine C-sample è quindi di stimare il numero di individui abitualmente dimoranti sul territorio comunale che non sono ancora registrati nell'anagrafe del comune (sottocopertura anagrafica) e il numero di quelli che non sono più dimoranti abitualmente nel comune pur essendo registrati come residenti nella sua anagrafe (sovracopertura anagrafica).[4] La sua prima esecuzione a regime è prevista per il 2016, dopo lo svolgimento di due edizioni pilota nel 2014 e nel 2015.

Alla base del processo di misurazione degli errori di copertura è posto il metodo noto come cattura-ricattura (o Dual System) che, nel caso in esame, utilizza l'integrazione tra i dati delle LAC/ANPR e quelli della C-sample per ottenere le stime.

L'accuratezza del metodo dipende dal verificarsi di alcune ipotesi sottostanti il modello,[5] tra le quali è importante sottolineare quella di indipendenza statistica tra le due rilevazioni, dove la probabilità di essere osservati nella seconda indagine non deve dipendere dal fatto che si sia stati osservati, o meno, in occasione della prima. In concreto, come illustrato nei successivi paragrafi, questa condizione viene garantita svolgendo la C–sample mediante un'esplorazione del territorio campione senza disporre di liste anagrafiche per contattare famiglie e individui.

La stima diretta $\hat{P}$ della popolazione P del comune, corretta per la sovra e la sottocopertura, è data dalla funzione

$$\hat{P} = P_L \frac{1-\hat{s}}{1-\hat{e}}$$

dove:

- P_L è la popolazione derivante dal conteggio dei record nella LAC/ANPR comunale;
- e è il tasso di errore di sotto-copertura comunale, indicando con $\hat{e}$ una sua stima definita come rapporto tra gli individui[6] rilevati dalla C-sample che non risultano in LAC/ANPR e il totale degli individui rilevati dall'indagine[7] effettuata su un campione areale statisticamente indipendente dai dati della LAC/ANPR;
- s è il tasso di errore di sovra-copertura comunale, indicando con $\hat{s}$ una sua stima determinata come rapporto tra gli individui in LAC risultati irreperibili sul territorio campione della C-sample (dopo una verifica effettuata mediante ritorno sul campo) e il totale degli individui LAC/ANPR che fanno riferimento allo stesso territorio.[8]
- Inoltre, se indichiamo con X un vettore di dati (amministrativi o campionari da altra fonte, come la D-sample o la stessa C-sample in anni precedenti) disponibili per tutti i comuni, una possibile stima indiretta della popolazione corretta per i comuni non selezionati nel campione può essere ottenuta mediante un modello di tipo *area level*.[9]

[4] Ulteriori obiettivi della C-sample sono la stima della sotto e sovracopertura anagrafica in termini di famiglie abitualmente dimoranti, nonché il conteggio delle abitazioni non occupate. In questo paragrafo saranno affrontati soltanto gli aspetti relativi agli individui, lasciando quelli concernenti famiglie e abitazioni non occupate a successivi approfondimenti.

[5] K. Wolter (1986), Some Coverage Error Models for Census Data, *Journal of the American Statistical Association* Volume 81, Issue 394.

[6] Deve essere considerata solo la parte di LAC riferita al territorio campione della C-sample.

[7] Totale preventivamente depurato dal numero di irreperibili, così come definiti nel capoverso successivo.

[8] Una stima alternativa di s può derivare dalla D-sample, in modo probabilmente più accurato, rapportando gli individui campione che risultano irreperibili al totale degli individui del campione LAC/ANPR.

[9] In questa sede il modello è presentato solo a fini illustrativi, essendo allo studio alternative rappresentabili con una formalizzazione differente.

$$P_i = \alpha + \beta \boldsymbol{X}_i + \theta_i + \varepsilon_i$$

con $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$, e dove θ_i rappresenta gli errori da disegno delle stime P_i e stimati sul campione.

Il modello in questione mette in relazione le variabili esplicative basate sui dati amministrativi e la popolazione stimata, calcolandone i parametri sui dati dei comuni campione, per i quali si possiede anche la stima diretta di *P*, e utilizzandolo con finalità predittive sui comuni non inclusi nel campione annuo. In questo modo è possibile stimare l'errore di copertura per ciascun comune e applicare un fattore di correzione ai conteggi di popolazione derivanti dall'enumerazione basata sul registro anagrafico.

Allo scopo di controllare in modo statistico e indipendente il conteggio di popolazione basato sul registro anagrafico, si propone di utilizzare le stime campionarie degli errori di sovra e sotto copertura del registro di popolazione a livello comunale per eseguire un test statistico di validazione dell'ipotesi nulla che la popolazione residente anagrafica di ciascun comune sia la popolazione vera,[10] fissando l'errore di prima specie (rigetto dell'ipotesi nulla quando questa è vera) e minimizzando l'errore di seconda specie (accettazione dell'ipotesi nulla quando questa è falsa). Ciò può essere fatto:

- basandosi sui risultati della indagine C-sample e su informazioni desunte dal SIM;
- calcolando annualmente per ciascun comune il rapporto delle verosimiglianze sotto le due ipotesi;
- determinando le aree di accettazione e di rifiuto dell'ipotesi nulla.

Si può ottenere così una "validazione" statistica del dato di popolazione anagrafica di ciascun comune e per ciascun anno. Per i comuni la cui popolazione risulterà "validata" dal test statistico la popolazione abitualmente dimorante[11] corrisponderà a quella anagrafica. Per i comuni la cui popolazione anagrafica risulterà "non validata" dal test statistico l'Istat determinerà un fattore correttivo della stessa struttura per età, sesso e cittadinanza (italiana e straniera) sulla base dei risultati della C-sample e delle informazioni reperibili dal SIM.[12] Inoltre, per i comuni del secondo tipo si renderanno necessarie operazioni di verifica e controllo delle anagrafi nell'anno successivo alla effettuazione dell'indagine C-sample sulla base di segnali di presenza o assenza originati da dati integrati da più fonti amministrative (SIM).

Tuttavia, per motivi di costo e peso organizzativo la C – sample può essere effettuata ogni anno solo su una parte dei comuni con popolazione inferiore a 50.000 abitanti, lungo un ciclo quinquennale al termine del quale questi saranno stati tutti sondati in via diretta almeno una volta.[13] Per questo motivo si ritiene che, per tutti i comuni sotto la suddetta soglia demografica, il test statistico possa dare luogo ad eventuale correzione dei dati di popolazione anagrafica solo a partire dal sesto anno di rilevazione C – sample.

Pertanto, nel primo quinquennio di esecuzione della C-Sample (2016 – 2020) la popolazione anagrafica non sarà corretta sulla base dei risultati del test statistico e il dato annuale di popolazione ufficiale sarà aggiornato per tutti i comuni italiani con

[10] Cioè la popolazione abitualmente dimorante secondo la definizione statistica internazionale.
[11] Ovvero quella censuaria altrimenti detta "legale".
[12] Sono in corso sperimentazione finalizzate a verificare la qualità delle stime riferite alla distribuzione della popolazione a livello comunale per sesso, classe di età e cittadinanza (italiana/straniera).
[13] Vedi paragrafo 3.3.

l'usuale metodo della popolazione calcolata, applicando i dati di flusso demografico e migratorio alla popolazione legale del 2011.[14] I comuni saranno comunque informati dell'esito del test e qualora un comune rientrasse nell'area di rigetto dell'ipotesi nulla, esso verrebbe informato della correzione che sarebbe stata operata sulla base della C-sample e delle informazioni desumibili dal SIM.

Alla base della scelta di non operare la correzione in questo primo ciclo vi è l'esigenza di procedere ad un periodo di "cura" del dato anagrafico, che possa far rientrare gran parte dei comuni entro le aree di accettazione dell'ipotesi nulla e rendendo "di fatto" via via sempre più residuale la presenza di divergenze che richiedano l'applicazione di correzioni nel conteggio di popolazione.

A partire dal 2021 la popolazione censuaria dei comuni che otterranno la validazione del dato anagrafico non sarà corretta e sarà posta uguale alla popolazione anagrafica, mentre la popolazione censuaria dei comuni che non otterranno la validazione sarà corretta per un fattore di stima basato sui risultati della C-sample e sulle informazioni desumibili dal SIM. Il prospetto che segue riassume la strategia proposta in merito al conteggio della popolazione abitualmente dimorante in ciascun comune.

Prospetto 2.1 – Schema sintetico del test statistico sul Registro di popolazione

In caso di accettazione dell'ipotesi nulla	
Con correzione (dopo il 2021)	Senza correzione (fino al 2021)
Il conteggio da LAC/ANPR viene accettato	Conteggio calcolato a partire dal censimento della popolazione 2011 aggiungendo e sottraendo i flussi della ANPR Il comune viene informato che il test statistico accetta il dato di popolazione anagrafica
In caso di rifiuto dell'ipotesi nulla	
Con correzione (dopo il 2021)	Senza correzione (fino al 2021)
Controllo accurato di LAC/ANPR, utilizzando altre fonti amministrative	Controllo accurato di LAC/ANPR, utilizzando altre fonti amministrative
Il conteggio da LAC/ANPR viene corretto di un fattore calcolato dall'Istat	Conteggio calcolato a partire dal censimento della popolazione 2011 aggiungendo e sottraendo i flussi della ANPR. Il comune viene informato di quale sarebbe stato il conteggio corretto dall'Istat a seguito di test statistico

2.2 Il disegno di campionamento della C-sample

Nel suo primo ciclo quinquennale (2016-2020) la C-sample sarà usata al solo scopo di valutare statisticamente la copertura anagrafica di ciascun comune, in modo da innescare opportune attività di revisione estensiva delle anagrafi comunali che ne innalzino e ne omogeneizzino la qualità complessiva entro la prossima scadenza censuaria fissata al 2021 dal citato Regolamento dell'Unione Europea.

Nel primo quinquennio i comuni con almeno 50.000 abitanti saranno sondati tutti gli anni in un campione delle loro sezioni di censimento, mentre i comuni al di sotto di tale soglia verranno ripartiti in cinque gruppi in funzione dell'errore di copertura delle

[14] Non appena ANPR sarà avviata, i dati di flusso saranno desunti esclusivamente da essa.

loro anagrafi, così come è stato riscontrato all'ultimo censimento e congiuntamente ad ulteriori evidenze desumibili dalle fonti amministrative integrate nel SIM. Secondo questo schema, nel quinto dei comuni sotto la soglia e sondati il primo anno saranno compresi quelli con il maggiore livello atteso di errore di sotto o sovracopertura, mentre negli anni successivi saranno sondati i comuni con livelli di errore atteso via via decrescenti. Lo scopo del criterio di selezione dei comuni ora delineato risiede nella necessità di consentire alle amministrazioni comunali di curare gli errori di copertura dove si manifestano in misura maggiore e mediante opportune attività di revisione dell'anagrafe condotte anche con l'ausilio di informazioni individuali tratte dal SIM e loro trasmesse da Istat. Alla fine della rilevazione condotta in ciascun anno l'Istat stimerà gli errori di copertura e calcolerà il fattore di correzione della popolazione anagrafica. Questo potrà essere comunicato all'amministrazione comunale.

A partire dal 2021 il fattore di correzione determinato dal test statistico verrà applicato a ciascun comune per il quale l'errore di copertura è superiore alla soglia di validazione. Per questo motivo il disegno di campionamento dell'indagine dovrà essere modificato rispetto all'impostazione dei primi cinque anni. Mentre i comuni con almeno 50.000 abitanti continueranno ad essere sondati tutti gli anni in un campione delle loro sezioni di censimento, i comuni sotto la soglia dimensionale dovranno essere selezionati con criterio casuale a rotazione in un ciclo quinquennale. In questo modo in ciascun anno del ciclo i comuni selezionati saranno il più possibile rappresentativi del complesso dei comuni sotto la soglia dimensionale, così da facilitare la produzione di stime accurate dell'errore di copertura.

Più in particolare, ai fini del campionamento a partire dal 2021, i comuni italiani saranno suddivisi in tre gruppi:

- comuni con almeno 50.000 abitanti;
- comuni tra 5.000 e 49.999 abitanti;
- comuni sotto i 5.000 abitanti.

I comuni sopra i 50.000 abitanti sarebbero visitati ogni anno dall'indagine attraverso un campione areale di sezioni di censimento,[15] con la condizione che una sezione estratta in un determinato anno non sia più coinvolta nei cinque anni successivi e che il campione annuale di sezioni sia sempre rappresentativo dell'intero comune. Sulla base delle sperimentazioni finora eseguite la dimensione campionaria attesa per i comuni di questa classe di ampiezza demografica è prevista in 3.000 individui per i comuni tra 50.000 e 149.999 abitanti e 10.000 individui per i comuni con almeno 150.000 abitanti. Ulteriori sperimentazioni permetteranno di determinare meglio tali numerosità.

Nella classe dei comuni compresi tra 5.000 e 49.999 abitanti, l'indagine coinvolgerà ogni anno un diverso quinto dei comuni, in modo che nell'arco di cinque anni tutti i comuni siano sondati. Quando coinvolto, un comune appartenente alla suddetta classe non sarebbe sondato per intero, ma solo su un campione di sezioni in numero tale che siano raggiunti dall'indagine circa 1.500 individui. Negli anni in

[15] Oppure attraverso un campione di numeri civici. Valutazioni sull'opportunità di utilizzare sezioni o indirizzi dipendono dalla qualità attesa della lista di indirizzi in rapporto all'effetto cluster indotto sulle stime, oltre al costo di un rilevatore che si muove su un terreno comunale più ampio nel caso degli indirizzi e meno ampio nel caso delle sezioni.

cui un comune non è incluso nell'indagine, gli errori presenti nella sua anagrafe sarebbero comunque predetti e corretti con il ricorso a modelli statistici, sulla base di caratteristiche strutturali note per quel comune (ad esempio, gli indici di mascolinità o di vecchiaia, i tassi di fecondità o di migrazione, l'età media, indicatori di qualità dell'efficienza operativa dell'anagrafe, ecc.). Tali informazioni verrebbero utilizzate come variabili esplicative degli errori di copertura, in modelli i cui parametri saranno stati precedentemente stimati per mezzo delle informazioni raccolte nei comuni campione, parte dei quali presentano caratteristiche strutturali simili a quello non campionato.

I comuni con meno di 5.000 abitanti (circa 5.500 enti per una popolazione approssimativa di 10 milioni), sono troppo piccoli perché un campione di ragionevole dimensione possa essere estratto da ciascuno di essi al fine di produrre stime dirette (o da modello) con la stessa cadenza di quelli della classe superiore. Per questo motivo è allo studio una loro preliminare aggregazione in entità corrispondenti a circa 5.000 abitanti anche al fine di valutare effettivi vantaggi nell'accuratezza delle stime ottenibili. Analogamente a quanto previsto per i comuni tra i 5.000 e i 49.999 abitanti, anche ciascun aggregato di comuni sarebbe sondato a rotazione in un ciclo di cinque anni, estraendo un numero di sezioni di censimento in tutti i comuni appartenenti all'aggregato in modo che, in media, siano coinvolti nell'indagine 700 individui. La stima dell'errore di copertura per il generico aggregato negli anni in cui questo non è sondato sarebbe ottenuta da modello, in modo analogo a quanto avverrà per i comuni con popolazione compresa tra 5.000 e 49.999 abitanti. Il passaggio dalla stima riferita all'aggregato a quella riferita a ciascun comune componente potrà avvenire mediante applicazione di stime sintetiche che sfruttano la regolarità del fenomeno tra i comuni.

2.3 Tecniche di rilevazione e organizzazione della C- Sample

La lista di campionamento delle sezioni di censimento nelle quali effettuare ciascun anno la rilevazione di controllo sarà prodotta a partire dalle basi territoriali tenute aggiornate dall'Istat e sarà fornita agli organi di rilevazione. Peraltro, considerate le finalità di controllo della rilevazione, è necessario che essa venga svolta senza l'ausilio di liste di persone o famiglie derivanti da qualsivoglia fonte (registri anagrafici o liste di public utilities), in modo da garantire l'indipendenza dei suoi risultati dai contenuti della lista da controllare (LAC/ANPR), così come richiesto dalle ipotesi teoriche del modello cattura-ricattura. Per questo motivo si ritiene necessario affidare le attività di rilevazione sul campo a soggetti indipendenti dalle amministrazioni comunali e più in generale indipendenti da enti che abbiano interessi diretti o indiretti alla determinazione della popolazione legale.

La tecnica di indagine prevede che i rilevatori esplorino gli indirizzi compresi nelle sezioni campione alla ricerca di popolazione abitualmente dimorante, secondo la tradizionale tecnica porta a porta che ha contraddistinto il censimento italiano fino al 2001.[16] I dati saranno acquisiti con tecnica CAPI direttamente dal rilevatore (indagine *paperless*) mediante un questionario elettronico breve contenente quesiti relativi a

[16] I rilevatori, di norma uno ogni 320 famiglie, muniti del materiale di ausilio alla rilevazione (mappe, itinerari di sezione, ecc.) e di un dispositivo di tipo tablet, percorreranno l'area loro assegnata e procederanno alla rilevazione utilizzando le applicazioni sw predisposte dall'Istat.

dati anagrafici di ciascun componente della famiglia e alcune informazioni utili per il calcolo degli errori di conteggio (ad esempio titolo di godimento dell'alloggio, stato civile, titolo di studio).

Attraverso il Sistema di Gestione della Rilevazione (SGR) i rilevatori saranno in grado di registrare tutte le informazioni necessarie per il monitoraggio in tempo reale delle attività (appuntamenti, note sulle famiglie non trovate per le quali è necessario un ritorno, ecc.). Gli applicativi del questionario elettronico e di SGR saranno sempre sincronizzati e in grado di aggiornare contestualmente i data base Istat nel caso di collegamento a Internet attivo, mentre nel caso di assenza di connessione a Internet la sincronizzazione e l'aggiornamento del data base avverranno ad opera del rilevatore nel momento in cui la connessione risulterà attiva.[17]

La rilevazione, avendo come obiettivo la misura degli errori di copertura di LAC/ANPR, dovrà essere eseguita sul campo in un periodo il più possibile vicino alla data di riferimento delle stesse (1 gennaio di ciascun anno). Si ritiene, quindi, che la rilevazione debba essere organizzata in modo che essa possa essere effettuata nel periodo tra gennaio e marzo. Contestualmente alle attività di rilevazione sul campo si procederà al *linkage* tra gli individui rilevati e quelli iscritti in LAC/ANPR. Per ciascuna sezione campionata verrà redatta la lista degli individui iscritti in LAC/ANPR ma non trovati durante la prima fase della rilevazione. In questo modo sarà possibile effettuare tempestivamente la successiva fase di ritorno sul campo guidata da lista e finalizzata alla ricerca degli individui presenti in LAC/ANPR ma non rilevati nella prima fase per eventuale errore dell'indagine di controllo.[18] Le stime relative agli errori di copertura di LAC/ANPR (sotto e sovra copertura) potranno essere prodotte entro dicembre, dopo aver svolto la fase di controllo e correzione, avvalendosi anche di dati rilevati mediante la D-sample.

Prospetto 2.2 - Cronogramma della C-sample annuale

Attività	Tempi
Estrazione sezioni campione	Entro 1° gennaio
Acquisizione LAC/ANPR per il linkage	Gennaio-marzo
Attivazione della rete di rilevazione	Entro terza settimana di gennaio
Ricognizione della sezione	Ultima settimana di gennaio
Rilevazione sul campo	Febbraio-marzo
Recupero con lista	Entro la quarta settimana di aprile
Controllo e correzione	Entro luglio
Produzione e diffusione stime	Entro dicembre

[17] Per approfondimenti sull'architettura informatica per la gestione della C-Sample si consulti l'allegato 3.
[18] Al momento si è stimato che per i comuni sotto i 50.000 abitanti l'ordine di grandezza medio dei ritorni sarà intorno al 4-6% del campione originario, mentre per i comuni con più di 50.000 abitanti sarà intorno all'8%.

3. IL COMPLETAMENTO DELL'OFFERTA INFORMATIVA DEL CENSIMENTO PERMANENTE

3.1 Le opportunità dell'integrazione tra dati amministrativi e statistici

Gli archivi di fonte amministrativa, prevalentemente di natura fiscale, previdenziale, camerale e assicurativa, contengono molte informazioni su caratteristiche demografiche e sociali relative ad individui, quali sesso, età, nazionalità, luogo di residenza, luogo di nascita e partecipazione al mercato del lavoro, oltre che informazioni sulla loro appartenenza a famiglie. L'Istituto, con l'obiettivo di rendere fruibile a fini statistici questo patrimonio informativo, ha costruito il Sistema Integrato di Microdati (SIM)[1] nel quale sono stati integrati archivi di fonte amministrativa in cui l'individuo è presente nella doppia veste di soggetto dichiarante (in quanto legato da relazioni di titolarità o partecipazione in unità economica) e di soggetto dichiarato (in quanto lavoratore, studente, pensionato eccetera). In questo modo è possibile ricostruire le connessioni logiche esistenti fra le unità, sfruttando la presenza di codici d'identificazione univoci delle persone fisiche nelle diverse fonti di input. Ad oggi il Sistema integra circa 60 diverse fonti amministrative per circa 500 milioni di record all'anno e nel prossimo futuro le informazioni saranno accompagnate da misure di qualità. Il SIM include le seguenti informazioni: Individui e famiglie con relative caratteristiche demografiche, occupazionali e di istruzione;

- luoghi di studio e di lavoro;
- unità frequentate dai singoli individui come luogo di lavoro, di studio e di abitazione;
- relazioni che legano le precedenti entità (ad esempio, il legame tra individui e scuola come studente piuttosto che insegnante).

Applicando tecniche di linkage e di georeferenziazione degli indirizzi è stato possibile collocare una parte rilevante degli individui nel luogo di dimora abituale e nel luogo dove svolgono l'attività di studio o di lavoro impiegando le informazioni presenti negli archivi (indirizzo di domicilio fiscale, indirizzo di residenza), sia a livello di persona fisica che di unità.

Lo studio delle fonti amministrative presenti nel SIM, con l'analisi delle definizioni, classificazioni e campo di osservazione e la misura della qualità dei dati rende possibile definire le variabili in esso contenute che potranno essere utilizzate, direttamente o indirettamente, a supporto del censimento permanente. A partire da questi studi sarà possibile:

- sostituire completamente variabili da rilevare al censimento con variabili presenti nelle fonti amministrative;
- iIntegrare nuove informazioni in aggiunta a quelle tradizionalmente rilevate;
- dare supporto alle attività di correzione/integrazione delle variabili rilevate, utilizzando informazioni presenti nel SIM per l'imputazione delle risposte mancate o incoerenti;
- individuare variabili ausiliarie a supporto delle strategie campionarie, al fine di

renderle più efficienti e di mirare le rilevazioni verso territori ove si presentino maggiori criticità.

In particolare i risultati degli studi relativi ai primi due punti potranno rendere possibile la riduzione delle informazioni da rilevare mediante questionario con l'indagine campionaria D-sample, adibita al completamento informativo in termini di variabili da stimare.[1]

3.2 La rilevazione D-Sample

La rilevazione campionaria D-sample è concepita per produrre stime di ipercubi di dati socio economici non presenti in anagrafe e in SIM e riferiti a individui, famiglie e loro abitazioni, al fine di soddisfare le esigenze informative nazionali e internazionali con riferimento a domini territoriali prefissati ad un livello comunale e sub-comunale. La sua progettazione è stata avviata nel 2013. La sua prima esecuzione a regime è prevista per il 2016, dopo lo svolgimento di un'edizione pilota nel 2015.

3.2.1 Il modello della D-sample

La D-sample si caratterizza come un *rolling sample* così come definito da L. Kish[2] in diversi suoi lavori. Pertanto si considera un disegno suddiviso in k campioni periodici non sovrapposti di famiglie, ciascuno dei quali con probabilità di selezione *f=1/F* dell'intera popolazione, tali che il cumulo di *k* periodi fornisca un campione dell'intera popolazione con una frazione di campionamento *f'=k·f*. Nel caso in cui sia *k=F*, la popolazione si intende interamente misurata in un periodo di *F* anni e il disegno assume il nome di *Rolling Census*.

Se consideriamo la variabile di interesse Y, il cui stimatore annuale viene indicato con $\hat{Y}_i$, *i=1,...,k*, è possibile ottenere la stima $\hat{Y}(W) = \sum_{i=1}^{k} W_i \hat{Y}_i$, con pesi *W=(W1,..., Wk)* tali che $\sum_{i=1}^{k} W_i$ =1. In questo modo la stima $\hat{Y}(W)$ cumula i campioni di *k* anni.

Secondo questo schema, se si assumesse $W_k=1$ e $W_i=0$, $i \neq k$, si starebbe considerando il solo campione dell'ultimo anno. Invece, nel caso in cui sia $W_i=1/k$, $\forall i$, si ottiene una media dei pesi su tutto il periodo. Di conseguenza lo stimatore può essere considerato una media di periodo riferita all'anno centrale del periodo medesimo. Altri schemi di pesi possono considerare l'aumento monotono dei pesi $(W_1<W_2<...<W_k)$ o la loro diminuzione $(W_1>W_2>...>W_k)$.

Se indichiamo con $\sigma^2_i = \sigma^2$ la varianza dello stimatore $\hat{Y}_i$ riferito all'*i-mo* anno, la varianza dello stimatore $\hat{Y}(W)$ è pari a $\sigma(W) = \sum_{i=1}^{k} W_i^2 \sigma_i^2$. Si dimostra facilmente che il valore minimo della varianza di $\hat{Y}(W)$ si ottiene quando il sistema dei pesi è $W_i=1/k$, $\forall i$.

Nel caso della D-Sample per il censimento permanente italiano si propone di

---

[1] Il censimento permanente dovrà garantire la produzione dei dati tradizionalmente offerti dal censimento decennale in Italia e di quelli richiesti dal Regolamento (CE) n. 763/08 del Parlamento europeo e del Consiglio del 9 luglio 2008. Per un esame delle variabili censuarie si veda l'Allegato 4.

[3] Leslie Kish, Space/Time Variations and Rolling Samples, *Journal of Official Statistics*, Vol. 14, No. 1, 1998, pp. 31-46.

adottare il modello fissando un periodo quinquennale di rotazione *(k=5) con F > k.* In questo modo la D – Sample non costituisce un caso di *Rolling Census* secondo la definizione di Kish, ma piuttosto un caso di *Rolling Sample*, alla stessa stregua del censimento francese. Inoltre si propone che i campioni annuali siano cumulati con peso uguale per ciascun anno in modo da garantire la maggior efficienza delle stime di periodo.

3.2.2 Il disegno di campionamento della D-sample

Il disegno di campionamento prevede ogni anno l'estrazione casuale, eventualmente con stratificazione, di un prefissato numero di famiglie dalle LAC di tutti i comuni italiani o dall'ANPR subordinatamente alla sua entrata in funzione. I campioni di anni successivi saranno coordinati in modo che una famiglia estratta nel campione dell'anno t non possa essere estratta per i successivi quattro anni. La dimensione del campione di famiglie annualmente estratte in ogni comune sarà determinata in modo da garantire l'efficienza delle stime degli ipercubi ai diversi livelli territoriali. Gli ipercubi ottenuti mediante il cumulo del campione su tre o cinque anni saranno riferiti all'anno centrale del periodo, in modo da garantire la massima efficienza possibile delle stime.[4]

Sono definiti i seguenti domini territoriali minimi rispetto ai quali è prevista la diffusione di ipercubi con precisione pianificata da disegno:

1. aggregazioni di comuni con meno di 5.000 abitanti;[5]
2. comuni tra 5.000 e 20.000 abitanti;
3. aree di censimento di comuni con uno o più centri abitati, ciascuno di almeno 20.000 abitanti.[6]

Allo stato attuale l'allocazione del campione tra i comuni è stata provvisoriamente effettuata con l'obiettivo di garantire che la stima di un fenomeno incidente nella popolazione con frequenza p=0,5 per cento sia compresa in un intervallo tra 0,32 per cento e 0,65 per cento con una probabilità del 90 per cento. I calcoli si fondano sull'uso della distribuzione ipergeometrica per i comuni fino a 10.000 abitanti e sulla distribuzione Poisson per i comuni di dimensione superiore. Come distribuzione dei comuni secondo la popolazione residente si è fatto riferimento ai dati di popolazione legale riferiti al 9 ottobre 2011, supponendo di aggregare[7] in circa 2.000 entità da 5.000 abitanti i comuni di dimensione inferiore alla soglia dei 5000 abitanti. I risultati delle elaborazioni effettuate sono riportati In Tabella 1 per classi di dimensione comunale. Secondo le ipotesi adottate la numerosità campionaria annuale si attesterebbe intorno 1,5 milioni di famiglie.

[3] Sono in corso sperimentazioni per valutare gli effetti sull'efficienza delle stime causati dal loro riferimento ad un anno successivo a quello centrale del periodo pluriennale. Tuttavia i risultati provvisori mostrano che spostare l'anno di riferimento delle stime ad un anno più recente di quello centrale provocherebbe un aumento della varianza campionaria delle stime dirette, vanificando in parte l'effetto del cumulo temporale del campione.

[5] Le aggregazioni di comuni di piccole dimensioni demografiche per la D –Sample saranno le stesse di quelle utilizzate per la C – Sample. Per i comuni sotto i 5.000 abitanti stime da modello possono essere comunque considerate a livello comunale.

[6] Le aree di censimento sono state già individuate e utilizzate come domini di stima nell'ambito del censimento del 2011.

[7] Per ora si è proceduto semplicemente a dividere per 5000 il totale della popolazione residente in comuni con popolazione inferiore a 5000 abitanti.

Tabella 3.1 - Allocazione comunale del campione di famiglie per classi di dimensione demografica comunale

CLASSI DI DIMENSIONE COMUNALE	Comuni	Domini	Popolazione	Frazione di campionamento %	Individui campione comunale	Famiglie campione comunale	Famiglie campione totale	Individui corrispondenti totale
(0-5.000)	5.702	2.065	10.324.330	9,6	480	200	412.973	991.136
(5.000-10.000)	1.185	1.185	8.380.615	8,9	629	262	310.781	745.875
(10.000-20.000)	698	698	9.591.763	5,8	797	332	231.801	556.322
(20.000-max)	507	2.274	31.137.036	4,0	548	228	518.951	1.245.481
TOTALE	8.092	6.222	59.433.744	6,0	569	237	1.474.506	3.538.814

Nel complesso il tasso di campionamento annuo sarebbe dell'ordine del 6 per cento delle famiglie residenti. In questo modo la popolazione intervistata in un decennio sarebbe approssimativamente pari al 60 per cento di quella nazionale. Peraltro si osserva che il tasso di campionamento nei piccoli comuni risulta più che doppio rispetto a quello nei più grandi. A titolo esemplificativo un comune di 50.000 abitanti sarebbe interessato da un campione annuo di 833 famiglie, mentre per il comune di Roma il campione raggiungerebbe le 45.000 famiglie.

Attraverso lo schema considerato, cumulando i campioni annuali per 5 anni, si può garantire la stima obiettivo per gli aggregati di comuni sotto i 5.000 abitanti, per i comuni compresi tra 5.000 e 35.000 abitanti e per le aree di censimento dei centri urbani di maggiori dimensioni. Per i comuni con popolazione compresa tra i 35.000 e i 100.000 abitanti (o più in generale per i territori inclusi in questo ambito) le stime dovrebbero essere ottenute cumulando i campioni annuali per tre anni , mentre per i comuni (o più in generale territori) con almeno 100.000 abitanti lo schema garantisce la possibilità di produrre stime annuali. Il prospetto 3 riassume lo schema di diffusione degli ipercubi secondo il quale la produzione di dati riferiti a domini territoriali (comunali, provinciali e regionali) superiori a 100.000 abitanti inizierebbe nel 2016 con riferimento allo stesso anno , quella riferita a domini territoriali tra 35.000 e 100.000 abitanti inizierebbe nel 2018 con riferimento all'anno precedente, quella riferita a domini territoriali con popolazione inferiore a 35.000 abitanti inizierebbe nel 2020 con riferimento a due anni prima. Tuttavia, occorre precisare che, per ciascun prefissato dettaglio territoriale, esiste comunque un limite al dettaglio tematico ammissibile, superato il quale la produzione degli ipercubi avrà accuratezza delle stime inferiore a quella desiderata. Questo fa si che sarà più facile rispettare l'accuratezza delle stime per un comune di 100.000 abitanti, dove il dettaglio tematico richiesto dagli ipercubi è inferiore, rispetto a regioni piccole come la Valle d'Aosta o il Molise, dove il dettaglio richiesto è molto superiore.[8]

[8] Uno degli obiettivi delle sperimentazioni in corso è quello di determinare la numerosità e l'allocazione del campione tale da garantire il maggior numero di ipercubi 'di qualità' previsti dal censimento permanente. Non è a priori escluso che, ad esempio per piccole regioni, si debba ricorrere al cumulo del campione su tre/cinque anni per qualche ipercubo, al fine di conseguire la necessaria accuratezza delle stime.

Prospetto 3.3 - Calendario di diffusione degli ipercubi derivati dalla D- Sample

ANNO DI DIFFUSIONE	Anno di riferimento (1 Gennaio) per dimensione dell'area di output		
(31 Dicembre)	> 100.000	da 35.000 a 100.000	< 35.000
2016	2016		
2017	2017		
2018	2018	2017	
2019	2019	2018	
2020	2020	2019	2018
2021	2021	2020	2019
2022	2022	2021	2020
2023	2023	2022	2021

3.2.3 Tecniche di rilevazione e organizzazione della D- Sample

Il campione di famiglie da intervistare sarà estratto dai registri anagrafici (LAC/ ANPR) riferiti al 1 gennaio di ciascun anno di rilevazione. Alle famiglie campionate l'Istat invierà una lettera informativa nella quale saranno specificati gli obiettivi della rilevazione, la normativa di riferimento e quella sulla tutela della privacy, l'invito ad auto-compilare il questionario, l'indirizzo internet e le relative credenziali per l'accesso alla compilazione web del questionario. Diversamente da quanto accaduto nel censimento del 2011 non è previsto il recapito alle famiglie del questionario in forma cartacea. Infatti la tecnica di rilevazione prevede una raccolta dei dati totalmente basata su acquisizione di questionari esclusivamente elettronici.

Le famiglie potranno restituire il questionario attraverso l'auto-compilazione via web in un intervallo di tempo durante il quale l'Istituto invierà alle famiglie uno o più solleciti con l'obiettivo di massimizzare la risposta diretta. In alternativa i cittadini potranno recarsi nei Centri Comunali di Raccolta (CCR) per la compilazione web assistita da rilevatori o per informazioni sulla rilevazione. Solo nei casi di mancata risposta diretta da parte della famiglia campionata saranno impiegati i rilevatori comunali che effettueranno a domicilio le interviste con tecnica CAPI necessarie al recupero dei questionari mancanti.

La rete di rilevazione sul territorio potrà essere incentrata sugli Uffici Comunali di Censimento (UCC) aventi modalità di costituzione e funzionamento simili a quelle adottate per il censimento del 2011. Essi dovranno avere caratteristiche di funzionamento continuativo nel tempo e curare la selezione, il reclutamento e la gestione dei rilevatori, nonché quella dei CCR. Tenuto conto dei domini territoriali stabiliti dal disegno campionario della D – Sample, i comuni con popolazione residente non superiore a 5.000 abitanti dovranno costituire UCC in forma associata.

Tenuto conto della dimensione campionaria annuale (circa 1,5 milioni di famiglie e convivenze), il fabbisogno standard medio annuo di rilevatori è stimato nella misura di 1 rilevatore ogni 550-600 unità di rilevazione.

Come già avvenuto per il censimento del 2011, tutte le operazioni di rilevazione saranno controllate dall'Istat e gestite dalla rete di rilevazione mediante un Sistema di Gestione della Rilevazione (SGR) accessibile via web dagli operatori della rete territoriale.

Poiché la tecnica di rilevazione è paperless, non è necessaria la fase di acquisizione dei questionari cartacei. Pertanto, terminata la rilevazione sul campo, si potrà subito passare alla fase di controllo e correzione.

- La D- Sample è basata su liste di famiglie residenti iscritte nel registro di popolazione al 1 gennaio dello stesso anno di riferimento della rilevazione. Ogni anno saranno coinvolti nella rilevazione: tutti i comuni con almeno 20.000 abitanti per un campione di famiglie estratto da LAC/ANPR e significativo a livello di aree di censimento;
- tutti i comuni tra 5.000 e 20.000 abitanti per un campione di famiglie estratto da LAC/ANPR;
- tutti i comuni sotto i 5.000 abitanti per un campione di famiglie estratto da LAC/ANPR, proporzionale alla propria popolazione rispetto a quella dell'agglomerato di appartenenza.

Poiché l'acquisizione annuale di LAC/ANPR si potrà concludere non prima del il 31 gennaio, l'estrazione del campione potrà essere effettuata nel mese di febbraio. Seguirà la spedizione postale delle lettere informative alle famiglie campionate che potrà concludersi presumibilmente entro marzo. Durante la prima fase di rilevazione sul campo alle famiglie saranno concesse 4 settimane per rispondere autonomamente via Internet o recandosi presso i CCR. Quindi all'inizio di maggio i rilevatori comunali potranno iniziare il recupero delle mancate risposte concludendolo entro lo stesso mese. I processi di controllo e correzione potranno essere terminati entro settembre e quelli di produzione delle stime, approntamento degli ipercubi e loro diffusione tramite data warehouse entro la fine dell'anno di riferimento dei dati.

Prospetto 3.4 - Cronogramma del ciclo tipo della D-sample

Attività	Tempi
Acquisizione LAC	Entro gennaio
Estrazione del campione e predisposizione lista	Entro febbraio
Spedizione lettera alle famiglie	Entro marzo
Risposta autonoma delle famiglie	Mese di aprile
Recupero mancate risposte	Mese di maggio
Controllo e correzione	Da giugno a settembre
Produzione e diffusione stime	Entro dicembre

CONCLUSIONI

L'obiettivo generale del censimento permanente, di produrre annualmente i tradizionali dati censuari a livello comunale e sub-comunale attraverso il massimo uso dell'informazione reperibile dalle fonti amministrative e l'impiego di indagini campionarie a rotazione, può essere raggiunto nel rispetto dei vincoli vigenti sia a livello nazionale che internazionale utilizzando le metodologie, tecnologie e soluzioni organizzative descritte nei paragrafi precedenti.

La nuova strategia può produrre:

- un considerevole abbattimento dei costi, stimabile in circa il 40 per cento in meno rispetto allo stanziamento censuario del 2011;
- un potenziamento delle attività di controllo e correzione dei dati, sfruttando le potenzialità offerte dall'integrazione di molteplici fonti di dati;
- una riduzione e diluizione nel tempo del carico organizzativo per le amministrazioni pubbliche e in particolare per i comuni;
- il rafforzamento in modo duraturo delle strutture del Sistan.

Le condizioni principali affinché il censimento permanente possa consentire il raggiungimento degli obiettivi esposti sono le seguenti:

- occorre dare corso alle progettate rilevazioni campionarie sperimentali nel biennio 2014-2015, in modo da anticipare le possibili criticità e valutare compiutamente le possibili alternative per pervenire a una definizione conclusiva delle metodologie, delle tecnologie e dell'organizzazione del censimento permanente;
- devono poter essere avviati a partire dal 2016, con data di riferimento il 1° gennaio, i primi cicli delle indagini C-sample e D-sample; con questa tempistica, infatti, sarà possibile completare il primo ciclo quinquennale delle rilevazioni campionarie in tempo utile a produrre, con riferimento al 2021, gli ipercubi di dati richiesti dal citato Regolamento dell'Unione europea.

Nondimeno, già a partire dal 2017, per i domini territoriali oltre i 100.000 abitanti, dal 2018, per i domini territoriali compresi fra i 35.000 e i 100.000 abitanti e dal 2019 per i domini territoriali sotto i 35.000 abitanti si avrà un aumento progressivo di dati censuari annuali da diffondere in modo da offrire al Paese informazioni rilevanti per la programmazione regionale e locale e per la valutazione *ex ante* ed *ex post* delle politiche pubbliche.

ALLEGATO 1
IL SISTEMA DI INTEGRAZIONE DI MICRODATI (SIM) DEMOGRAFICI, ECONOMICI E SOCIALI

1. Aspetti generali

L' esperienza di utilizzo di fonti amministrative a fini statistici è iniziata in Istat negli anni Novanta con la realizzazione progressiva del sistema ASIA, dei registri statistici dedicati alle unità economiche. L'esperienza si è estesa dal 2009 in poi alle di fonti focalizzate sull'entità individuo nei suoi aspetti demografici e sociali quali sesso, età, cittadinanza, composizione familiare, luogo di residenza, luogo di nascita e partecipazione al mercato del lavoro.

L'acquisizione delle Liste Anagrafiche Comunali (LAC) a fini censuari e degli archivi di natura fiscale, previdenziale e assicurativa, nei quali l'individuo è presente nella doppia veste di soggetto dichiarante (in quanto legato da relazioni di titolarità o partecipazione in unità economica) e di soggetto dichiarato (in quanto lavoratore, studente, pensionato eccetera), ha consentito di analizzare le caratteristiche demografiche e sociali di segmenti di popolazione sempre più ampi. In questo modo è stata avviata la realizzazione del Sistema di Integrazione di Microdati (SIM) che rappresenta uno dei principali progetti caratterizzanti le linee strategiche dell'Istituto per il triennio 2013 – 2015 (Progetto Statistico 2015), con l'obiettivo di integrare e rendere utilizzabili a fini statistici le informazioni sugli individui presenti nelle diverse fonti anagrafiche e non anagrafiche.

Il processo di integrazione tra gli archivi esistenti ha consentito di realizzare una vista unificata e riconciliata del patrimonio informativo delle diverse fonti di input mediante l'analisi delle connessioni logiche esistenti tra di esse.

2. Il modello concettuale

La realizzazione del SIM si è sviluppata attraverso diverse fasi di modellazione concettuale.

In primo luogo sono stati definiti gli schemi origine svolgendo le seguenti attività:

Successivamente si è definito lo schema globale attraverso le seguenti operazioni:

- stabilire i punti di connessione logica tra le diverse sorgenti;
- definire i criteri di armonizzazione delle singole fonti rispetto ad attributi comuni;
- individuare le regole e le modalità di integrazione tra i dati.

La disponibilità di un codice d'identificazione univoco della persona fisica integrato con altre caratteristiche (quali cognome e nome, data e luogo di nascita)

ha consentito livelli di riconoscimento tra le fonti organizzate a livello di individuo superiori in media al 98 per cento.

La presenza di archivi che registrano il legame diretto tra i codici di identificazione delle unità giuridiche e le singole persone ha consentito il riconoscimento, anche in questo caso con livelli superiori al 98 per cento, degli individui legati da rapporti di *partnership* o di lavoro con unità economiche.

Inoltre la disponibilità di informazioni relative alla localizzazione sul territorio sia a livello di persona fisica (indirizzo di domicilio fiscale, indirizzo di residenza), sia a livello di unità, congiuntamente all'applicazione di tecniche di linkage e di georeferenziazione degli indirizzi rende possibili analisi relative al luogo dove si dimora abitualmente e al luogo dove si svolge l'attività di studio o di lavoro per un ampio sottoinsieme di individui (lavoratori e studenti).

Dall'insieme delle relazioni definite all'interno dello schema globale sono state determinate, per i diversi ambiti di studio, delle schematizzazioni intermedie nelle quali è specificata, in termini di relazione e di attributi, la rappresentazione concettuale della porzione della realtà oggetto di interesse e in cui sono formalizzate le regole di interpretazione e le metodologie di trattamento dell'insieme delle informazioni provenienti dalle diverse fonti e relative allo stesso soggetto.

L'attività di integrazione per fini statistici di informazioni di origine amministrativa presenta elementi di forte complessità legati ai seguenti fattori:

1. non sempre si dispone di una conoscenza totale dei contenuti informativi delle singole fonti e soprattutto dei processi che le hanno prodotte;
2. le fonti di input sono soggette ad un costante cambiamento sia rispetto ai contenuti che ai sistemi classificatori adottati;
3. alcune fonti sono legate alle entità di interesse in maniera indiretta e a volte le relazioni che legano le unità logiche che esse rappresentano e la popolazione statistica sono complesse;
4. poiché i dati delle fonti sono in generale tra di loro indipendenti, le informazioni in esse contenute possono essere reciprocamente incoerenti;
5. le varie fonti possono presentare disallineamenti semantici;
6. il riferimento temporale dei dati è spesso differente tra le fonti;
7. gli identificatori riferibili nelle varie fonti ad uno stessa entità possono presentare problemi di correttezza formale che causano falsi casi di mancato accoppiamento.

3. I processi di produzione e le strutture dei dati

Il modello concettuale descritto nel paragrafo precedente è rappresentato nella figura seguente in termini di processi elaborativi e di strutture dati.

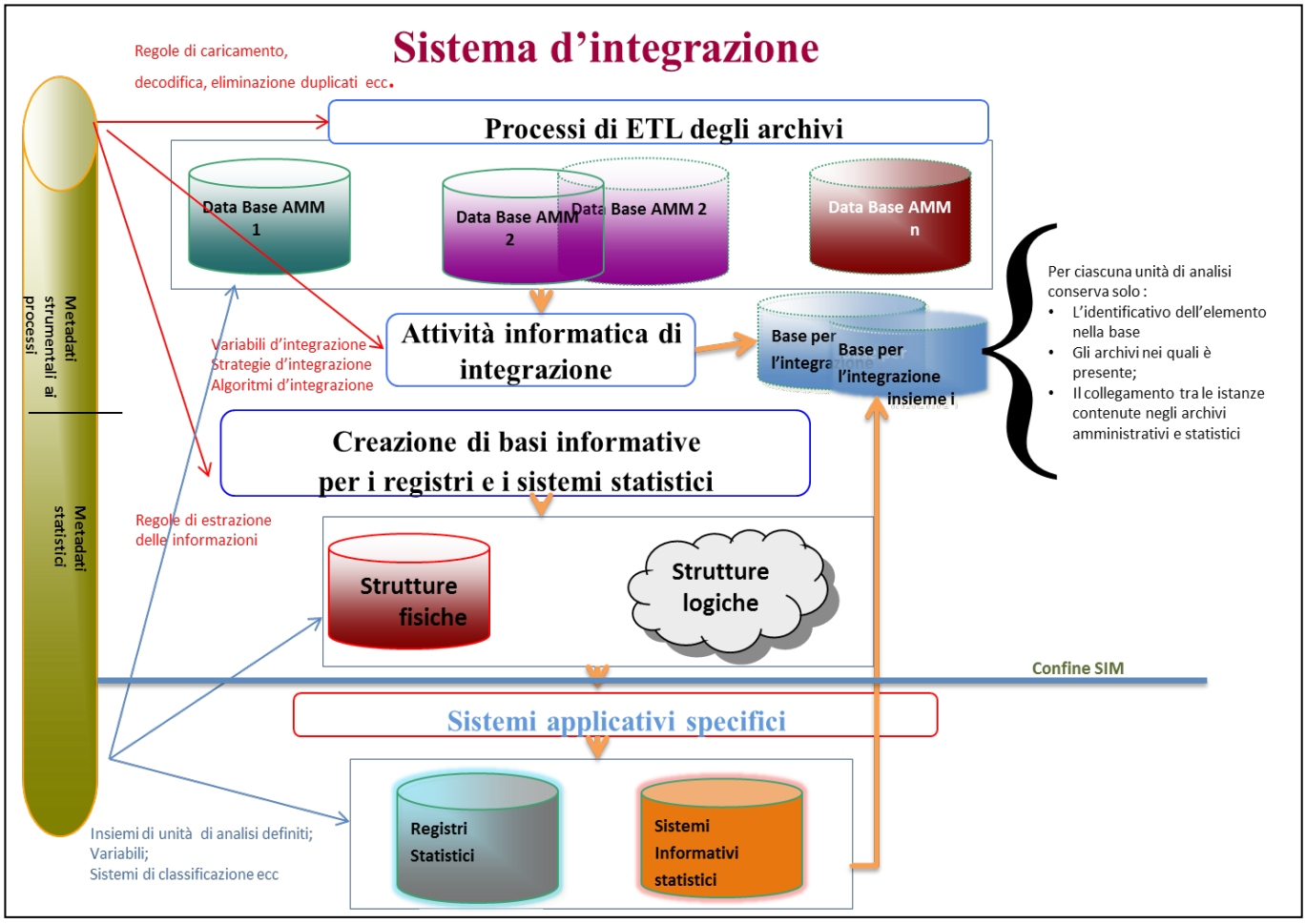
Nel modello sono individuati tre macro-processi:

- il processo di popolamento (ETL) dei Data Base degli archivi amministrativi;
- il processo d'integrazione;
- il processo di creazione di basi informative per i registri e i sottosistemi statistici di microdati integrati.

Il processo di popolamento viene disegnato a partire dalla definizione concettuale degli schemi origine. Esso fa riferimento all'insieme di operazioni di trasformazione

delle informazioni finalizzate a rendere i dati provenienti da sorgenti diverse omogenei tra di loro e aderenti alla logica applicativa del sistema d'integrazione. Gli archivi di input sono organizzati in una o più strutture dati definite sulla base degli insiemi di unità di analisi che in essi sono rappresentate. Ogni istanza (riga) di una struttura dati che rappresenta una determinata unità di analisi è individuata univocamente da una chiave interna il cui valore si conserva stabilmente nel tempo sulla base della immutabilità dei valori di un insieme di variabili scelte come caratterizzanti dell'elemento stesso. Tra le diverse strutture dati relative ad una determinata fonte sono definiti dei legami fisici che rappresentano le connessioni logiche esistenti tra le diverse tipologie di popolazioni in essa rappresentate. Anche il valore di questi legami si mantiene stabile nel tempo a parità delle caratteristiche della relazione stessa.

Figura 1: Sistema di integrazione per SIM



Il processo d'integrazione fa riferimento ai procedimenti di linkage veri e propri e applica le regole d'integrazione le quali definiscono:

- le strategie d'integrazione come ad esempio l'eventuale suddivisione del processo in più passaggi elaborativi, l'insieme di variabili da utilizzare ad ogni passaggio, le eventuali variabili di blocco eccetera;

- gli algoritmi di confronto da applicare;
- le metodologie di accettabilità dei collegamenti creati.

Il processo d'integrazione genera le "basi per l'integrazione", cioè strutture di dati intermedie che sono alimentate, ad ogni iterazione del processo, da nuove informazioni provenienti dalle fonti di input e rispetto alle quali le variabili di integrazione contenute nei diversi data base amministrativi e relative a uno specifico insieme di unità di analisi sono confrontate; la loro strutturazione fisica, pur seguendo una logica comune, dipende dal particolare sottosistema di integrazione a cui si riferiscono. In generale ciascuna istanza di una base per l'integrazione memorizza:

- l'archivio di provenienza;
- Il codice identificativo dell'unita di analisi nel sottosistema di integrazione;
- Il progressivo dell'unità di analisi all'interno dell' archivio di provenienza ;
- Il valore della variabile d'integrazione presenti nell'archivio per quell'elemento;
- Il riferimento temporale.

Scopo principale delle basi per l'integrazione è garantire la conservazione nel tempo del codice d'identificazione di una determinata unità elementare all'interno del sottosistema integrato e di mantenere i collegamenti, definiti attraverso il processo di linkage, tra le diverse fonti.

La creazione di basi informative per i sistemi statistici è uno strato software che produce strutture dati costituenti il punto di partenza di sistemi applicativi specifici. L'organizzazione fisica di tali strutture dipende dalle caratteristiche della porzione di realtà oggetto di studio, così come da esse e dagli obiettivi specifici che s'intendono raggiungere dipendono le regole di estrazione delle informazione dal SIM. Tali regole devono esplicitare:

l'insieme o sottoinsieme statistico d'interesse;

i principi di eleggibilità degli elementi nell'insieme;

gli archivi da referenziare e per ciascuno di essi l'insieme delle variabili;

le modalità di estrazione delle informazione in relazione al tempo (es. i valori più frequenti o più recenti rispetto ad un intervallo temporale, tutti i valori che ricadono in un intervallo temporale, i valori più vicini ad un determinato istante temporale ecc.).

Dai sistemi applicativi specifici, consistenti nell'applicazione di regole volte a determinare l' effettiva appartenenza delle unità estratte alla specifica popolazione d'interesse e nella implementazione di metodologie di stima o di scelta tra l'insieme di informazioni omogenee disponibili , deriva la creazione di elenchi di unità elementari con le loro caratteristiche fisse o variabili nel tempo che costituiscono i registri o i sistemi informativi statistici. Questi ultimi, avendo origine da un sistema d'integrazione in cui ogni unità è univocamente determinata, possono costituire a loro volta l'input di altri sistemi informativi specifici.

Trasversalmente ai macro-processi in cui si sviluppa il SIM, si colloca il Sistema dei Metadati che è suddiviso in due componenti:

- metadati statistici intesi nel senso tradizionale d'informazioni organizzate che descrivono i contenuti delle diverse fonti di input ed dei sistemi integrati ed i loro livelli di qualità;
- dati strumentali ai processi intesi come una serie di informazioni strutturate che non solo documentano ma anche regolano il comportamento dei sistemi applicativi al fine di garantire un elevato grado di flessibilità nella gestione dei

dati, nella modellazione degli applicativi e nelle logiche di elaborazione.

Sulla base della tipologia dei dati amministrativi disponibili, allo stato attuale, sono attivi nel SIM sette sottosistemi, alcuni dei quali alimentano già da tempo la produzione dei registri statistici dell' Istituto :

1. il sottosistema integrato delle unità economiche, relativo alle persone fisiche e giuridiche che possiedono una partita IVA (unità economiche);
2. il sottosistema di integrazione degli individui ;
3. il sottosistema di integrazione dei luoghi degli individui, nel quale sono confrontate le informazioni registrate nelle diverse fonti e relative alla localizzazione sul territorio delle persone fisiche;
4. il sottosistema integrato dei luoghi delle unità economiche, che costituisce la base informativa del Registro statistico delle Unità Locali delle Imprese e per il quale è previsto un ampliamento rispetto ad altre tipologie di unità quali istituzioni pubbliche e istituzioni non profit;
5. il sottosistema integrato delle relazioni tra unità, come le relazioni Impresa/Impresa (Eventi di trasformazione), Impresa/Unità giuridica, Impresa/Unità locali, Impresa/Gruppo di imprese;
6. il sottosistema di integrazione individuo/Unità economiche, che è alla base del sistema informativo dell'occupazione;
7. il sottosistema integrato che descrive le relazioni tra individui.

Nel corso del 2013 e per i primi mesi del 2014 è programmata l'acquisizione di 226 forniture relative complessivamente a 125 archivi. Attualmente 58 di essi sono soggetti ai processi di trattamento, di validazione e di integrazione dei dati del SIM. La tavola seguente riporta le principali caratteristiche dimensionali di ciascun sottosistema del SIM finora realizzato.

Tavola 1 – Numero di archivi trattati nei sottosistemi del SIM

Sottosistemi SIM	N. di archivi	Somma delle unità presenti (milioni)	Percentuale
Unità economiche	45	74,9	77,6
Individui	40	366,5	69,0
Luoghi degli individui	12	163,2	20,7
Luoghi delle unità economiche	23	30,7	39,7
Relazioni tra unità economiche	17	25,0	29,3
Relazioni tra individui e unità economiche	28	370,0	48,3
Relazioni tra individui	5	36,8	8,6
Numero complessivo di archivi in SIM	58		

ALLEGATO 2
LA QUALITÀ DELLE FONTI ANAGRAFICHE

In occasione del 15° Censimento generale della popolazione l'Istat ha dato l'avvio con la rilevazione delle LAC all'impiego statistico su scala nazionale dei dati anagrafici individuali che hanno consentito: a) l'invio dei questionari alle famiglie; b) la tracciatura delle attività dei rispondenti e degli operatori censuari; c) il confronto censimento/anagrafe contestuale all'attività censuaria; d) la determinazione dell'ammontare e della struttura della popolazione legale in tempi molto più rapidi rispetto ai censimenti passati.

La disponibilità delle LAC consente anche di disporre di un ampio patrimonio informativo, in particolare sulla qualità dei dati anagrafici, per calibrare in maniera efficiente ed efficace le metodologie per la realizzazione del censimento permanente.

Dai controlli quantitativi e qualitativi a cui sono stati sottoposti i dati di fonte anagrafica e dalle risultanze censuarie, è emerso che il divario tra il dato anagrafico comunale e il dato censuario, talvolta di non modesta entità, è da imputare:

- alla diversa natura delle due fonti delle quali l'anagrafe risente dei tempi e delle modalità delle procedure amministrative basate sulle dichiarazioni rese dai cittadini, mentre il censimento è un'operazione statistica condotta in un periodo di tempo relativamente breve da diversi attori sul territorio con l'obiettivo di determinare l'ammontare della popolazione abitualmente dimorante in ciascun comune ad una prefissata data;
- al livello di accuratezza, consistenza, completezza e non ridondanza con cui l'amministrazione comunale e i suoi servizi demografici provvedono alla tenuta del registro anagrafico;
- all'impegno dei rilevatori e degli altri operatori comunali, nonché dei rispondenti nel garantire l'aderenza alla realtà dei dati rilevati al censimento.

Analizzando la qualità dei registri anagrafici comunali, è emerso che essa viene influenzata da diversi fattori:

- la dimensione demografica del comune;
- la sua localizzazione territoriale;
- i controlli e i vincoli di qualità imposti dai software utilizzati per l'acquisizione e la gestione delle pratiche anagrafiche rese dai cittadini.

Nei paragrafi seguenti si forniscono elementi di valutazione della qualità dei dati contenuti nelle LAC, dei tempi di loro acquisizione telematica da parte dell'Istat, degli ulteriori miglioramenti attesi dalla realizzazione dell'ANPR.

1. La qualità delle LAC

Nel 2011 l'Istat ha eseguito per la prima volta la rilevazione totale delle LAC con univoco riferimento dei dati al 31.12.2010. Una seconda acquisizione è stata effettuata in concomitanza con la data di riferimento del censimento (8.10.2011), al fine di aggiornare i dati presenti sul Sistema di Gestione della Rilevazione (SGR) con le variazioni anagrafiche verificatesi rispetto alla precedente rilevazione e consentire agli

UCC di effettuare il confronto tra dati anagrafici e risultanze censuarie. Queste due acquisizioni sono state possibili a seguito di una estesa sperimentazione avvenuta nel corso del 2010 (1.850 comuni con 30 milioni di individui registrati in anagrafe) che aveva l'obiettivo di definire, sviluppare e consolidare standard tecnici sostenibili affinché gli uffici di anagrafe potessero produrre in formato elettronico archivi formalmente e qualitativamente validi e trasmetterli all'Istat in tempi relativamente brevi. Questi standard sono rimasti invariati per le due rilevazioni del 2011, così come per l'ultima acquisizione LAC condotta nei primi mesi del 2013 con riferimento al 1 gennaio, secondo quanto previsto dal Programma Statistico Nazionale vigente.

L'insieme di variabili inserite nel tracciato record delle LAC ha permesso e permette tutt'ora di assolvere a più funzioni: gestione del processo di acquisizione, conseguimento degli obiettivi istituzionali legati all'uso dei dati anagrafici, verifiche della qualità inter e intra variabili per singolo individuo o per loro aggregazione in famiglie o convivenze.[1]

Per ciascuna rilevazione i dati acquisiti dai comuni hanno tutti la stessa data di riferimento, in modo da minimizzare le duplicazioni di famiglie e individui tra comuni (circa 100 mila individui sul totale dei comuni per ciascuna delle acquisizioni).

Prospetto 1 – Le rilevazioni delle LAC finora effettuate dall'Istat

Data di riferimento dei dati di LAC	Comuni interessati	Comuni rispondenti	N. individui acquisiti
31.12.2009 (a)	1.850	1.746	31 milioni
31.12.2010	8.093	8.092	61 milioni
08.10.2011 (b)	2.510	2.437	49 milioni
01.01.2013	8.092	8.001	61 milioni

(a) Rilevazione sperimentale effettuata nel 2010.
(b) solo i comuni con almeno 5.000 abitanti e i comuni della Provincia Autonoma di Bolzano.

Nel complesso delle quattro rilevazioni, più del 93 per cento dei comuni (e le rispettive Software House che ne gestiscono gli aspetti tecnico-informatici) ha adempiuto a quanto richiesto dall'Istat producendo e inviando la LAC al primo tentativo, senza riscontrare alcuna difficoltà.

La tempestività con la quale i comuni hanno inviato i dati anagrafici è testimoniata dai cronogramma di seguito riportati e riferiti alle sole acquisizioni avvenute durante le fasi del 15° Censimento generale della popolazione. Essi sintetizzano il processo di acquisizione dei dati delle LAC. Per l'acquisizione del 2011 (Grafico 1) a distanza di 30 giorni dalla data di inizio della rilevazione erano disponibili più del 92 per cento delle LAC; per la seconda acquisizione circa il 98 per cento dei comuni con almeno 5.000 abitanti ha inviato la LAC nell'arco di un mese.

Nel corso del tempo la procedura di predisposizione e invio delle LAC è risultata

[1] Per ogni individuo iscritto nell'anagrafe di ciascun comune italiano sono riportati in formato elettronico le seguenti variabili: codice provincia, codice comune, tipo di residenza (famiglia o convivenza), codice della famiglia di appartenenza, codice della convivenza di appartenenza, identificativo dell'individuo in anagrafe, cognome, nome, codice fiscale, sesso, data di nascita, codice provincia di nascita, codice comune di nascita, codice stato estero di nascita, codice stato di cittadinanza, numero di componenti la famiglia anagrafica o la convivenza, relazione di parentela, stato civile, data di iscrizione in anagrafe e indirizzo di residenza.

sempre più di facile gestione per i referenti degli uffici di anagrafe. Essi hanno potuto avvalersi di una applicazione web (STARLAC), perfezionata dall'Istat nel corso delle diverse rilevazioni, che consente sia di gestire tutti gli aspetti tecnico organizzativi della rilevazione sia di garantire l'acquisizione controllata dei dati inviati. Per ogni file di dati inviato dai comuni l'applicazione verifica l'incidenza, a livello individuale e familiare, dei dati mancanti o errati per singole variabili e delle incongruenze fra gruppi di variabili logicamente connesse. L'applicazione considera acquisiti solo i file comunali che non violino soglie di accettabilità prestabilite e, ai comuni che hanno effettuato correttamente l'invio, restituisce le seguenti analisi di qualità:

- informazioni di sintesi; per ciascuna regola di controllo implementata è possibile visualizzare il valore assoluto e percentuale dei record che la violano;
- informazioni di dettaglio; per ciascuna regola di controllo violata sono visualizzabili i dati errati per singolo individuo, in modo da consentire ai responsabili degli Uffici di Anagrafe di intervenire direttamente sulle schede individuali anagrafiche per migliorare la qualità dei dati.

La possibilità di accedere tramite l'applicazione STARLAC alla reportistica sulla qualità dei dati è stata data ai comuni già in occasione della rilevazione sperimentale del 2010 ed stata confermata anche nell'ultima acquisizione del 2013. Le informazioni disponibili rappresentano un'importante occasione per l'avvio di un processo virtuoso volto al miglioramento della qualità dei dati anagrafici.

Il grafico 3, realizzato aggregando per gruppi tematici i numerosi controlli di qualità dei dati anagrafici eseguiti da STARLAC contestualmente al caricamento della LAC, illustra il progressivo miglioramento nel corso del tempo della qualità delle variabili acquisite.

Grafico 1 - Cronogramma dell'acquisizione delle LAC al 31.12.2010

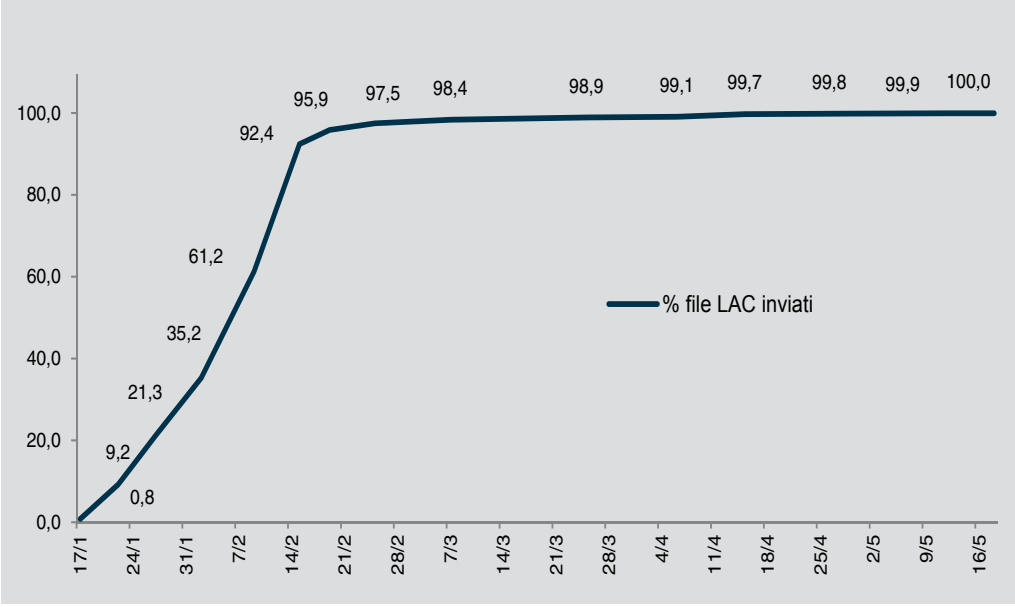


Grafico 2 - Cronogramma dell'acquisizione delle LAC al 08.10.2011

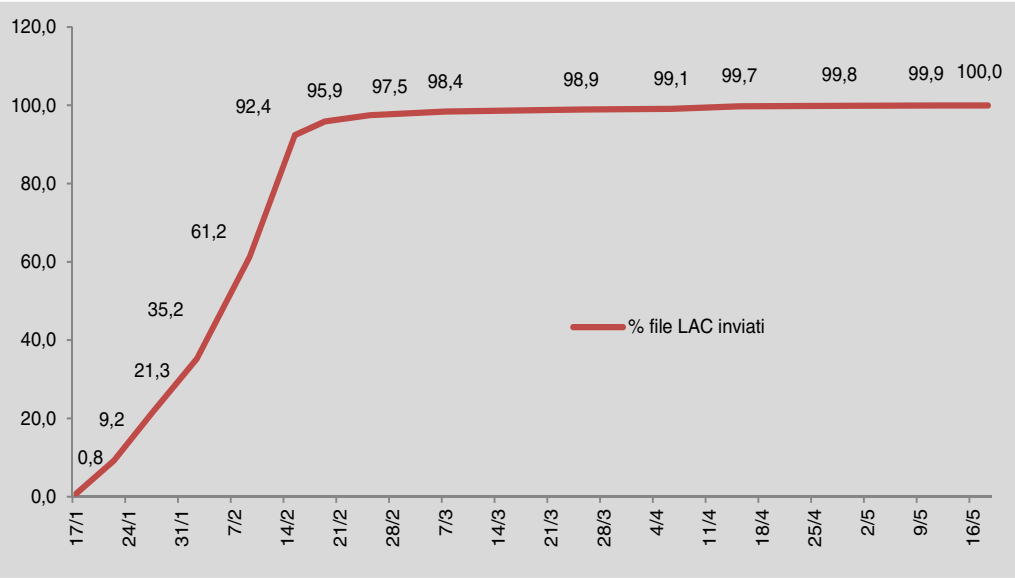
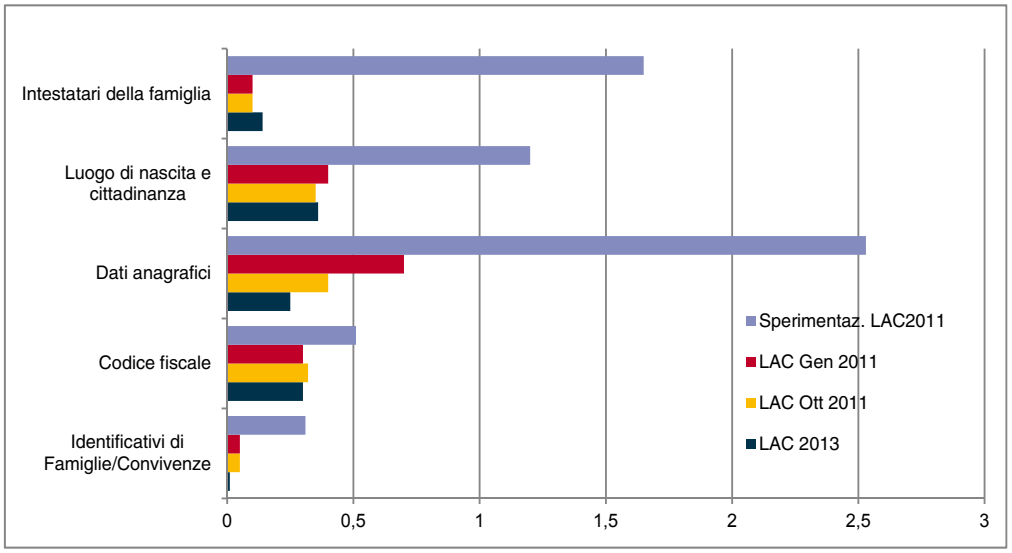


Grafico 3 – Percentuale di individui errati per Rilevazione LAC e per gruppi tematici di variabili sottoposte a controlli di qualità



A conferma dell'importanza dei dati di fonte anagrafica per il censimento permanente, un ulteriore riscontro è rappresentato dai risultati dell'analisi eseguita sulle principali variabili demografiche presenti in LAC (sesso, cittadinanza, data di nascita e relazione di parentela). Queste variabili sono state messe a confronto con le stesse informazioni riportate nei questionari di rilevazione dai rispondenti censiti come residenti e presenti in LAC. A riguardo, la tabella 1 mostra che:

- i valori mancanti nelle LAC variano tra il massimo di 4,66 per cento con riferimento allo stato civile e il minimo dello 0,97 per cento per la relazione di parentela;
- i tassi di concordanza tra valori presenti in LAC e nei questionari sono supe-

riori al 95 per cento per tutte le variabili di interesse, ad eccezione dello stato civile per il quale la concordanza è stata riscontrata nel 89 per cento dei casi.

Questi risultati mostrano l'elevato grado di affidabilità dei dati di LAC e confermano che l' elemento di incertezza delle anagrafi è rappresentato dai loro livelli di copertura rispetto alla popolazione "abitualmente dimorante".

Tabella 1 – Individui censiti presenti in LAC per esito del confronto con i questionari di censimento secondo le principali variabili anagrafiche – Valori percentuali

VARIABILE	Concordanza	Discordanza	Dato LAC mancante o fuori range	Totale
Sesso	96,3	0,1	3,6	100,0
Cittadinanza	96,0	0,2	3,8	100,0
Anno di nascita	95,4	1,2	3,4	100,0
Relazione di parentela (a)	95,4	3,7	0,1	100,0
Stato civile	89,0	6,4	4,7	100,0

(a) La relazione di parentela è stata valutata per i soli individui residenti in famiglia.

2. I miglioramenti di qualità attesi dalla realizzazione di ANPR

Entro il 2014, il sistema anagrafico nazionale, al momento strutturato in quattro partizioni (INA, anagrafe comunale, AIRE centrale e AIRE comunale) dovrà essere riordinato in un'unica anagrafe denominata Anagrafe Nazionale della Popolazione Residente (ANPR). Essa sarà istituita presso il Ministero dell'interno e determinerà un'accelerazione del processo di automazione amministrativa rendendo più efficiente la gestione dei dati anagrafi della popolazione e riducendone i costi.

Con il DPCM attuativo è in corso di definizione un piano per il graduale subentro dell'ANPR alle anagrafi comunali, da completare entro il 31 dicembre 2014. Fino alla completa attuazione di questo piano di unificazione, l'ANPR acquisirà automaticamente in via telematica i dati contenuti nelle anagrafi comunali assoggettate alla revisione anagrafica di cui all'art. 46 del vigente Regolamento anagrafico; il sistema sarà organizzato secondo modalità funzionali e operative che garantiscano la univocità dei dati individuali.

L'ANPR assicurerà la disponibilità dei dati anagrafici e dei servizi per l'interoperabilità con le banche dati tenute dai comuni per lo svolgimento delle funzioni di competenza e assicurerà alle pubbliche amministrazioni e agli organismi che erogano servizi pubblici l'accesso ai suoi contenuti.

Se i dati di ANPR saranno aggiornati in tempo reale, la possibilità di disporre all'inizio di ogni anno di un'unica fonte di dati a livello nazionale consentirà di eliminare le seguenti principali criticità riscontrate nelle LAC:

- i disallineamenti dovuti alle pratiche di cambio di residenza;
- le doppie iscrizioni in anagrafe dovute ai tempi di perfezionamento delle pratiche;
- le mancate cancellazioni per decesso e per espatrio documentate da specifico atto amministrativo.

Il passaggio dalle acquisizioni delle LAC a quella di ANPR con riferimento ad un a data prefissata determinerebbe effetti positivi anche sulla copertura del censimento. L'Istat ha quantificato che se in occasione del Censimento della Popolazione 2011 ci fosse stata ANPR in luogo delle LAC, sarebbero stati conteggiati circa 500 mila irreperibili in meno (in termini relativi una percentuale tra il 20 e il 25 per cento) causati dai mancati aggiornamenti e dai disallineamenti anagrafici che sono stati riscontrati nelle liste comunal, in particolare a causa dei cambi di residenza. Ulteriori e non meno rilevanti vantaggi derivanti da ANPR consisteranno:

- nella possibilità di anticipare le attività di conteggio della popolazione per il censimento permanente e per le altre rilevazioni Istat sulla popolazione grazie all'abbreviazione dei tempi di sua acquisizione rispetto a quelli, pur contenuti, delle LAC;
- nell'eliminazione degli errori sistematici di riclassificazione di alcune variabili commessi in fase di predisposizione dei file LAC da parte di alcune Software House;
- nel totale annullamento del problema del cambiamento dei codici identificativi in occasione della sostituzione del software di gestione dell'archivio di un comune.

ALLEGATO 3
L' ARCHITETTURA INFORMATICA PER LA GESTIONE DELLE RILEVAZIONI C-SAMPLE E D-SAMPLE

1. Introduzione

La messa in esercizio dei sistemi web a supporto dei censimenti del 2010 – 2011 ha portato allo sviluppo e al progressivo raffinamento di sofisticate applicazioni per la gestione della rilevazione e la compilazione online dei questionari censuari, caratterizzate da elevata scalabilità ed efficienza. Parallelamente si è consolidato anche un ampio bagaglio di conoscenze ed esperienze sulla gestione di infrastrutture hardware e software, in grado di far fronte a carichi elaborativi di grande entità, quali appunto quelli delle rilevazioni censuarie.

Ciò che ha differenziato in maniera sostanziale i censimenti dell'agricoltura del 2010 e dell'industria e servizi del 2012 da quello della popolazione ed abitazioni del 2011 è stata la differente modalità di gestione dell'infrastruttura hardware e software di base, basata esclusivamente su risorse (umane e strumentali) interne all'Istat nei primi due casi e sull'adozione di un'infrastruttura di web hosting per il terzo.

L'architettura informatica generale e gli applicativi sviluppati per i tre censimenti sono stati costruiti come sistema integrato , parametrizzato e standardizzato in grado di supportare tutte le attività legate all'acquisizione e in alcuni casi al controllo dei dati, minimizzando le attività di sviluppo di codice specifico e massimizzando il riuso di software.

In modo più specifico il sistema integrato di gestione del 15° Censimento della popolazione e delle abitazioni è composto da tre moduli: il portale di documentazione per la rete di rilevazione, il Sistema di Gestione della Rilevazione (SGR) e il sistema di acquisizione mediante questionario elettronico QUOL. Dal punto di vista applicativo esso si è dimostrato robusto e in grado di supportare in modo adeguato le varie fasi di rilevazioni multicanale, multi questionario, con una rete di rilevazione complessa e distribuita sull'intero territorio nazionale.

Premesso quindi che si desidera evitare, per quanto possibile , eccessivi cambiamenti sia dell'infrastruttura hardware, sia del software di base e specifico messi a punto nel corso di questi anni, si pone la questione di verificare l'adattabilità di quanto realizzato per i Censimenti del 2010 - 2011 al Censimento Permanente, tenuto conto delle sue caratteristiche generali e di quelle specifiche delle due indagini C-sample e D-sample.

I sistemi realizzati per i tre Censimenti dell'ultima tornata si sono basati su una infrastruttura centralizzata che ha consentito l'accesso e l'utilizzo dell'intero sistema a tutte le persone in possesso di un *personal computer* o *tablet* di ultima generazione, connessione a *internet* con *browser* a scelta tra i più diffusi sul mercato per

la navigazione, senza necessità di nessuna installazione aggiuntiva di *software* o particolari configurazioni del proprio dispositivo elettronico. Sebbene le infrastrutture centralizzate abbiano alcuni inconvenienti, quali la maggiore criticità di un corretto dimensionamento per garantire un'adeguata affidabilità del servizio, la scelta di questo approccio è stata motivata da ragioni organizzative, economiche, di complessità e consistenza.

Questo tipo di infrastruttura può essere ragionevolmente riproposta per la gestione della rilevazione D-sample che mantiene, pur se su scala ridotta, le stesse caratteristiche metodologiche e organizzative della rilevazione del censimento della Popolazione 2011, mentre risulta poco adatta o comunque insufficiente per la gestione della raccolta dei dati sul campo prevista nell'indagine C-sample. In questo caso, infatti, è necessario garantire al rilevatore la possibilità di acquisire i dati in formato digitale sul proprio dispositivo portatile anche in zone non coperte da rete di trasmissione dati. Per consentire ciò è necessario prevedere una modalità locale di acquisizione dati mediante lo sviluppo di software aggiuntivo da installare sui laptop o tablet, definire le modalità di trasferimento dei dati da server locale a server centrale e sviluppare le opportune procedure.

2. Soluzioni hardware per le rilevazioni Capi C-sample e D-sample

Come illustrato nel capitolo 3 del documento di base, la rilevazione C-sample deve essere condotta da rilevatori che non appartengono al personale comunale in quanto finalizzata alla determinazione degli errori di copertura del registro di popolazione (LAC/ANPR), mentre la rilevazione D-sample si avvale dei Comuni come organi intermedi di rilevazione . Questa differenza di organizzazione della rete porta necessariamente a considerare le due indagini come totalmente indipendenti, almeno dal punto di vista della dotazione della strumentazione tecnologica necessaria nella fase di raccolta sul campo.

A seguito dell'analisi del mercato corrente, sono state identificate due tipologie di dispositivo potenzialmente utilizzabili per le rilevazioni: tablet e PC portatili con touch screen. Di seguito vengono elencate le caratteristiche tecniche dei dispositivi disponibili al momento sul mercato che garantiscono la compatibilità con le specifiche funzionali e tecniche previste:

Tablet	PC Portatili
• Sistema operativo Android • Sistema operativo Windows 8 (ibrido) • Schermo da 7" a 10"	• Sistema operativo Windows (7 o superiore) o Linux • Schermo da 10"-11" • Eventualmente convertibili (con touch screen separabile)

La seguente tabella, commentata nel seguito, riporta un confronto tra le tipologie dei dispositivi sulla base di alcuni dei requisiti che si ritengono significativi per la scelta.

	Tablet	PC Portatili
Costi	□	
Peso Maneggevolezza	□	
"Semplicità" dello sviluppo – riuso del software esistente		□
Sicurezza		□

Costi: i tablet attualmente disponibili sono in generale molto più economici dei PC portatili con touch screen.

Peso Maneggevolezza: i tablet sono più leggeri e maneggevoli dei PC; quindi più comodi da trasportare per i rilevatori. Va però considerato che gli schermi da 7" potrebbero rivelarsi troppo piccoli per alcuni utenti e uno schermo da 10" ridurrebbe il vantaggio dei tablet in termini di costi e manegevolezza.

"Semplicità" dello sviluppo – Riuso del software: il software utilizzato per gli ultimi censimenti è stato concepito per la distribuzione e l'utilizzo su web quindi è una *web application* che richiede un *application server* Java per essere eseguito. In quanto tale, non è portabile sui tablet. L'utilizzo dei tablet richiede quindi uno sviluppo applicativo ex-novo, con tecnologie totalmente diverse.

Sicurezza: il sistema operativo Android presenta ancora delle incertezze dal punto di vista della possibilità di prevedere diversi profili utente con limitazioni sulle funzionalità utilizzabili.

Nel caso della C-sample l'effort di nuovo sviluppo su tablet è piuttosto contenuto in quanto il questionario è composto sostanzialmente dalla tradizionale Lista A (dati identificativi dei componenti della famiglia) e da poche informazioni aggiuntive su ciascun componente della famiglia o della convivenza. Inoltre la C-sample si configura nel tempo esclusivamente come un'indagine CAPI, con un dimensionamento fissato a priori del numero dei dispositivi portatili necessario a condurre la rilevazione. In queste condizioni particolare rilevanza assumono fattori come il basso costo dei dispositivi, la loro leggerezza e maneggevolezza. Infine il tipo di dati raccolti rende meno vincolante la mancanza di strumenti consolidati di protezione dei dati sui tablet.

Al contrario nel caso della D-sample l'applicazione in locale potrebbe essere realizzata facilmente riutilizzando buona parte del codice prodotto per la generazione del questionario elettronico del censimento del 2011. Inoltre le interviste CAPI, essendo previste solamente per il recupero delle mancate risposte, sono destinate a diminuire nel tempo a favore dell'autocompilazione del questionario da parte delle famiglie e della sua acquisizione via internet. Infine la tipologia delle informazioni richieste nel questionario (personali e in parte sensibili) rende assolutamente necessaria l'adozione di adeguate misure per garantire la sicurezza e la protezione dei dati, sia quando questi si trovano sul dispositivo portatile, sia durante il trasferimento da dispositivo portatile a server centrale.

Le considerazioni fatte fanno propendere per adottare dei tablet con sistema operativo Android o Windows8 per la C-sample e dei Pc portatili con sistema operativo Windows7 o superiore o Linux per le interviste CAPI della D-sample.

3. L'infrastruttura tecnologica per la gestione delle rilevazioni C-sample e D-sample

A fronte della positiva esperienza maturata durante il 9° Censimento generale dell'industria e dei servizi 2, si prevede di confermare per il Censimento permanente della popolazione l'uso di un'infrastruttura completamente interna all'Istat (in house), con un'architettura del tipo *multi-tiers* già utilizzata nel 2011 per i servizi censuari online e basata su tre livelli:

FRONTEND-TIER: composto da "bilanciatori geografici", opportunamente dimensionati per sostenere gli accessi ai "server applicativi" secondo i requisiti previsti dall'Istituto per il censimento permanente.

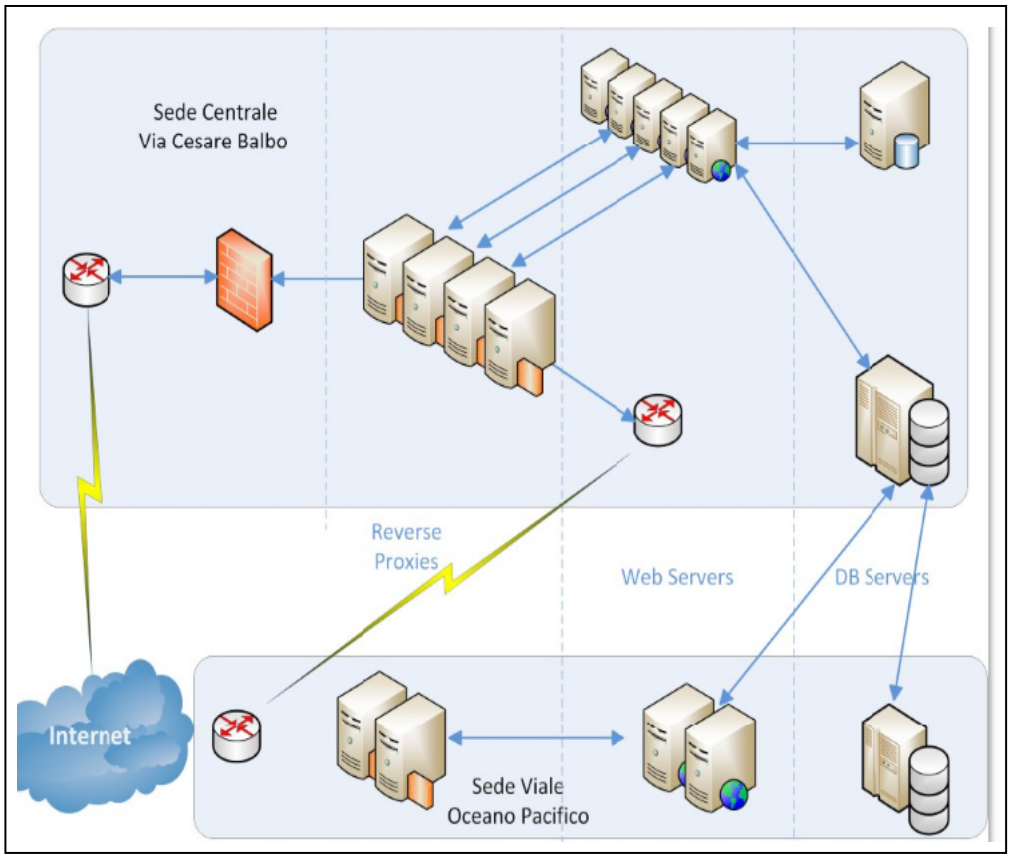
APPLICATION-TIER: costituito da una batteria di Server applicativi, opportunamente dimensionati per sostenere gli accessi alle componenti dinamiche e applicative dei servizi web sia per la raccolta che per la gestione censuaria, secondo i requisiti previsti dall'Istituto.

DATABASE-TIER: che rappresenta lo strato per la gestione della persistenza dei dati; particolare cura dovrà essere riposta nel trattamento dei dati sensibili se verranno richiesti nella rilevazione censuaria.

In particolare si prevede di utilizzare una serie di server di primo livello (*reverse proxy*), operanti in parallelo, che ricevono le richieste Internet attraverso il firewall e inoltrano le chiamate ai server di secondo livello (*web server Tomcat*) con un meccanismo di bilanciamento delle richieste.

A loro volta i web server di secondo livello evadono le richiese interrogando il database server Oracle (EXADATA). Tale infrastruttura è quella standard in Istituto e potrebbe essere potenziata attraverso l'inserimento di una coppia di sistemi per garantire un primo livello di *disaster recovery*.

L'infrastruttura proposta è riportata nella figura seguente:



Tra i problemi riscontrati nel passato censimento della popolazione risulta evidente la crisi del 9 ottobre 2011 che ha immediatamente costretto il fornitore (Telecom) a scalare l'architettura *hardware* e *middleware* per far fronte all'emergenza.[1] Nel giro di meno di 24 ore, quella che era nata come un'architettura di *hosting* è stata potenziata e trasformata in una *cloud* privata, con pile applicative dedicate, ma riconfigurabili tra il questionario online (QUOL) e il sistema di gestione della rilevazione (SGR) a seconda delle esigenze correnti.

L'infrastruttura proposta per il Censimento permanente, pur non essendo una *cloud* privata, supera le rigidità tipiche di una configurazione stabilita a priori e non modificabile. Infatti, in caso di carichi anomali sul sistema dovuti ad un eccesso di connessioni contemporanee non preventivato in fase di dimensionamento dell'infrastruttura, sarà possibile scalare sulla componente applicativa. La scalabilità orizzontale, intendendo con questo la possibilità di aggiungere nuovi server in parallelo per aumentare la potenza elaborativa, è piuttosto agevole e dipende solo dalla disponibilità al momento di nuovi server fisici. Relativamente alla scalabilità verticale, intendendo con questo la possibilità di potenziare le risorse hardware delle singole macchine, i server sono già dimensionati con un numero di risorse cospicue e certamente molto al di sopra della media.

I vincoli della pila applicativa, almeno per quanto riguarda la parte sistemi Linux, sono soddisfatti da: Java + procedure pl/sql, Webserver Tomcat, Dbms Oracle.

[1] È stato riconosciuto che i malfunzionamenti iniziali non erano in alcun modo provocati dal software prodotto da Istat.

La scalabilità sulla componente Db è sicuramente più complessa da realizzare, in quanto richiede l'acquisto di una coppia di macchine Exadata con relative licenze Oracle , macchine sulle quali risiedono la maggior parte dei database dell'Istituto. Tuttavia l'esperienza dei passati censimenti dimostra che il DBMS non è mai andato in sofferenza, neppure nelle ore di maggior carico di utenti connessi contemporaneamente. In ogni caso, Il dimensionamento puntuale e il perfezionamento delle configurazioni delle componenti architetturali saranno frutto del processo di stress test applicativo.

Al fine di soddisfare le esigenze specifiche di affidabilità e robustezza del servizio censuario, dovrà essere garantita la presenza di ulteriori tre ambienti :

AMBIENTE DI SVILUPPO: con dimensionamento adeguato alle sole esigenze funzionali disponibile sulla intranet, che rispecchia l'architettura dell'ambiente di esercizio.

AMBIENTE DI TEST: con dimensionamento adeguato alle sole esigenze funzionali, disponibile su internet per consentirne l'uso anche a soggetti esterni all'Istat e per supportare sia la fase di test dei sistemi sia la fase di formazione, che rispecchia l'architettura dell'ambiente di esercizio.

AMBIENTE DI DISASTER RECOVERY: ambiente replicato per garantire la business continuity del servizio censuario.

Lo scenario corrisponde sostanzialmente ad una replica della soluzione adottata dapprima per il censimento dell'agricoltura, quindi per la "coda" del censimento della popolazione e infine, con opportuni raffinamenti, per l'intero censimento dell'industria e dei servizi. Ne consegue che si tratta di una soluzione ormai ben consolidata e ampiamente sperimentata, soprattutto per quanto riguarda l'organizzazione della modalità di interazione tra i team di sviluppo e quelli dei sistemi.

Tra i punti di forza di questa opzione vi è senz'altro la semplicità nella gestione dei processi di popolamento dei DB e di tutte quelle attività di modifica ed elaborazione diretta sul DB di produzione (attraverso script realizzati ad-hoc e non previsti dal sistema SGR) le quali, come si evince dall'esperienza pregressa sul Censimento della Popolazione rappresentano una componente significativa ed ineliminabile delle attività di gestione del sistema.

Un altro indiscusso vantaggio riguarda la maggiore flessibilità e facilità del processo di patching e più in generale di deploy di nuove versioni del software.

Tra i vantaggi si deve inoltre annoverare la flessibilità delle tecnologie per lo sviluppo. Infatti, questa soluzione non impone vincoli sostanziali nella scelta delle tecnologie utilizzate per lo sviluppo, sia per quanto riguarda l'architettura del sistema (application server, DBMS, etc.) che per le specifiche librerie e framework usati nei vari componenti software. Questo aspetto consente il completo riuso del software esistente e lascia la possibilità di modifiche e/o aggiunte in caso di necessità future.

Di contro, l'unico svantaggio di una soluzione basata esclusivamente su risorse interne risiede nella impossibilità pratica di garantire i livelli di servizio tipici del web, in particolare il ben noto 24/7, che corrisponde, in caso di guasti o malfun-

zionamenti di qualsiasi natura, alla possibilità da parte dei tecnici di intervenire sui sistemi 24 ore su 24 e 7 giorni su 7.

Tuttavia la continuità operativa nell' ambito del censimento permanente va intesa come requisito stringente solo per i servizi che non prevedono la possibilità di operare off-line. In particolare per la rilevazione C-sample, gli intervistatori una volta scaricato l'itinerario di sezione potranno operare nella modalità off-line non solo nel caso in cui la copertura della rete per la trasmissione dati non sia disponibile, ma anche nel caso di indisponibilità dei servizi lato server gestiti direttamente dall'Istat. Questa considerazione risulta solo parzialmente applicabile per la rilevazione D-sample, poiché la compilazione spontanea delle famiglie sarà possibile solo tramite un applicativo web che necessita della presenza dell'infrastruttura server disponibile.

Comunque è da tenere in considerazione che l'Istituto sta affrontando a livello generaleil problema della continuità operativa come richiesto dall'AGID (Agenzia per l'Italia Digitale).

4. Il software per il questionario elettronico C-sample e D-sample

Il software di acquisizione sviluppato per il censimento della popolazione (QOL) e per quello dell'industria e servizi e censimento delle istituzioni non profit (QCIS) è stato progettato dall'inizio con l'idea di essere riutilizzato in diversi contesti. Il software presenta ormai un alto grado di maturità, provata sul campo: le due applicazioni sono state utilizzate per 6 tipologie diverse di questionari, con circa 20 diverse tipologie di domande; sono stati raccolti complessivamente 8.200.000 questionari con QOL e 700.000 con QCIS senza nessun problema di scalabilità o di sicurezza riscontrato in esercizio, nonostante la mole e la delicatezza delle operazioni. Infine il software presenta ormai un supporto completo per l'accessibilità, aspetto particolarmente importante per un servizio web da fornire direttamente ai cittadini e assolutamente non banale da garantire in presenza di tipologie di domande complesse (ad esempio tabelle) e comportamento dinamico (ad es. conferma sulle domande).

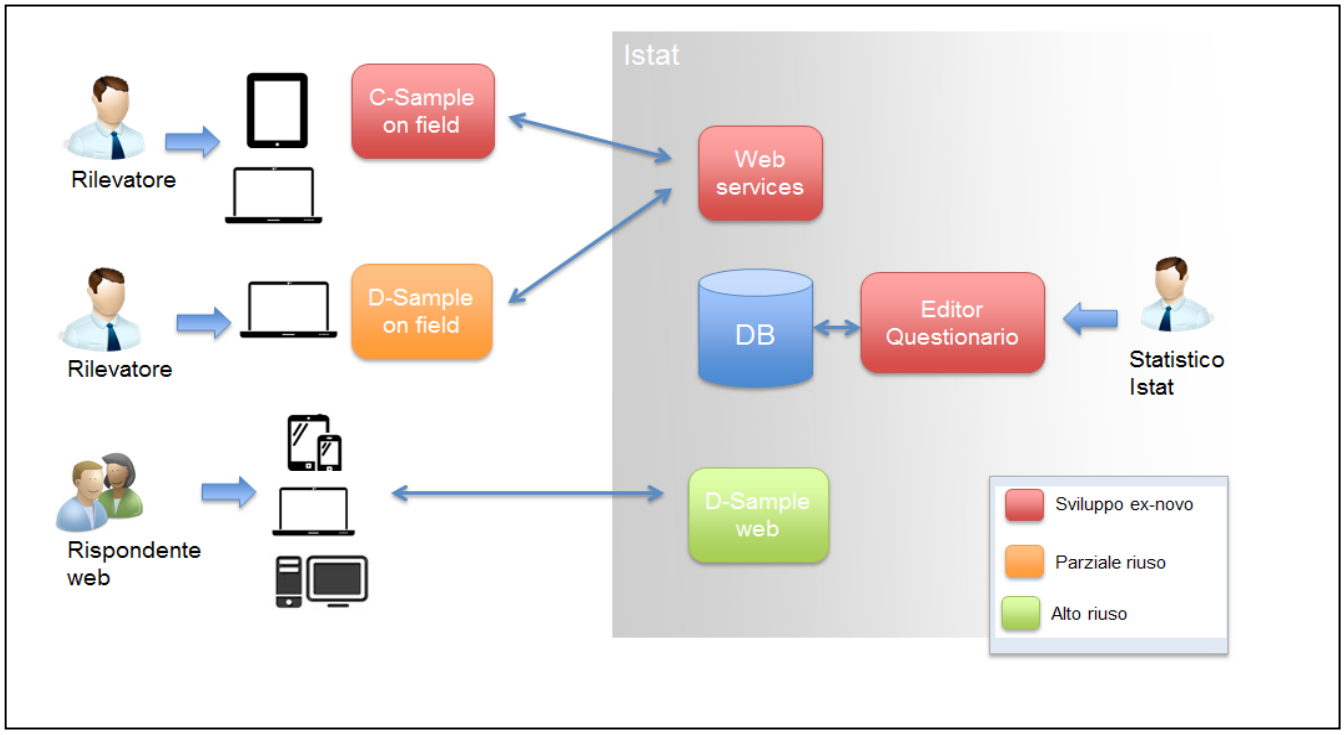
Sulla base di queste considerazioni si ritiene che il riuso di base del software esistente sia scontato, almeno per l'acquisizione D-Sample. Ulteriori sviluppi di software sono però necessari per coprire i seguenti aspetti:

- supporto C-Sample: anche se i dati raccolti sono un sottoinsieme del questionario del 2011, è necessario effettuare degli interventi per rendere l'applicazione auto consistente;
- modalità offline;
- comunicazione con server, salvataggio dati in locale e aggiornamento dell'applicazione da remoto;
- generalizzazione e realizzazione di un modulo per la definizione dei questionari. Nella versione attuale del software la definizione dei questionari è effettuata con un approccio "a basso livello", ovvero inserendo direttamente nel DB i metadati del questionario. Sebbene questo approccio si sia rivelato sostenibile anche per utenti non informatici, è evidente che il completamento della generalizzazione del software insieme ad una interfaccia utente per la definizione dei metadati

renderebbe più agevole il lavoro ad utenti statistici e più ottimizzato il processo. Inoltre, si aumenterebbe la disponibilità di software generalizzati per l'acquisizione potenzialmente utilizzabili dall'Istituto in altri contesti di produzione.

Considerando tutti gli aspetti analizzati, si elabora una proposta di architettura software rappresentata in figura e strutturata come segue:

- acquisizione C-Sample realizzata sia come app per tablet Android che come web app Java eseguita offline su web server locale distribuito su Pc portatile del rilevatore;
- acquisizione D-Sample basata sul progetto QOL-QCIS, ovvero web app Java eseguita in due modalità: online su web server centrale e offline su web server locale distribuito su PC portatili dei rilevatori;
- le due applicazioni comunicano con il server centrale mediante un livello di web services ;
- un componente "Editor del questionario" consente di definire i metadati del questionario attraverso un'interfaccia grafica.



La figura rappresenta i vari componenti architetturali, insieme agli utenti coinvolti e alle tipologie di dispositivo utilizzate. Per ogni componente dell'architettura viene rappresentato, con un diverso colore, il grado di riuso del software (nessuno - sviluppo ex-novo, parziale, alto).

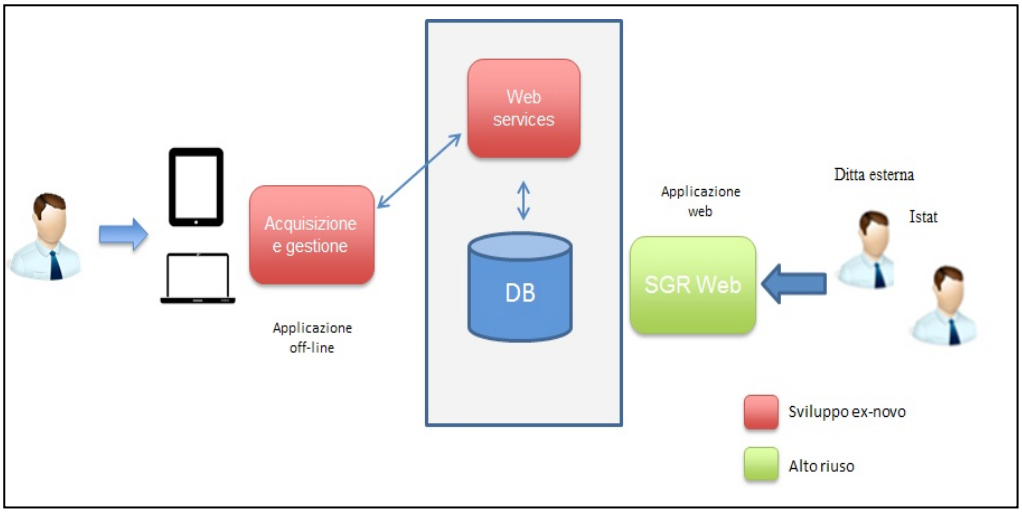
5. Il software per il monitoraggio C-sample e la gestione della D-sample

Le proposte architetturali riguardanti il sistema di acquisizione hanno un impatto considerevole sul Sistema di Gestione della Rilevazione (SGR). La scelta di dotare di tablet i rilevatori per la C-sample porta come conseguenza di dover garantire la possibilità di effettuare, tramite lo stesso strumento, le operazioni sul campo che sono loro richieste. Di fatto sarà, quindi, necessario sviluppare una sistema di gestione apposito per i rilevatori, strettamente integrato con il sistema di acquisizione, in modo che il rilevatore inserisca, oltre i dati del questionario, anche le informazioni necessarie al monitoraggio.

Accanto a questa nuova infrastruttura tecnologica deve comunque essere garantita un'applicazione che permetta il monitoraggio e le operazioni di back office. Per questo sistema si può riutilizzare quanto fatto per SGR dei passati censimenti. Tali software hanno il vantaggio di essere stati sviluppati nell'ottica della generalizzazione e quindi possono essere facilmente adattati alla nuove esigenze della C-sample. In particolare il sistema di gestione della rilevazione dovrà garantire all'Istituto un attento monitoraggio della ditta esterna che effettuerà la rilevazione; i classici report di monitoraggio saranno integrati con dei nuovi output maggiormente finalizzati al controllo degli operatori e corredati da indicatori di benchmark per la valutazione della bontà della rilevazione sul campo. Inoltre, l'uso dei tablet da parte dei rilevatori apre delle nuove vie al monitoraggio e potrebbe consentire, qualora fosse richiesta, la tracciatura non solo delle operazioni ma anche della localizzazione territoriale del rilevatore stesso al momento della compilazione del questionario. Di contro però l'applicazione su tablet non permette un monitoraggio in tempo reale, in quanto i dati sono scaricati sui server dell'Istituto solo a richiesta del rilevatore e questo avviene tipicamente a fine giornata lavorativa. Il sistema web, analogamente a quanto avvenuto nei passati censimenti, sarà utilizzato dalla ditta di rilevazione per effettuare tutte le operazioni di back office come ad esempio la creazione e l'assegnazione della rete di rilevazione.

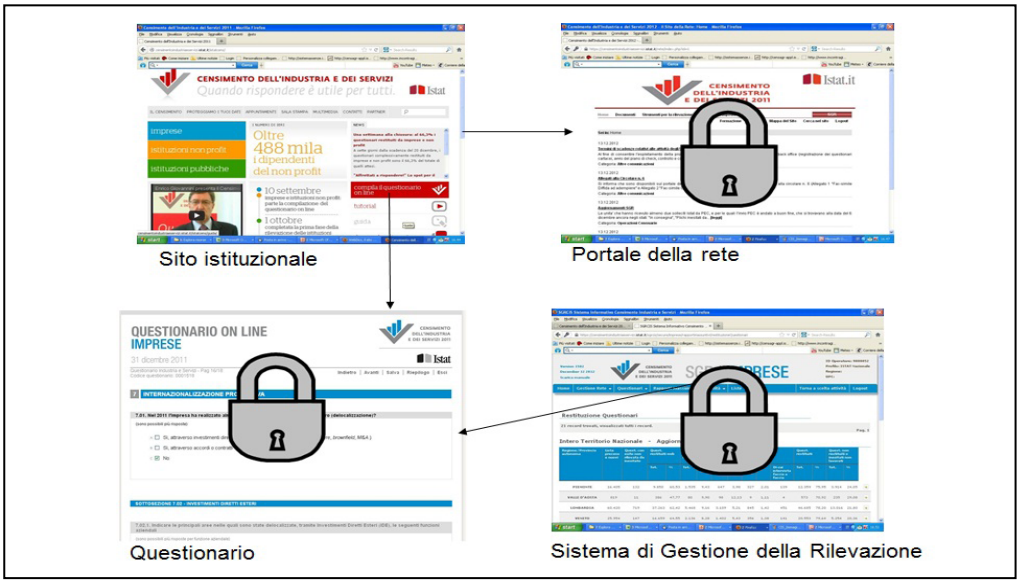
Il Sistema di Gestione della Rilevazione, utilizzato per i passati censimenti, dovrà quindi essere diviso in due sottosistemi ben distinti e con funzionalità e architettura diversa. Di fatto bisognerà prevedere un applicazione web che permetta un monitoraggio dell'andamento della rilevazione e la gestione di tutte le operazioni di back end tipiche della rilevazioni, nonché un'applicazione off-line che gestisca le operazioni sul campo esclusive del rilevatore. Analogamente a quanto fatto in altre rilevazioni, quali prezzi al consumo e forze di lavoro, il rilevatore scaricherà dal server i dati di partenza sul quale svolgere la rilevazione , come ad esempio l'itinerario di sezione.A fine giornata il rilevatore dovrà caricare i dati acquisiti localmente sul server centrale.

La figura sottostante, riprende lo schema architetturale mostrato per l'acquisizione e sottolinea la necessità dell'integrazione tra i sistemi di acquisizione e di monitoraggio e l'effort informatico necessario per intraprendere questa soluzione.



Per quanto concerne la D-sample, la proposta di utilizzare per l'acquisizione dei dati una applicazione web java che giri sia in modalità off-line che in modalità on-line garantisce un maggior riuso del software di gestione delle rilevazioni censuarie del 2010 - 2011. Infatti, in questo caso, può essere sufficiente sviluppare un'unica applicazione web che gestisca tutte le fasi della rilevazione. L'unico punto di attenzione deve essere posto sullo sviluppo dell'applicazione off line di acquisizione. Infatti, in questo caso si deve prevedere l'inserimento di alcune informazioni normalmente demandate al sistema di monitoraggio, analogamente a quanto proposto per la versione off-line su tablet.

Inoltre, essendo la D-sample presumibilmente gestita dai comuni, è utile garantire lo stesso livello di servizi offerto per il passato censimento della popolazione. La piattaforma informatica proposta è quindi basata su tre applicazioni web distinte e integrate tra loro, tutte ad accesso protetto e destinate alle diverse tipologie di utenze coinvolte: famiglie, rete di rilevazione e Istat. La figura seguente riporta a titolo esemplificativo la piattaforma informatica messa a disposizione per il censimento industria e servizi.



6. Certificazione del software e test per il dimensionamento dell'infrastruttura

Lo scopo primario della certificazione e dei test per il dimensionamento dell'infrastruttura è quello di dotare l'Istituto di un supporto specializzato nella revisione qualitativa e prestazionale dell'architettura hardware e software necessaria per la gestione informatica del Censimento permanente, in particolare prima che esso venga messo in esercizio e nel monitoraggio durante le fasi di particolare criticità del servizio.

Il servizio di certificazione del software si pone come obiettivo il miglioramento della qualità e della sicurezza delle applicazioni sviluppate, tramite un processo continuo di ricerca di possibili criticità e vulnerabilità di sicurezza presenti nel codice e verifica del rispetto delle "best practice" internazionali (OWASP, NIST, ecc.) in materia di sviluppo. Le attività di certificazione del software prevedono una forte interazione tra i differenti gruppi di sviluppo codice e si giovano della possibilità di avviare la collaborazione fin dalle prime fasi del lavoro, in particolare nela fase di code review e per permettere di ridurre al minimo l'impatto di eventuali richieste di correzione codice. In generale si prevede di articolare il processo di certificazione nelle seguenti fasi:

1. **Code Review:** Attività di analisi del codice (statica e dinamica) in ambiente di test con modalità white box, cioè nella modalità di analisi del sistema da certificare con una conoscenza completa dei dettagli implementativi del sistema (simulazione dell'attività di un attaccante interno):
   - **Analisi statica**
     Analisi del codice sorgente senza necessità di compilare o eseguire il codice sorgente, volta a identificare possibili migliorie nel codice avvalendosi delle *best practice* note in letteratura.
   - **Analisi dinamica**
     Analisi a runtime del codice installato e configurato volta a identificare problemi relativi all'esecuzione del programma.
2. **Vulnerability Assessment (VA)**: Analisi tramite strumenti, automatici e non, della presenza di vulnerabilità note in letteratura all'interno del sistema da certificare.
3. **Penetration Test (PT)** : Verifica a *runtime*, su apposito sistema di test, delle criticità individuate utilizzabili per un attacco al sistema da certificare, in modalità black box: cioè nella modalità di analisi del sistema da certificare senza alcuna conoscenza pregressa di materiale che non sia pubblicamente disponibile (simulazione dell'attività di un utente finale).

L'esperienza acquisita in materia di certificazione è stata utile per acquisire all'interno dell'Istituto tecniche di programmazione che si sono rivelate di grande utilità nelle successive rilevazioni.

Il servizio di test per il dimensionamento dell'infrastruttura si pone come obiettivo l'individuazione del carico massimo dell'infrastruttura, prima che le prestazioni (come ad esempio, i tempi di risposta) degradino oppure che i parametri di performance diventino inaccettabili. Le attività di test per il dimensionamento dell'infrastruttura prevedono una forte interazione con il gruppo dei sistemisti per verificare l'adeguatezza delle infrastrutture predisposte all'interno dei carichi stimati di utenti e di tipologia di funzioni richieste durante l'andamento della rilevazione. In generale si prevede di

articolare il processo di test nelle seguenti attività:

- **Test di carico e stress** che valuta le prestazioni del sistema sotto un carico di utenti incrementale.
- **Endurance testing** effettuato per valutare il comportamento dell'applicazione sottoposta ad un carico medio di utenti per un periodo prolungato.
- **Test di scalabilità** che ha l'obiettivo di determinare la scalabilità dell'infrastruttura hardware e software rispetto ad un incremento di utenti o dei volumi di dati scambiati.

ALLEGATO 4
VARIABILI E DATI DA PRODURRE CON IL CENSIMENTO PERMANENTE

Nella progettazione di un censimento, la esplicitazione dei contenuti informativi da rilevare, delle definizioni e delle classificazioni da adottare, deve tener conto di molteplici esigenze, al fine di:

- agire in conformità a quanto previsto nei regolamenti europei e nelle raccomandazioni internazionali;
- garantire la produzione di dati utili al Paese nel rispetto della tradizione censuaria nazionale;
- essere coerenti con la produzione di dati di altre indagini statistiche ufficiali;
- garantire qualità e coerenza rispetto a possibili diversificazioni di fonti, di metodi e di tecniche di acquisizione.

Il Censimento della popolazione del 2011 è stato il primo normato da un regolamento della Unione Europea.[1] Il regolamento, che per la tornata del 2021 non dovrebbe subire modifiche sostanziali,[2] riporta in allegato i core topics (variabili obbligatorie) che, a vari livelli di dettaglio territoriale (fino a regionale o fino a comunale), devono essere resi disponibili ad Eurostat entro 27 mesi dalla fine dell'anno di riferimento del censimento secondo il piano di diffusione di cui al regolamento di attuazione della Commissione n.519/2010[3] e le classificazioni riportate nel regolamento di attuazione della Commissione n.1201/2010.[4]

Le variabili core europee sono state selezionate a partire dalle Raccomandazioni Internazionali dell'UNECE[5] che tradizionalmente guidano i censimenti demografici al fine di garantire la comparabilità dei dati prodotti nei diversi paesi, sincronizzare le operazioni di rilevazione e di produzione dei dati censuari, armonizzare i concetti utilizzati.

Considerato che sono in fase di valutazione i risultati dei lavori svolti da nove *task force* costituite nel 2012 da UNECE per l'aggiornamento delle raccomandazioni internazionali e che non sono ancora riprese presso Eurostat le attività per l'aggiornamento del citato regolamento della Commissione Europea, allo stato attuale la definizione dei contenuti informativi del Censimento permanente deve necessariamente fare riferimento a quanto stabilito dalle raccomandazioni internazionali e dai regolamenti europei per la tornata censuaria del 2011.

[1] Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on population and housing censuses, Official Journal L 218 , 13/08/2008 P. 0014 – 0020.
[2] A tale riguardo si rimanda al documento: Eurostat, Directorate F: Social Statistics "Issue paper for the DSS Board and DSS Discussion", Doc. Eurtostat/F/13/DSS Borad/01/EN, scheda 13, pagine 63 – 67.
[3] COMMISSION REGULATION (EU) No 519/2010 of 16 June 2010 adopting the programme of the statistical data and of the metadata for population and housing censuses provided for by Regulation (EC) No 763/2008 of the European Parliament and of the Council.
[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.
[5] United Nations Economic Commission for Europe "Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing", New York and Geneva 2006.

Prospetto 1 - Core topics (Regolamento UE 763/08)

Popolazione	Alloggi
- Luogo di dimora abituale	- Sistemazione abitativa
- Luogo di lavoro	- Tipo di alloggio
- Sesso	- Alloggio per commune
- Età	- Stato occupazionale dell'alloggio
- Stato civile	- Proprietà
- Condizione professionale o non professionale	- Numero di occupant
- Attività lavorativa	- Superficie e/o numero di stanze
- Settore di attività economica	- Disponibilità sistema di alimentazione idrica
- Posizione nella professione	- Disponibilità gabinetto
- Grado di istruzione	- Disponibilità vasca/doccia
- Luogo di nascita	- Impianto di riscaldamento
- Cittadinanza	- Tipo di edificio
- Residenza all'estero e anno di arrive in Italia	- Epoca di costruzione
- Relazione di parentela	
- Titolo di godimento	
- Località	

Fonte: Regolamento UE 763/08

Con riferimento alle esigenze nazionali, occorre inoltre considerare che il piano di rilevazione del 15° Censimento della popolazione e delle abitazioni non prevedeva solo core topics resi obbligatori dalla UE, ma anche alcune variabili che rientrano fra i non core topics[6] proposti nell'ambito delle Raccomandazioni UNECE del 2010 (ad esempio: cittadinanza italiana dalla nascita, luogo abituale di studio, dimora abituale cinque anni prima del Censimento). Questa scelta è dovuta sia all'esigenza di mantenere aggiornate le serie storiche di informazioni richieste negli anni passati da utilizzatori nazionali (ad esempio: frequenza asilo nido/scuola materna, specifica del titolo di studio più elevato) sia alla necessità interna di garantire la disponibilità dei dati necessari al processo di controllo e correzione.

La progettazione dei contenuti informativi del Censimento permanente, oltre a considerare i vincoli normativi ancora vigenti, dovrà tener conto:

- della qualità dei dati raccolti in occasione del Censimento del 2011 con riferimento a quelle variabili che hanno rappresentato una novità rispetto alle esperienze passate (ad. esempio luogo di nascita dei genitori, energia rinnovabile, disabilità);
- dell'opportunità di continuare a raccogliere informazioni complesse (domande a risposta aperta o che richiedono uno sforzo di memoria) e che comportano elevati costi in termini di tempo e di risorse nella fase di elaborazione dati (ad esempio descrizione del titolo di studio più elevato conseguito);
- della necessità di prevedere quesiti utili solo ad alcune fasi del processo di controllo e correzione (ad esempio la data del matrimonio).

Nella progettazione del censimento, a maggior ragione perché permanente, si dovrà tener fortemente conto delle nuove possibilità offerte dall'integrazione statistica tra gli archivi amministrativi e, per alcune variabili, dai *Big Data*, quali fonti sostitutive della tradizionale rilevazione mediante questionario.

[6] Variabili non obbligatorie ma raccomandate a livello internazionale.

Le potenzialità informative a fini censuari offerte dalle diverse fonti di dati sono illustrate nel seguito, dove sono riportati gli esiti di lavori avviati e in corso in Istat proprio per indagare su questi aspetti.

Archivi amministrativi

Nel 2012 è stato costituito un gruppo di lavoro[7] con l'obiettivo di valutare la potenziale sostituzione o integrazione di dati amministrativi nella rilevazione censuaria e il loro uso a fini di controllo e/o correzione dei dati. In sintesi i principali risultati fin qui raggiunti sono i seguenti:

- la ricognizione ha individuato 26 archivi ed un Sistema Informativo di interesse; è stato esaminato il possibile contributo informativo di ciascuno di essi per le 7 aree tematiche del Foglio individuale del questionario di Censimento 2011; alcune informazioni sono risultate direttamente confrontabili (per esempio luogo di nascita); altre richiedono un ulteriore approfondimento data la loro presenza in più di un archivio;
- sono ancora in fase di studio i possibili archivi di interesse per la Sezione I – Notizie su famiglia e alloggio;
- la sezione più complessa del tradizionale questionario di censimento è senz'altro quella sulla condizione professionale o non professionale; nelle basi date amministrative acquisite dall'Istat si rileva una copertura asimmetrica tra occupati nel settore privato (dipendenti, parasubordinati e autonomi) e occupati nel settore pubblico; per i primi, nell'ambito del 9° Censimento dell'industria e dei servizi l'Istat ha realizzato la base di dati ASIA-Occupazione che è una fonte di informazioni ormai ufficiale e validata, costituirà un supporto importante per questa Sezione del questionario ; un archivio analogo è in fase di sviluppo per gli occupati nel settore pubblico, a partire dalla validazione dell'archivio principale di fonte INPDAP acquisito di recente dal DICA;
- per le variabili anagrafiche dell'individuo si è riscontrato che i dati dell'Anagrafica Tributaria sono più completi di quelli della Lista Anagrafica Comunale. In particolare il test condotto ha mostrato che l'incidenza di valori mancanti per alcune variabili presenti in LAC è consistente, ma recuperabile usando la base dati dell'Anagrafe Tributaria, come mostrato dal prospetto seguente.

Prospetto 2 - Valori mancanti e valori recuperati in alcune variabili del Censimento

Aggregati	Valori mancanti	Valori recuperati
Codice fiscale	878.597	727.743
Luogo di Nascita (prov/com/estero)	20,7 mln	19,6 mln

Fonti: SGR per LAC, Base dati Anagrafe Tributaria

[7] GdL per la valutazione del potenziale uso sostitutivo ed integrativo di dati amministrativi a fini di produzione di dati censuari (Del. N.18 DICA/2012).

Sistemi Integrato di Microdati (SIM)

Il sottosistema di integrazione "Relazioni tra individui e unità economiche" è già da ora in grado di riprodurre annualmente il registro ASIA-Occupazione con riferimento alle imprese. Nel 2014 il sottosistema potrebbe essere sviluppato per comprendere anche le unità del settore non profit e quelle del settore pubblico insieme ai loro occupati.

Nel sottosistema di integrazione "Luoghi degli individui" sono già disponibili le informazioni registrate nelle diverse fonti amministrative e relative alla localizzazione sul territorio delle dimore e delle attività delle persone fisiche, siano esse studenti o lavoratori. Inoltre, nell'ambito del progetto ARCHivio Integrato di Microdati Economici e DEmosociali (ARCHIMEDE), il sottosistema citato è stato di recente utilizzato per produrre con riferimento al 2011 la matrice Origine/Destinazione tra tutti i comuni italiani dei movimenti casa-luogo di lavoro e casa-luogo di studio. I risultati così ottenuti saranno presto messi a confronto con quelli derivati dall'elaborazione del questionario di famiglia del censimento del 2011, in modo da valutare il grado di sostituibilità dei quesiti che nel questionario di famiglia sono dedicati al pendolarismo.

Big Data

Sono dati digitali desumibili dalle tracce rilasciate nell'uso dei social media, dei motori di ricerca, dalle chiamate tramite cellulari, dalla posta elettronica, dai dispositivi GPS o dagli acquisti online. Il loro utilizzo a fini statistici è diventato relativamente consueto con riferimento a domini circoscritti (in termini di popolazione, spazio, tempo), mentre non è altrettanto consueto il loro uso nella produzione di statistiche ufficiali. Dato che la rilevanza dei *Big Data* per la statistica ufficiale è stata riconosciuta anche in sede internazionale dall'High Level Group for the modernisation of statistical production and services, sia in quanto tali, sia integrati con indagini campionarie o fonti di dati amministrativi, l'Istat ha previsto di dedicare al tema studi e progetti sperimentali. È in questa cornice di accrescimento dell'innovazione tecnologica che si colloca la recente firma di un protocollo di ricerca con il CNR.

Rispetto ai contenuti informativi del Censimento permanente, il Protocollo rileva, in quanto prevede di sperimentare nei prossimi mesi metodi di integrazione tra la matrice Origine/Destinazione della mobilità tra luogo di abituale dimora e luogo di lavoro o di studio, realizzata dall'Istat nell'ambito del progetto Persons and Places, e le informazioni derivabili dai tabulati della telefonia mobile. Più in particolare la sperimentazione è finalizzata a sviluppare metodi di statistical matching, in modo da poter completare la matrice Origine/Destinazione con distribuzioni statistiche di dati inerenti la frequenza e la durata dei movimenti individuali tra comuni e località.

In conclusione, alla luce degli approfondimenti e sperimentazioni fin qui condotte, è possibile tracciare nel prospetto seguente le principali corrispondenze tra aree tematiche della sezione II del questionario di famiglia utilizzato per il Censimento della popolazione del 2011 e gli archivi amministrativi o i sistemi statistici di microdati integrati.

Prospetto 3 - Macro corrispondenze tra aree tematiche della sezione II del questionario e archivi amministrativi e Sistemi Informativi statistici

	Sezioni del questionario						
DENOMINAZIONE	Notizie anagrafiche	Stato civile e matrimonio	Cittadinanza	Presenza e dimora precedente	Istruzione e formazione	Condizione professionale	Luogo di studio e lavoro
Liste Anagrafiche Comunali/ ANPR	X	X	X				
Sistema Persons &Places	X			X	X		X
Anagrafe Nazionale studenti non universitari					X		
Anagrafe Nazionale Studenti universitari					X		
Sistema ASIA – Occupazione						X	
Inps Lavoratori domestici						X	
Inps Archivio Autonomi in agri-coltura						X	
INPDAP (Pubblica Amministra-zione)						X	
Casellario dei pensionati						X	