istat
working
papers

N.23
2016

Primi risultati della sperimentazione condotta su fonti amministrative capaci di valutare i segnali di dimora abituale in Italia e l'individuazione di sottopopolazioni critiche

F. Borrelli, A. Chieppa, S. Di Domenico, G. Gallo, S. Rosati, V. Tomeo
Coordinatore del gruppo di sperimentazione: G. Garofalo

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N.23
2016

Primi risultati della sperimentazione condotta su fonti amministrative capaci di valutare i segnali di dimora abituale in Italia e l'individuazione di sottopopolazioni critiche

F. Borrelli, A. Chieppa, S. Di Domenico, G. Gallo, S. Rosati, V. Tomeo
Coordinatore del gruppo di sperimentazione: G. Garofalo

Istat

Istat Working Papers

Primi risultati della sperimentazione condotta su fonti amministrative capaci di valutare i segnali di dimora abituale in Italia e l'individuazione di sottopopolazioni critiche

N. 23/2016

ISBN 978-88-458-1938-4

Primi risultati della sperimentazione condotta su fonti amministrative capaci di valutare i segnali di dimora abituale in Italia e l'individuazione di sottopopolazioni critiche

A cura di: F. Borrelli, A. Chieppa, S. Di Domenico, G. Gallo, S. Rosati, V. Tomeo

Coordinatore del gruppo di sperimentazione: G. Garofalo

Sommario

Il prossimo censimento della popolazione in Italia segnerà la definitiva transizione dal tradizionale conteggio porta a porta a un sistema basato su registri, combinando l'utilizzo di registri ufficiali di popolazione con altre fonti amministrative relative a specifiche tematiche.
Attraverso l'uso di un sistema integrato di registri è possibile identificare nella grande quantità di informazioni amministrative alcuni pattern utili e creare profili di sottopopolazioni di interesse.
L'Istituto Nazionale di Statistica ha condotto una sperimentazione per identificare la popolazione abitualmente dimorante attraverso l'utilizzo di dati amministrativi, giungendo a interessanti risultati.
E' stata valutata la qualità dei registri e sono stati individuati dei pattern nei dati; attraverso questi è stato possibile classificare gli individui in specifici gruppi, alcuni dei quali rappresentano delle sottopopolazioni critiche che devono essere considerate attentamente nel definire una nuova strategia censuaria.
E' stata quindi definita una metodologia sperimentale per giungere a un conteggio della popolazione abitualmente dimorante.

Parole chiave: censimento, innovazione, fonti amministrative.

Abstract

The next Population Census round in Italy will mark the definitive transition from the traditional "door-to-door" enumeration to a "register-based" system, which combines official population registers with other subject-specific administrative sources. By using an integrated system of registers it is possible to identify useful patterns in huge amounts of administrative data and create subpopulation profiles. The Italian National Institute of Statistics (ISTAT) carried out an initial trial to identify the usually-resident population by using administrative data, which produced useful results. ISTAT analyzed the quality of the registers and identified patterns in the data. These patterns enabled ISTAT to classify individuals into specific groups, which also represent the "critical" subpopulation to be considered when defining the new Census strategy. ISTAT then defined a preliminary workflow for deriving usually-resident population counts.

Keywords: census, innovation, administrative sources.

Indice

1. Premessa: gli obiettivi del progetto Archetipo e del Work Package 1 (WPC1)

Al fine di definire il disegno strategico del Censimento permanente della popolazione e delle abitazioni[1], il 12 agosto 2015 l'ISTAT ha costituito un gruppo di lavoro inter-dipartimentale denominato ARCHETIPO (ARCHivi E sisTema di Indagini integrate per il Censimento permanente della POpolazione)[2]. L'attività di questo gruppo si è inserita nella più ampia prospettiva di porre al centro della produzione statistica dell'Istituto l'integrazione degli archivi amministrativi e delle indagini statistiche. In tale ambito, il WPC1 ha ricevuto il mandato di valutare e individuare le fonti amministrative, e le relative variabili, già disponibili o realisticamente acquisibili in ISTAT, capaci di supportare, sulla base della definizione di *usual resident population* del Regolamento CE N. 1260/2013, l'identificazione di una dimora abituale differente dalla iscrizione in anagrafe degli individui residenti in Italia. Inoltre, il WPC1 era incaricato di analizzare la "tempestività" delle fonti prese in esame e individuare le azioni da intraprendere sia per incrementare il patrimonio informativo delle fonti amministrative, sia per migliorare la qualità e la tempestività di quelle già disponibili.

In sintesi, l'attività di ricerca del WPC1 si è basata sulla seguente ipotesi di lavoro: data una struttura informativa di "iscrizione anagrafica" degli individui residenti in Italia, occorre valutare i dati amministrativi utili per identificare le sottopopolazioni con dimora abituale in Italia differente dalla iscrizione anagrafica o con assenza di iscrizione anagrafica.

Sulla base di questi obiettivi, è stata effettuata una prima sperimentazione i cui risultati sono riportati nei paragrafi che seguono. Si precisa però che, in questa prima fase, il test ha avuto l'obiettivo di valutare i segnali presenti nelle fonti amministrative, e nella base integrata dell'Istituto, con riferimento all'individuazione di una dimora abituale in Italia, senza operare, per il momento, alcuna verifica rispetto all'esatto comune italiano dove la dimora abituale stessa dovrebbe essere allocata.

2. Il contesto attuale e il nuovo approccio per l'identificazione della popolazione abitualmente dimorante in Italia

Obiettivo del Censimento della popolazione è la fornitura di dati ufficiali sulla popolazione abitualmente dimorante in Italia ad un elevano livello di dettaglio territoriale. Secondo la definizione del Parlamento e del Consiglio d'Europa per popolazione abitualmente dimorante su un territorio si intende l'insieme delle persone *"..who usually spend their daily rest for at least the last 12 months or intend to live for at least 12 months"* in una specifica area geografica all'interno dei confini nazionali (Regolamento N. CE 1260/2013). Questa definizione, che è adottata sia per i censimenti che per la produzione delle statistiche demografiche di stock e di flusso, si basa non solo sul criterio di "continuità" di almeno dei 12 mesi ma include anche le persone che hanno "l'intenzione" a stare per almeno 12 mesi. Ciò rappresenta un punto critico rilevante, soprattutto in riferimento alla popolazione più mobile sul territorio rispetto alla quale il criterio dell'"intenzionalità" può cambiare anche nell'arco di pochi giorni, come hanno spesso osservato anche i ricercatori dell'Eurostat (Lanzieri, 2013).

In Italia, la prossima tornata censuaria segnerà il passaggio definitivo da un'indagine di tipo tradizionale a una rilevazione basata sull'uso dei registri amministrativi. L'ISTAT negli ultimi anni ha utilizzato le fonti amministrative prevalentemente per la produzione di statistiche economiche e, in parte, per determinare le statistiche demografiche di flusso. In occasione del Censimento del 2011, le fonti amministrative hanno avuto un ruolo centrale nella strategia censuaria in tre attività specifiche: a) l'acquisizione delle Liste Anagrafiche Comunali (LAC) per l'invio dei questionari di Cen-

[1] Il Censimento permanente è previsto dall'articolo 3 del decreto legge n. 179 del 18 ottobre 2012, convertito con modifiche dalla legge n. 221 del 17 dicembre 2012.

[2] ISTAT-Deliberazione, DGEN, n. 78.

simento alle famiglie[3]; b) l'operazione contestuale al Censimento di confronto censimento-anagrafe per la validazione dei dati di popolazione legale; c) il confronto tra le LAC e l'archivio dei Permessi di Soggiorno per derivare liste ausiliare geo-referenziate al fine di migliorare la copertura della rilevazione dei cittadini stranieri provenienti dai paesi terzi dell'Unione europea.

Nonostante l'importante sforzo attuato nel corso del Censimento 2011 per aumentarne la copertura, la rilevazione di alcuni gruppi di popolazione risulta ancora estremamente complessa.

In seguito alla Post Enumeration Survey 2011, che ha coinvolto più di 330 mila individui, l'ISTAT ha stimato che circa 650 mila persone abitualmente dimoranti in Italia non sono state censite al 2011 e la quota di cittadini stranieri sfuggiti al Censimento è pari a circa l'80% (ISTAT, 2014). Questo risultato, oltre ai costi economici e al carico statistico sui rispondenti, connesso alla realizzazione di un Censimento in modalità tradizionale, ha portato l'ISTAT a valutare la possibilità di una nuova strategia censuaria basata sull'uso combinato di dati amministrativi e indagini campionarie (Crescenzi F. et al. 2015).

Il calcolo della popolazione abitualmente dimorante a partire da un sistema integrato di archivi, eventualmente supportato da indagini statistiche ad hoc, può rappresentare una soluzione ottimale per risolvere i problemi connessi all'uso dei soli archivi anagrafici: la sottocopertura di tali registri di popolazione può essere corretta attraverso l'integrazione con dati provenienti da altri archivi amministrativi che rivelano segnali "stabili" di presenza delle persone sul territorio; allo stesso modo, l'assenza di ulteriori segnali amministrativi per le persone iscritte nei registri di popolazione consente di identificare la sottopopolazione critica a "rischio di sovracopertura". Rimane, comunque, aperta la questione di alcuni sottogruppi specifici di popolazione, come i senza fissa dimora e i senza tetto, per i quali è pressoché impossibile rinvenire segnali di presenza da altri archivi amministrativi: la soluzione di una indagine *ad hoc* sembra essere l'unico percorso perseguibile ai fini di un conteggio esaustivo.

Per sfruttare al meglio il sistema integrato di archivi nella produzione degli aggregati di popolazione è fondamentale combinare la conoscenza *a priori* degli esperti di studi sulla popolazione e di esperti nella gestione delle fonti amministrative con le evidenze che emergono dai dati: è quindi particolarmente indicato un processo di *Knwoledge Discovery from Databases* guidato da tali esperti, al fine di far emergere dai dati amministrativi informazioni "nuove" e "rilevanti", utili per i conteggi di popolazione.

3. Gli obiettivi specifici della sperimentazione per l'identificazione della popolazione abitualmente dimorante in Italia

Il primo obiettivo della sperimentazione è definire un *workflow* preliminare su come "processare" i dati amministrativi da cui derivare i conteggi sulla popolazione abitualmente dimorante in Italia: ciò comporta, in primo luogo, una selezione e una valutazione della qualità dei dati amministrativi. Il secondo obiettivo consiste nel fare emergere dai dati disponibili *patterns* rilevanti e le associazioni tra le informazioni derivate dai registri; questo permette di ricavare le variabili più pertinenti tra quelle esistenti, o di determinare nuove variabili derivate rispetto a quelle originarie. Infine, la scoperta di *patterns* può orientare la classificazione di gruppi specifici di individui che potrebbero rappresentare "sottopopolazioni critiche" da tenere in considerazione al momento della definizione della nuova strategia censuaria.

[3] Si fa presente che l'acquisizione delle LAC non è stata effettuata solo in occasione del Censimento 2011 ma rappresenta una attività prevista nel PSN con cadenza annuale.

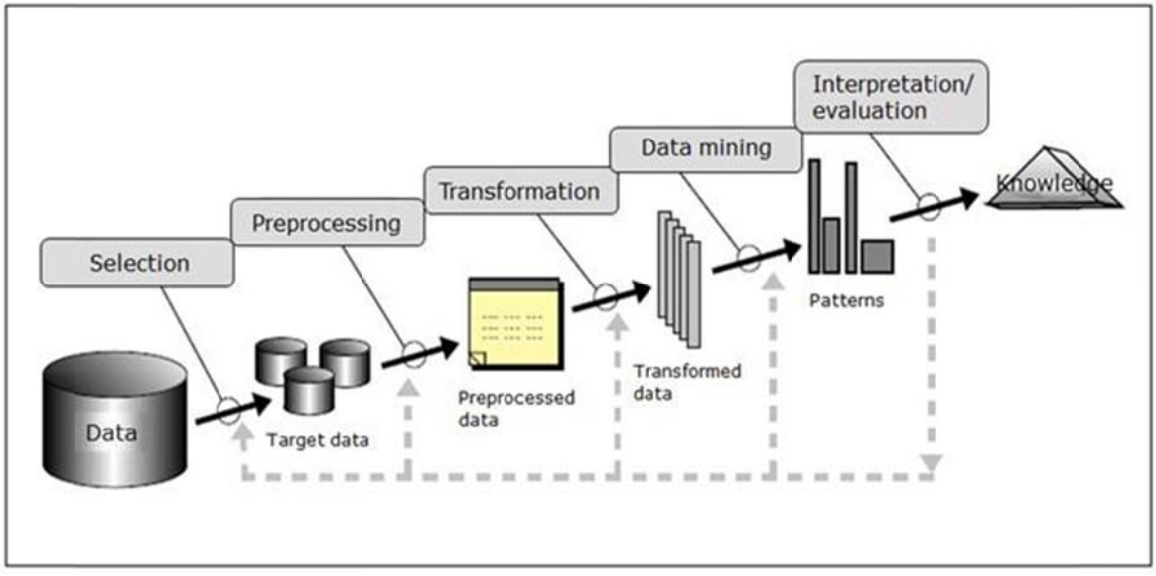
3.1 Il processo di Knowledge Discovery from Databases per la definizione del sistema integrato di archivi

Nella prospettiva di un Censimento permanente in Italia, basato sull'uso degli archivi amministrativi per integrare l'informazione dei registri anagrafici, sono necessarie maggiori informazioni: sia sulla qualità delle fonti amministrative disponibili, sia sulle implicazioni legate alla definizione dei processi di produzione degli output richiesti dagli *stakeholder* nazionali e internazionali.

È necessaria una attenta esplorazione e una dettagliata descrizione di quanto i dati a disposizione offrano: quali sono gli archivi amministrativi, oltre ai registri di popolazione, disponibili in ISTAT e utili a soddisfare gli obiettivi censuari? È plausibile identificare una gerarchia tra le fonti amministrative? È possibile derivare variabili di classificazioni pertinenti ai fini censuari dalle variabili contenute negli archivi amministrativi? Quali sono i gruppi di popolazione che emergono dall'integrazione tra archivi amministrativi?

Quando si maneggiano database complessi, come è il caso di un sistema integrato di archivi amministrativi, non è possibile effettuare direttamente una analisi esplorativa sulle tabelle esistenti: la fase di *data mining* è possibile solo all'interno di un adeguato processo che preveda anche opportune estrazioni e trasformazioni dei dati. Per questo motivo l'approccio scelto per condurre le sperimentazioni è quello comunemente noto come *Knowledge Discovery from Database* (KDD), il cui obiettivo è proprio l'estrazione dai dati di nuova informazione, utile e pertinente ("*the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*", Fayyad et al., 1996). Il KDD è strutturato in diverse fasi, ciascuna delle quali egualmente importante ai fini del risultato finale. La *prima fase* consiste nell'attività di estrazione e di preparazione dei micro-dati; la *fase intermedia* è caratterizzata dall'analisi dei dati (*data mining*); infine, la *fase finale* è dedicata all'elaborazione e interpretazione dei risultati. Poiché il processo di KDD si basa sul presupposto che esistano "modelli latenti" nei dati, una volta che tale informazione emerge nei risultati si rende spesso necessario rivedere e aggiornare le fasi iniziali: in questo modo si genera un ciclo di miglioramento continuo e di messa a punto del processo che permette di pervenire a risultati via via più rilevanti ed efficaci.

Figura 1 – Knowledge Discovery from Databases scheme



Fonte: Fayyad, Piatetsky-Shapiro, Smyth, 1996

Questo tipo di approccio, soprattutto quando è guidato da esperti di studi di popolazione e da esperti di dati amministrativi, può rappresentare uno strumento potente di conoscenza e valutazione delle fonti amministrative disponibili in ISTAT. La sperimentazione descritta nei paragrafi seguenti ne è una prova.

3.2 Le fonti amministrative e gli attributi rilevanti per il calcolo della popolazione abitualmente dimorante

Il Sistema Integrato di Microdati (SIM) è un repository di dati amministrativi integrati costruito con lo scopo di sostenere i processi di produzione statistica, sia per le statistiche sociali sia per le statistiche economiche (Di Bella, Ambroselli, 2014). L'integrazione avviene attraverso l'assegnazione di un codice ID univoco e costante che permette di identificare ciascun individuo e unità economica all'interno dei diversi archivi e di costruire le relazioni tra le diverse fonti.

Le indicazioni fornite dagli esperti di fonti amministrative e da coloro che usano tali fonti per la produzione di statistiche sulla popolazione sono state essenziali nella selezione degli archivi da prendere in considerazione. Tra tutti quelli disponibili, sono stati scelti:

- Base dati della popolazione iscritta in anagrafe
- Archivi del lavoro dipendente e del lavoro autonomo
- Archivio degli studenti della scuola dell'obbligo
- Archivio degli studenti universitari
- Casellario dei pensionati
- Trattamenti non pensionistici
- Dichiarazioni fiscali
- Archivio dei permessi di soggiorno.

Nella Tavola 1, viene riportata l'informazione relativa al numero di record per alcuni degli archivi relativi a lavoro e studio più importanti, riferita al 2012: questa informazione permette di disporre di una approssimazione del volume di tali fonti.

Tavola 1 – Volume dei principali archivi relativi ad attività di lavoro e studio. Anno 2012

Tipo di Archivio	2012
INPS - EMens	17.283.049
INPS - DMAG	1.190.049
INPS - Autonomi agricoli	467.293
INAIL - Lavoratori interinali	604.085
INPS - Parasubordinati Collaboratori	1.685.545
INPS - Lavoro Domestico	978.334
INPDAP - Assicurati	2.982.862
MIUR - Anagrafe Studenti	8.411.156
MIUR - Universitari	611.198

Fonte: Nostra elaborazione su dati Istat Descrizione della fonte

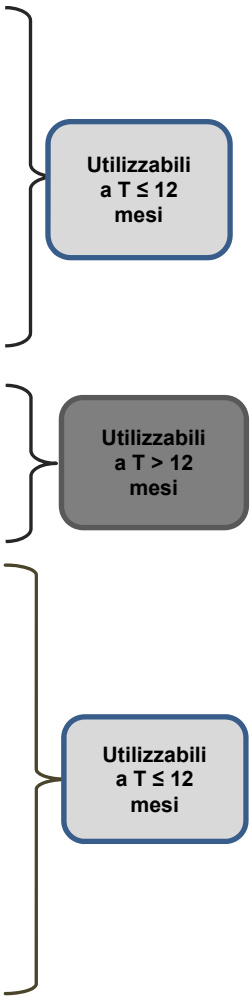
Nel Prospetto 1 è riportato l'elenco dettagliato dei singoli archivi succitati importabili dal SIM; le fonti sono state classificate in base a un criterio di rilevanza in cui si considera che lavorare o frequentare un corso di studio rappresenta una garanzia forte di presenza sul territorio.

La base dati della popolazione residente è costituita prevalentemente dall'Anagrafe Virtuale Statistica (ANVIS) e, solo in minima parte, dalle LAC. La popolazione di ANVIS è costruita partendo dal set di microdati della popolazione legale al 9 ottobre 2011 a cui si aggiungono i seguenti dati individuali di flusso:

- i microdati degli esiti di SIREA (revisione post-censuaria delle anagrafi effettuata con riferimento al 9 ottobre 2011, ai sensi dell'Art.46 del Regolamento anagrafico), sia in addizione sia in sottrazione;
- i record derivanti dalle procedure di contabilità demografica MIDEA (MIcro-DEmographic Accounting) applicate al set dei microdati degli eventi della dinamica demografica (flussi comunali di iscrizione per nascita; cancellazione per decesso; iscrizione/cancellazione da/per l'estero; iscrizione/cancellazione da/per altri Comuni italiani), verificatisi nei periodi: 9 ottobre 2011-31 dicembre 2011, anno 2012, anno 2013 e anno 2014.

Prospetto 1 – Fonti amministrative utilizzate e nuove fonti in via di acquisizione, periodicità dei dati, tempi di lavorazione e rilevanza

Tipologia della fonte	Fonte	Periodicità dei dati	Tempistica	Rilevanza rispetto alla DA
Anagrafica	ANVIS	R	(da valutare)	ALTA
	LAC	A	T+3	ALTA
Previdenziali e assicurative	INPS - EMens	M	T+11	ALTA
	INPS - DMAG	R	T+11	ALTA
	INPS - Autonomi agricoli	R	T+11	ALTA
	INPS - Parasubordinati	R	T+10	ALTA
	INPS - Lavoratori domestici	R	T+5	ALTA
	INPS - Casellario dei Pensionati	A	T+10	MEDIA
	Ex-INPDAP – Lavoratori PA	R	T+10	ALTA
	INPS - Trattamenti non pensionistici	A	T+11	MEDIA
	INAIL- Lavoratori interinali	R	T+4	ALTA
Istruzione	MIUR - Laureati	A	T+12	ALTA
	MIUR - Anagrafe degli studenti	R	T+14	ALTA
	MIUR - Esiti scolastici	A	T+21	ALTA
	MIUR - Anagrafe Universitari	R	T+14	ALTA
	MIUR - Personale Scuole Statali	A	T+14	ALTA
Dichiarazioni Fiscali	MEF - Banca dati reddituale	A	T+21	ALTA
	UNICO/730/770 - Familiari a carico	A	T+18	ALTA
	Cedolini Stipendiali	M	T+5	ALTA
	Anagrafe Tributaria	A	T+3	ALTA
Altre	Permessi di soggiorno	R	T+2	MEDIA
Nuove Fonti *(a partire dal 2016)*	Separazioni e divorzi	A	T+10	BASSA
	Matrimoni	A	T+10	BASSA
	Acquisizione cittadinanza italiana	A	T+9	BASSA
	Richieste di asilo	A	(non conosciuta)	MEDIA
	Visti di ingresso in Italia	A	T+6	MEDIA
	Schedari consolari	A	T+9	ALTA
	Consumi elettrici	A	(non conosciuta)	ALTA
	Consumi gas	A	(non conosciuta)	ALTA
	Contratti di locazione	A	(non conosciuta)	ALTA
	Certificazione Unica	A	T+6	ALTA
	Comunicazioni obbligatorie	A	(non conosciuta)	ALTA
	Dottorati di ricerca	A	(non conosciuta)	ALTA



Note: R= Range temporale con date di inizio e di fine; M=mensile; A= Annuale

Fonte: nostra elaborazione, Istat

Poiché l'archivio di ANVIS risulta ancora in una fase di sviluppo all'interno dell'ISTAT, si è reso necessario, ai fini della sperimentazione, includere altri 347 mila individui iscritti in anagrafe di cui si ha avuto riscontro solo nelle LAC.

3.3 Il problema della tempistica delle fonti utilizzate e l'acquisizione di nuove fonti

Nell'ambito delle fonti utilizzate, l'istante di riferimento dei dati e i tempi di acquisizione di ciascuna delle fonti amministrative rappresenta una criticità di processo importante per l'ISTAT. Se si assume come dato il tempo di lavorazione e di integrazione delle fonti disponibili, rispetto al quale un maggior impiego di risorse tecnologiche e una maggiore efficienza da parte dell'ISTAT possono determinare notevoli guadagni di tempo, l'aspetto più rilevante è rappresentato dalla loro utilizzabilità in termini di "ritardo data" rispetto ad un mese di riferimento, che presumibilmente, rappresenta la data di riferimento del Censimento. La disponibilità e l'utilizzabilità dei dati amministrativi devono essere compatibili con la tempistica degli output previsti dai regolamenti europei. Il Prospetto 1 mostra, per ciascuna tipologia di fonte, la periodicità dei dati, la tempistica (intesa come somma del periodo di riferimento dei dati e il tempo necessario per l'acquisizione delle fonti) e la rilevanza

(bassa, media o alta) rispetto alla definizione della dimora abituale, secondo il regolamento europeo (regolamento CE n. 1260/2013). Dal Prospetto 1 è possibile osservare che le fonti anagrafiche e quelle previdenziali e assicurative (da cui si evincono prevalentemente gli individui che partecipano al mercato del lavoro, ma anche i pensionati e gli individui che beneficiano dei trattamenti non pensionistici), sono utilizzabili per la produzione statistica con un ritardo data pari a soli 12 mesi; invece, per i dati provenienti dal Ministero dell'Istruzione, se si escludono le informazioni sui laureati, si osserva un ritardo data che va dai 14 mesi dell'anagrafe degli studenti ai 21 mesi degli esiti scolastici. Così anche per alcune fonti delle dichiarazioni fiscali, in particolare quelle della Banca dati reddituale e le informazioni dei modelli UNICO/730/770, si registrano tempi di utilizzazione delle informazioni con un ritardo data di circa 20 mesi.

In definitiva, tempi di utilizzabilità dei dati che superano i 12 mesi rappresentano una criticità forte per l'uso del sistema integrato dei dati amministrativi per fini censuari, soprattutto se si considerano le tempistiche di rilascio dei dati per Eurostat. Ad esempio, il *framework Regulation* dell'Unione europea (n. 763 del 2008), che è in vigore anche per i censimenti del 2021, stabilisce che i dati di popolazione legale del prossimo Censimento della popolazione debbono essere consegnati ad Eurostat entro il 31 marzo del 2024. Ciò comporta che, ipotizzando un processo di produzione del Censimento permanente fortemente orientato sull'impiego dei archivi amministrativi, un ritardo di utilizzabilità dei dati pari a 18-20 mesi potrebbe comportare un forte slittamento in avanti degli output censuari.

Inoltre, una valutazione accurata sia sulla tempistica sia sul contributo informativo dovrà essere effettuata anche sulle nuove fonti che, in parte già acquisite nel corso del 2016 (ad esempio, gli schedari dell'anagrafe consolari per gli italiani residenti all'estero) in parte di prossima acquisizione, arricchiranno il sistema integrato delle fonti amministrative disponibili in ISTAT (Prospetto 1). A questo riguardo, oltre agli schedari consolari, di particolare rilevanza sono le informazioni riguardanti i contratti di locazione e le utenze (gas ed elettricità), le comunicazioni obbligatorie (riguardanti le instaurazioni, le proroghe, le trasformazioni e le cessazioni dei rapporto di lavoro), le certificazioni uniche (che arricchiranno le fonti sulle dichiarazioni fiscali), i richiedenti asilo e i visti di ingresso (che vanno ad integrare le informazioni sugli stranieri Non Ue presenti sul territorio).

3.4 Segnali diretti sulla presenza di individui sul territorio italiano

L'informazione relativa a ciascun individuo contenuta negli archivi amministrativi può costituire un *segnale della presenza* di tale individuo sul territorio italiano: a seconda dell'archivio, possono essere necessarie delle regole di estrazione specifiche (ad esempio, nel caso del Casellario dei pensionati, il segnale viene estratto se il *flag* relativo alla dimora all'estero non è valorizzato).

Negli archivi amministrativi di partenza sono presenti attributi specifici delle singole fonti: il lavoro di selezione ha riguardato anche questi aspetti, sono stati scelti solo gli attributi rilevanti alla determinazione della popolazione dimorante abitualmente (vedi Allegato 1 – descrizione delle variabili e delle riclassificazioni); inoltre si è provveduto a uniformare la struttura e i codici relativi a fonti diverse, in modo da rendere possibile l'integrazione tra tutti i dati a disposizione.

Gli archivi dell'INPS relativi al lavoro e quelli del MIUR relativi alla frequenza di corsi scolastici e universitari (denominati "archivi di attività") sono particolarmente rilevanti ai fini della individuazione della popolazione abitualmente dimorante: tali archivi offrono un notevole dettaglio informativo, legato *in primis* alla durata della attività, ma poi anche alla localizzazione (Comune e indirizzo) della stessa e ad alcuni attributi specifici dell'attività svolta (tipologia di contratto, corso di studi, ecc.) che possono essere d'aiuto nella valutazione della forza del segnale di presenza sul territorio. Per questo motivo, gli archivi di attività rivestono una importanza maggiore (nel Prospetto 1 vengono contrassegnati da un valore "Alto" della rilevanza) nell'ambito di questa analisi. Nella Tabella 2 sono riportati gli attributi considerati per ciascuna fonte di questo tipo.

Tavola 2 – Attributi degli archivi relativi ad attività di lavoro e/o studio

Nome dell'attributo del Database integrato	Descrizione
Codice individuo	Identificativo univoco di ciascun individuo che permette l'integrazione tra le varie fonti e anni differenti
Codici Provincia e Comune	Luogo di lavoro o di studio a seconda del tipo di fonte da cui proviene l'informazione
Durata dell'attività – Sequenza di presenze mensili	Dettaglio dei mesi, da gennaio a dicembre degli anni considerati, a cui l'informazione si riferisce
Lavoro – Qualifica	Operaio, impiegato, dirigente,…
Lavoro – Orario di lavoro	Tempo pieno, parziale orizzontale,…
Lavoro – Durata del contratto	Indeterminato, determinato, stagionale
Studio – Tipologia di corso	Scuola dell'obbligo, Università
Unità economica di riferimento	Unità locale o scuola
Variabili specifiche di fonte	es. "stato: frequentante/trasferito/abbandono" per la fonte "Anagrafe degli studenti", "convivenza con il datore di lavoro" per la fonte "Rapporti di lavoro domestico"

Fonte: nostra elaborazione su dati Istat

Di particolare importanza è l'attributo relativo alla "durata dell'attività": la circostanza che questa informazione sia riportata in diversi formati negli archivi di base contenuti nel SIM (date di inizio e fine, trimestri ecc.) ha comportato la necessità di calcolare una durata standardizzata su base mensile.

3.5 Segnali indiretti calcolati sulla base delle relazioni familiari tra individui

I segnali possono essere direttamente associati ad un individuo oppure rappresentare un "segnale indiretto", che può essere considerato meno robusto di quello diretto e che è l'espressione di una relazione familiare esistente tra l'unità analizzata e un'altra persona.

I segnali indiretti sono stati ricavati a partire dalle dichiarazioni fiscali, ossia, nello specifico, dal "Quadro Familiari a carico" dei modelli UNICO Persone Fisiche e 730 (Redditi 2012). Le relazioni principali che si possono individuare sono quelle tra "coniugi" e tra genitore certo (il dichiarante) e "figlio/i". Nella base dati sono contenuti, inoltre, anche i dati relativi agli "altri familiari" a carico (il coniuge legalmente ed effettivamente separato; i discendenti dei figli; i genitori (compresi i genitori naturali e quelli adottivi); i generi e le nuore; il suocero e la suocera; i fratelli e le sorelle (anche unilaterali); i nonni e le nonne) a patto che vivano con il dichiarante o che ricevano da lui assegni alimentari non risultanti da provvedimenti dell'Autorità giudiziaria.

Altri tipi di segnali indiretti sono quelli che si ricavano dalle LAC e che si riferiscono alla composizione del nucleo familiare registrato in anagrafe. In questo caso, le relazioni principali sono quelle tra intestatario della famiglia e "coniuge" e tra intestatario e "figlio/i".

3.6 Variabili demografiche

Le informazioni anagrafiche utili all'analisi esplorativa sono state ottenute integrando le informazioni contenute nei diversi archivi inseriti nel SIM. In particolare, i dati utilizzati per ricostruire il profilo demografico degli individui hanno riguardato:

- Data di nascita
- Sesso
- Cittadinanza
- Paese di nascita

Qualora la Cittadinanza non fosse presente si è fatto riferimento all'informazione sul Paese di nascita.

In presenza di informazioni anagrafiche incongruenti nelle diverse fonti per uno stesso individuo sono state scelte le informazioni più frequenti e, in seconda battuta, quelle più recenti.

La fonte dell'Anagrafe Tributaria, insieme alle LAC, rappresenta la più completa rispetto alle variabili anagrafiche Data di nascita, Sesso e Paese di nascita (è assente l'informazione relativa alla

Cittadinanza, presente invece nelle LAC). Si evidenza, inoltre, che ai fini dell'analisi della effettiva localizzazione degli individui in termini di dimora abituale, per ora non considerata nell'ambito di questa sperimentazione, risulta molto utile la presenza in questa fonte anche delle variabili "Comune" e "Indirizzo" del domicilio fiscale.

3.7 L'analisi dei segnali e profili di continuità nella presenza dalle fonti di lavoro e studio

Una volta che dal SIM sono stati scaricati i dati derivanti dalle fonti amministrative prescelte e sono poi stati strutturati in maniera adeguata all'analisi della popolazione dimorante in Italia, possono essere estratti i segnali di presenza attraverso opportune regole deterministiche. Questi segnali vengono quindi analizzati e valutati in combinazione tra loro.

Ogni analisi su segnali provenienti dagli archivi amministrativi deve essere preceduta dalla scelta del periodo di riferimento: lo stesso concetto di presenza su un territorio è imprescindibilmente legato a un intervallo di tempo. La scelta del periodo di riferimento rappresenta un fattore cruciale ai fini dell'identificazione e dell'analisi dei segnali di presenza: più ampia è la finestra temporale maggiore è la possibilità di valutare il peso dei segnali sia in termini di continuità e di stabilità nel tempo sia a livello di numerosità e di ripetizione connesso ai processi migratori, alla mobilità lavorativa, ai percorsi di studio etc. Ad esempio, nella Tavola 3 è possibile verificare che dal confronto della localizzazione dei segnali relativi ad un anno (2011, nell'esempio) con un altro periodo di tre anni (dal 2011 al 2013) scaturisce un notevole incremento degli individui presenti su più di un comune: la percentuale di individui localizzati in un solo comune scende, nel periodo considerato, dal 89% al 77%.

Tavola 3 – Distribuzione percentuale dei segnali rispetto alla localizzazione comunale, in diversi periodi di tempo

N. Comuni dove si esplica l'attività di lavoro/studio	Periodo di riferimento	
	2011	2011-2012-2013
Un solo comune	89%	77%
Più comuni	11%	24%

Fonte: nostra elaborazione su dati Istat

Ai fini della sperimentazione, si è tenuto conto dalla definizione di dimora abituale del Regolamento 1260/2013 dell'Unione europea che, ai fini della *usual residence population*, specifica come riferimento temporale la permanenza in un luogo per almeno 12 mesi prima della data di riferimento e di almeno 12 mesi, dopo la stessa data, come espressione dell'"intenzione" di vivere stabilmente nello stesso territorio. Sulla base di queste indicazioni, per la sperimentazione è stato scelto il periodo che va da gennaio dell'anno T a dicembre dell'anno T+1 e che ha come data di riferimento il 31 dicembre dell'anno T.

Nel database utilizzato per le sperimentazioni sono stati estratti segnali in cui si integra l'informazione delle fonti su diversi periodi temporali (Prospetto 2).

Prospetto 2 – Struttura informativa del segnale di presenza in un determinato periodo P

Nome attributo	Codice territorio + codice individuo	Fonti di presenza del segnale nel periodo considerato	Attributi specifici	{mese1--mese24}
Descrizione	{Identificativo}	{Sequenza fonti: ogni posizione una fonte specifica; 1=Presenza nella fonte}	{Altre info a corredo}	{Presenza/Assenza mensile}
Esempio	*Indiv. 18 di Aglié*	*Emens (pos.1)+Università (pos.9)*	*Tempo indeterminato*	*Presente tutti i mesi*
Dati di esempio	001-001-0000018	1000000010	0----1-----	11111111111111111111

Fonte: nostra elaborazione su dati Istat

Il periodo considerato per la valutazione dei segnali ha una lunghezza di 24 mesi, come detto in precedenza. Ogni segnale è da associare ad uno specifico individuo e ad una determinata localizzazione territoriale: ad esempio, se nel periodo considerato per un individuo con codice "0000018"

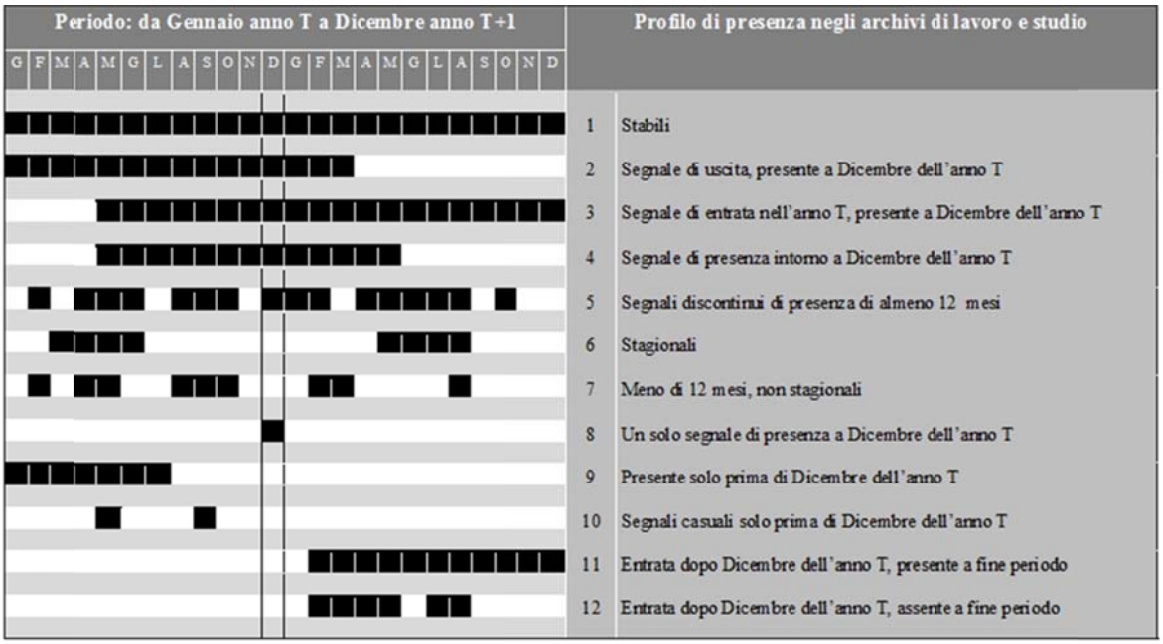
viene rilevato un record in un archivio di lavoro e un altro in una fonte relativa allo studio, entrambi però localizzati nel comune di Agliè, disporremo di un unico segnale localizzato in quel comune specifico. Questo segnale, però, viene contrassegnato sia da un attributo che permette di tracciare l'individuo in entrambi gli archivi di partenza, sia da un attributo relativo alla durata della presenza espressa in termini di attività lavorativa di studio. Nel caso in cui non è disponibile il dettaglio territoriale, viene usato un codice fittizio che rappresenta la mancanza di informazione.

Rispetto alla durata del segnale nel periodo, viene considerata la durata complessiva risultante dai vari archivi che lo hanno generato. Nell'esempio riportato, se l'individuo risulta iscritto per tutto l'anno T all'università e per tutto l'anno T+1 ha un contratto di lavoro dipendente, l'attributo che rappresenta la durata del segnale sarà costituito da una sequenza di 24 digit tutti uguali a 1.

Così come per i segnali diretti e indiretti si è fatto riferimento ad una maggiore robustezza dei primi in termini di capacità di rilevare la presenza sul territorio, le fonti sono state classificate in modo gerarchico considerando i segnali di attività relativi alle fonti lavoro, studio e università come più rilevanti. Lavorare o frequentare un corso di studio rappresenta una garanzia forte di presenza sul territorio. Allo stesso tempo i registri di attività dispongono di informazioni più dettagliate come la durata e il tipo di attività svolta.

L'analisi dei segnali di attività condotta sulla base della finestra temporale prescelta ha permesso di costruire uno schema di presenza mensile strutturato lungo una sequenza di 24 caratteri (tanti quanti sono i mesi considerati), ciascuno dei quali indica se, in un determinato mese, una persona è presente oppure no per quella specifica attività e in quale territorio essa si esplica (Prospetto 3).

Prospetto 3 – Schema sui profili di continuità dei segnali di lavoro e studio



Fonte: nostra elaborazione, Istat

È possibile osservare che esistono più modelli di continuità che possono essere raggruppati secondo i seguenti profili:

- Profilo 1 - Segnali stabili, che esprimono una presenza stabile e continua lungo l'intero periodo di riferimento;
- Profili 2 e 3 - Segnali "in entrata" e segnali "in uscita", che esprimono una presenza di più di dodici mesi, che comprende dicembre dell'anno T; possono riferirsi a persone che hanno cambiato lavoro, hanno iniziato o perso un'attività o che si muovono sul territorio;
- Profilo 4 - Segnali "intorno a dicembre dell'anno T", che rappresentano una presenza continua, di almeno dodici mesi, dal 2012 al 2013;

- Profilo 5 - Segnali "discontinui", che esprimono una presenza di più di dodici mesi ma con più di un'interruzione mensile;
- Profilo 6 - Segnali "stagionali", che riguardano casi il cui modello di attività si ripete una sola volta per anno.
- Profili 7 e 8 - Altri segnali, che rappresentano una presenza discontinua ma inferiore a dodici mesi, oppure una presenza occasionale di un solo mese;
- Profili da 9 a 12 - Segnali "continui" ma concentrati solo nell'anno T, escludendo il periodo di riferimento, oppure tutti nell'anno T+1 (come nel caso dei profili 9 e 11); segnali "discontinui", come nei profili 10 e 12.

L'associazione di questi segnali alle variabili demografiche degli individui ha consentito la costruzione di un database specifico da utilizzare ai fini della sperimentazione.

3.8 Le operazioni di linkage tra gli archivi e la classificazione degli individui in sottopopolazioni

L'analisi del linkage tra gli archivi amministrativi e quelli anagrafici ha permesso di determinare una serie di sottogruppi di popolazione utili ai fini del conteggio. Il Prospetto 4 mostra il processo di identificazione delle sottopopolazioni e il loro ammontare, prendendo in esame il periodo che va da gennaio 2012 a dicembre 2013 ma con il riferimento dei dati al 31 dicembre 2012. Tale processo è stato sviluppato attraverso tre fasi:

1) nella prima fase, il Registro di Popolazione ANVIS è stato confrontato prima con i segnali di attività provenienti dagli Archivi di Lavoro e di Studio (LS), poi con i segnali di presenza nelle Liste Anagrafiche Comunali (LAC) e nell'archivio dei Permessi di Soggiorno (PS);
2) nella seconda fase, solo per gli individui presenti in ANVIS e privi di segnali di Lavoro e di Studio è stata verificata l'esistenza di segnali provenienti dagli archivi del Casellario dei Pensionati (CP) e dei Trattamenti Non Pensionistici (TNP);
3) nella terza fase, sono stati utilizzati anche i segnali indiretti provenienti dalle Dichiarazioni Fiscali (DF) e dalle famiglie anagrafiche identificate nelle LAC (LACfam).

La *prima fase* del processo permette così di identificare cinque gruppi di popolazione:

- Individui presenti in ANVIS per i quali si hanno segnali negli archivi di LS (Gruppo A1);
- Individui non presenti in ANVIS con segnali da archivi di LS e registrati nelle LAC (Gruppo A2);
- Individui presenti in ANVIS senza segnali da archivi di LS (Gruppo B);
- Individui non presenti né in ANVIS né nelle LAC, ma che presentano segnali provenienti dagli archivi di LS (Gruppo C);
- Individui registrati solo nell'archivio dei Permessi di Soggiorno, assenti da ANVIS e dagli archivi di LS ma presenti nelle LAC (Gruppo D1);
- Individui registrati solo nell'archivio dei Permessi di Soggiorno, assenti da ANVIS, dagli archivi di LS e dalle LAC (Gruppo D2).

Nella *seconda fase*, la sola popolazione del Gruppo B è stata messa in relazione con gli archivi del Casellario dei Pensionati e dei Trattamenti Non Pensionistici. I risultati di questa analisi hanno evidenziato la presenza di due sottogruppi di popolazione:

- Individui presenti in ANVIS, senza segnali di LS, con segnali negli archivi del CP o dei TNP (Gruppo B1);
- Individui presenti in ANVIS, senza segnali di LS, senza alcun segnale né nel CP né nei TNP (Gruppo B2).

Infine, nella *terza fase*, per la popolazione del Gruppo B2 è stata verificata la presenza di segnali indiretti all'interno delle fonti delle Dichiarazioni Fiscali al fine di identificare individui dichiarati a carico per fini fiscali (il segnale indiretto si esplicita nella relazione di parentela con il dichiarante) come: "coniuge", "figlio/i" o "altri familiari". Successivamente, è stata utilizzata l'informazione sulle famiglie anagrafiche identificate nelle LAC in modo da riconoscere le persone registrate come

"coniugi", all'interno della stessa famiglia, di individui (intestatari della famiglia) che presentano anche di segnali di lavoro e studio (appartenenti, quindi, ai Gruppi A1 e A2). Tale analisi ha restituito due sottogruppi di popolazione:

- Individui che sono presenti in ANVIS con solo segnali indiretti provenienti dalle fonti DF e LACfam (Gruppo B2.1);
- Individui presenti in ANVIS senza alcun segnale diretto dalle fonti non anagrafiche e neanche indiretto proveniente dalle fonti DF e LACfam (Gruppo B2.2).

Sull'insieme delle sottopopolazioni individuate è possibile compiere delle specifiche valutazioni quantitative in riferimento alla dimora abituale degli individui. In tal senso, è plausibile ritenere che gli individui appartenenti ai Gruppi A1, A2, B1 e B2.1 presentino segnali sufficientemente robusti da poter essere considerati "eleggibili" per il registro della popolazione abitualmente dimorante in Italia. Questo ammontare è pari a 58,1 milioni di individui.

Al contrario, le sottopopolazioni dei Gruppi B2.2, C, D1 e D2 sono costituite dai casi "dubbi" (o incerti) in termini di dimora abituale sul territorio e contabilizzano in totale 4,6 milioni di individui. Questo ammontare è oggetto di specifici approfondimenti, condotti nei paragrafi che seguono, con l'obiettivo di valutarne ulteriormente il grado di eleggibilità.

Prospetto 4 – Schema di processo e identificazione dei gruppi di popolazione in base alla loro eleggibilità o meno ad essere inclusi nella popolazione abitualmente dimorante in Italia – Data di riferimento al 31.12.2012 (*dati in migliaia*)

I FASE						II FASE			III FASE			
Fonte						Fonte			Fonte			
ANVIS	LS	LAC	PS			CP-TNP			DF	LACfam		
60.742,7	37.704,3	61.251,6	3.378,4			20.763,8			26.648,8	4.559,4		
Presenza				*Gruppo*	*Individui*	*Presenza*	*Gruppo*	*Individui*	*Presenza*		*Gruppo*	*Individui*
Sì	Sì	-	-	**A1**	**36.271,0**							
No	Sì	Sì	-	**A2**	**347,1**							
Sì	No	-	-	B	24.471,7	Sì	**B1**	**14.506,7**				
						No	B2	9.965,0	Sì	Sì	**B2.1**	**6.939,1**
									No	No	**B2.2**	**3.026,0**
No	Sì	No	-	**C**	**1.086,2**							
No	No	Sì	Sì	**D1**	**106,5**							
No	No	No	Sì	**D2**	**351,1**							

Legenda

Eleggibili

Dubbi

ANVIS	*Anagrafe Virtuale Statistica*
LS	*Archivi di Lavoro e Studio*
LAC	*Liste Anagrafiche Comunali*
PS	*Permessi di Soggiorno*
CP-TNP	*Casellario Pensionati - Trattamenti non Pensionistici*
DF	*Dichiarazioni Fiscali*
LACfam	*Famiglie anagrafiche delle LAC*

Fonte: nostra elaborazione su dati Istat

3.9 La stabilità dei risultati in riferimento a due istanti di tempo: 2012 e 2013

Al fine di valutare l'evoluzione dei segnali su più anni e verificare se i gruppi di sottopopolazione identificati al 31 dicembre 2012 subiscono variazioni significative nel tempo, è stata effettuata una ulteriore sperimentazione, utilizzando come finestra temporale il periodo che va da gennaio 2013 al 31 dicembre 2014, ponendo come istante di riferimento il 31 dicembre 2013. Di seguito si riporta il confronto nei due istanti di tempo secondo i principali gruppi individuati nel processo di creazione delle sottopopolazioni critiche (Tavola 4).

L'integrazione con gli archivi anagrafici delle fonti amministrative prese in esame ha permesso di identificare l'insieme degli individui eleggibili ad essere inclusi nella popolazione abitualmente dimorante in Italia alla data di riferimento considerata. Questo insieme di "possibili dimoranti abitualmente" ammonta a 62,6 milioni di individui nel 2012 (alla data del 31 dicembre) e a 62,4 milioni nel 2013. A partire da questo insieme di individui, è possibile identificare tre sottogruppi principali:

1. La sottopopolazione presente negli archivi anagrafici senza segnali da altre fonti, con ammontare pari a 3,0 milioni di individui nel 2012 e 3,9 milioni nel 2013 (Gruppo B2.2);
2. La sottopopolazione presente negli archivi anagrafici che presenta anche segnali in altre fonti, con 58,1 milioni di individui nel 2012 e 57,1 nel 2013 (Gruppi A1+A2+B1+B2.1);
3. La sottopopolazione non presente in ANVIS ma con segnali da altre fonti, con 1,5 milioni di individui nel 2012 e 1,4 milioni nel 2013 (Gruppi C+D1+D2).

È possibile osservare che, dal 2012 al 2013, il gruppo degli individui iscritti in ANVIS senza segnali di lavoro studio (Gruppo B), aumenta di oltre 1,2 milioni (in termini relativi, più del 5%). La distribuzione delle sottopopolazioni del Gruppo B mostra nel biennio 2012-2013 un aumento degli individui senza segnali nelle fonti di lavoro e studio. Questa variazione è da imputare all'incremento dei seguenti aggregati: circa 245 mila individui con segnali derivanti dal casellario dei pensionati e dai trattamenti non pensionistici (Gruppo B1); altri 993 mila individui del Gruppo B2[4], di cui particolarmente significativo appare l'incremento (+653 mila individui) caratterizzato dalle nascite verificatesi successivamente alle dichiarazioni fiscali del 2012[5] e dai lavoratori (autonomi e dipendenti) che nel 2013 non partecipano più al mercato del lavoro (Gruppo B2.2). La contrazione 2012-2013 nella partecipazione al mercato del lavoro si registra anche per il sottogruppo della popolazione C (-80 mila individui dal 2012 al 2013) che rappresenta la potenziale sottocopertura degli archivi anagrafici.

Tavola 4 – Confronto tra i risultati dell'identificazione dei gruppi di popolazione con T centrati al 31.12.2012 e al 31.12.2013

Gruppo	Descrizione	2012	2013	Var.%
A1	Presenti in ANVIS con segnali di LS	36.270.975	35.052.647	-3,4
A2	Non presenti in ANVIS con segnali di LS, presenti nelle LAC	347.134	216.999	-37,5
B	Presenti in ANVIS senza segnali di Lavoro e Studio	24.471.690	25.709.372	5,1
	di cui:			
B1	Presenti nel CP o TNP	14.506.721	14.751.634	1,7
B2	Non presenti nel CP e TNP	9.964.969	10.957.738	10,0
	di cui:			
B2.1	Come "coniuge"	3.348.647	3.365.550	0,5
	Come "figlio a carico"	3.231.972	3.364.909	4,1
	Come "altro famigliare a carico"	77.790	77.789	0,0
	Come "Coniuge" nel nucleo LAC di intestatari con segnali di LS	280.646	293.861	4,7
B2.2	Presenti in ANVIS senza segnali da altre fonti	3.025.914	3.855.629	27,4
C	Con segnali di LS non presenti né in ANVIS né in LAC	1.086.162	1.003.136	-7,6
D1+D2	Con PS senza segnali in ANVIS, senza segnali di LS	457.574	425.887	-6,9

Fonte: nostra elaborazione su dati Istat

Nel complesso, la sperimentazione condotta sui due anni mostra una sostanziale stabilità dei risultati non solo osservando le singole poste degli aggregati presi in esame ma anche dall'analisi dei profili dei gruppi (Tavola 5).

[4] Una quota di questo incremento, pari all'8,8%, è rappresentata da individui detentori di Partita IVA presenti solo nelle fonti di lavoro del 2012. L'informazione relativa al 2013 non era disponibile al momento dell'analisi.

[5] Il confronto con i segnali indiretti ricavati dalle dichiarazioni fiscali per il 2013 è stata effettuato utilizzando i segnali delle dichiarazioni relative al 2012 in quanto non ancora disponibili al momento dell'analisi.

La distribuzione per sesso, età e cittadinanza dei singoli gruppi identificati al 2012 e al 2013 appare identica e presenta delle differenze solo in corrispondenza di due sottopopolazioni: il gruppo B2.2 degli individui presenti in ANVIS senza segnali da altre fonti, che presenta una significativa differenza nella classe di età "minori di 18" e di cui si è già parlato sopra; il gruppo C degli individui non presenti in ANVIS con segnali di studio e di lavoro, che mostra significative differenze in riferimento alla composizione per cittadinanza. Su quest'ultimo punto è il caso di osservare che dal 2012 al 2013 aumenta il numero di cittadini italiani non iscritti in anagrafe con segnali di lavoro e studio (circa 37 mila unità in valore assoluto) e, invece, diminuisce quello di stranieri (circa 170 mila unità in meno).

Tavola 5 – Confronto tra le variabili anagrafiche dei gruppi di popolazione identificati al 2012 e al 2013

Gruppo	Descrizione	Anno T	Sesso		Cittadinanza		Classi di Età					
			M	F	Italiani	Stranieri	<18	18-25	25-35	35-45	45-65	>=65
A1	Presenti in ANVIS con segnali di LS	2012	54,4	45,6	91,2	8,8	18,9	10,0	15,7	20,9	31,7	2,8
		2013	54,3	45,7	91,1	8,9	18,0	10,1	15,4	20,6	32,9	2,9
B	Presenti in ANVIS senza segnali di Lavoro e Studio	2012	39,8	60,2	93,4	6,6	12,3	2,6	5,4	7,4	22,6	49,8
		2013	40,6	59,4	92,6	7,4	13,6	2,7	5,8	7,6	21,9	48,3
B1	Presenti nel CP o TNP	2012	43,6	56,4	98,7	1,3	0,3	0,4	1,7	2,1	17,4	78,0
		2013	43,9	56,1	98,5	1,5	0,4	0,5	1,9	2,6	16,5	78,0
B2	Non presenti nel CP e TNP	2012	34,2	65,8	85,5	14,5	30,0	5,9	10,9	15,3	29,8	8,1
		2013	36,2	63,8	84,7	15,3	31,3	5,7	11,0	14,3	29,2	8,4
B2.1	Non presenti nel CP e TNP a "carico" nelle DF e "coniugi" nelle LAC	2012	27,3	72,7	89,4	10,6	35,6	5,4	9,0	13,5	27,6	8,9
		2013	28,1	71,9	89,2	10,8	34,5	5,8	9,3	12,8	28,0	9,6
B2.2	Presenti in ANVIS senza segnali da altre fonti	2012	50,4	49,6	76,2	23,8	16,7	7,1	15,4	19,6	35,0	6,1
		2013	51,3	48,7	76,1	23,9	25,5	5,6	14,2	17,2	31,4	6,1
C	Con segnali di LS non presenti né in ANVIS né in LAC	2012	60,1	39,9	10,4	89,6	10,3	16,7	30,8	21,8	19,1	1,3
		2013	60,7	39,3	17,5	82,5	9,2	16,4	28,7	22,2	20,7	2,8
D1+D2	Con PS senza segnali in ANVIS, senza segnali di LS	2012	51,3	48,7	0,4	99,6	3,2	16,8	27,6	22,3	23,4	6,6
		2013	52,6	47,4	0,4	99,6	3,3	17,3	27,2	22,2	23,4	6,7

Fonte: nostra elaborazione su dati Istat

3.10 L'analisi delle sottopopolazioni critiche individuate al 31.12.2012

La sperimentazione ha consentito di individuare alcune specifiche sottopopolazioni, per le quali la dimora abituale in Italia risulta "dubbia", che richiedono ulteriori analisi di approfondimento. Nel corso dei lavori preparatori per il Censimento 2011, l'Istituto Nazionale di Statistica spagnolo (INE) individuò un aggregato di popolazione molto simile a quello emerso da questa sperimentazione che orientò le scelte sulla strategia censuaria da adottare. Integrando i dati dell'archivio anagrafico con i dati provenienti dagli archivi amministrativi disponibili, l'INE valutò che circa il 2,2% della popolazione rappresentava dei "casi dubbi" dal momento che per questa quota di popolazione non era possibile rinvenire "segnali" di presenza significativi dagli archivi amministrativi (Argüeso et al., 2014). Questa attività preliminare portò l'INE ad adottare un modello di rilevazione che combinava l'uso dei registri amministrativi con indagini di campo.

In quest'ottica anche per l'Italia risulta necessario, in vista della prossima tornata censuaria, approfondire l'analisi sulle principali caratteristiche sociodemografiche e la distribuzione sul territorio degli individui per i quali si rileva una "dimora abituale dubbia".

3.10.1 Gli individui presenti negli archivi anagrafici senza alcun segnale in altre fonti amministrative

Al 2012, circa 3 milioni di persone compongono la sottopopolazione che risulta iscritta in anagrafe ma senza segnali di studio o di lavoro in altre fonti amministrative (Gruppo B2.2). Tre individui su quattro possiedono la cittadinanza italiana. La distribuzione di genere appare ben bilanciata sia per gli italiani che per stranieri; l'età mediana è di poco superiore ai 40 anni, ma gli stranieri sono di oltre 6 anni più giovani degli italiani (Tavola 6).

La distribuzione per classi di età di questa sottopopolazione riflette perfettamente quella del totale della popolazione residente in Italia al 2012. Nel complesso, la quota di cittadini italiani in età più adulta (35-64 anni) è di circa 14 punti percentuali più elevata rispetto a quella degli stranieri.

Inoltre, l'analisi sulla distribuzione per età consente di far emergere una quota di bambini con meno di 6 anni molto consistente (poco meno del 15% sia per gli italiani che per gli stranieri) da imputare prevalentemente alle nascite verificatesi successivamente alle dichiarazioni fiscali (Cfr. par. 3.10). In questo caso, si tratta complessivamente di oltre 400 mila bambini che andrebbero sottratti a questa sottopopolazione, dato che essi appartengono certamente ad un nucleo familiare e, in quanto tali, non possono rappresentare casi con dimora abituale "dubbia". La loro assenza dagli archivi amministrativi è da imputare semplicemente ad un ritardo nella disponibilità della fonte delle dichiarazioni fiscali.

Tavola 6 – La struttura demografica degli individui presenti negli archivi anagrafici senza alcun segnale nelle altre fonti

PAESE DI CITTADINANZA	Valori assoluti (Migliaia)	Valori Percentuali	Rapporto di genere	Classi di età (%)						
				Meno di 6 anni	6-14 anni	15-34 anni	35-64 anni	65 e oltre	Totale	Età Mediana
Totale	**3,023**	**100,0**	**101,3**	14,3	1,5	22,7	55,3	6,3	100,0	**40,6**
Italiani	2,223	73,5	101,4	13,4	0,8	21,5	58,5	5,8	100,0	42,1
Stranieri (*di cui*):	800	100,0	101,3	17,0	3,5	26,6	44,7	8,1	100,0	35,7
Romeni	134	16,7	90,3	20,0	2,4	34,5	40,8	2,4	100,0	31,9
Marocchini	72	9,1	159,6	19,7	2,9	22,5	46,9	8,0	100,0	36,7
Albanesi	60	7,5	96,1	16,0	1,4	19,0	40,6	23,0	100,0	48,9
Cinesi	34	4,3	107,7	40,2	10,0	15,2	31,3	3,2	100,0	13,2
Ucraini	20	2,6	45,3	10,5	2,3	24,8	56,9	5,5	100,0	40,6
Tunisini	17	2,2	192,7	18,8	6,4	24,0	47,5	3,3	100,0	34,5
Egiziani	17	2,2	201,5	22,0	12,4	23,2	40,8	1,6	100,0	29,8
Polacchi	17	1,9	48,2	7,7	3,4	26,2	59,2	3,5	100,0	38,0
Filippini	15	2,2	78,7	27,2	5,3	22,3	40,2	5,1	100,0	30,5
Nigeriani	15	1,9	99,1	24,3	2,4	38,2	34,0	1,1	100,0	29,9
Altri	396	49,5	98,1	12,2	3,2	27,1	47,5	9,9	100,0	37,6

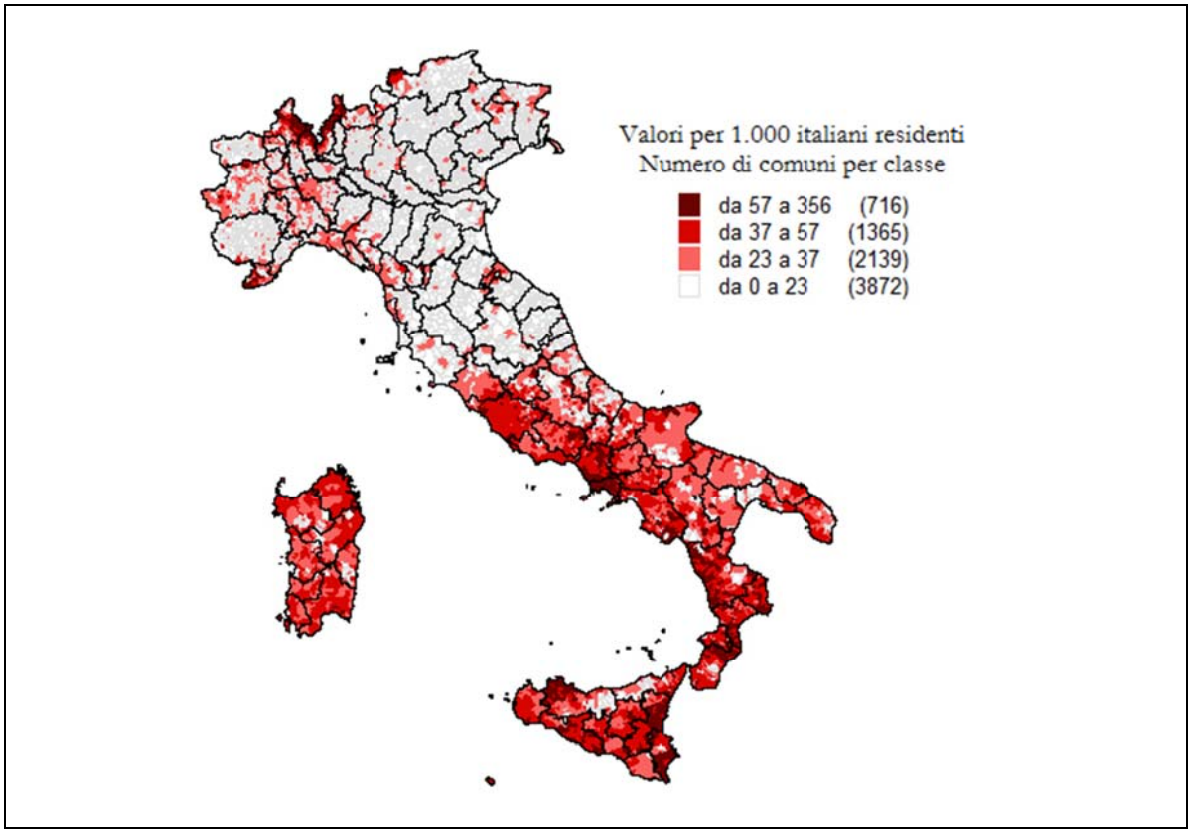
Fonte: nostra elaborazione su dati Istat

La distribuzione territoriale di questa sottopopolazione mostra una forte concentrazione nei comuni del centro e del sud Italia dove, per altro, la disoccupazione di lunga durata e la perdita di posti di lavoro generati dalla crisi del 2008 sono stati più consistenti.. Tuttavia emergono differenze significative in base alla cittadinanza.

Gli italiani si concentrano nei comuni di Lazio, Campania, Sicilia e Calabria, ma quote medio-alte di persone con cittadinanza italiana si trovano anche nei comuni del Nord, lungo i confini di Svizzera, Francia e Austria (Figura 2). È molto probabile che questo gruppo di individui identifichi il fenomeno dei lavoratori frontalieri.

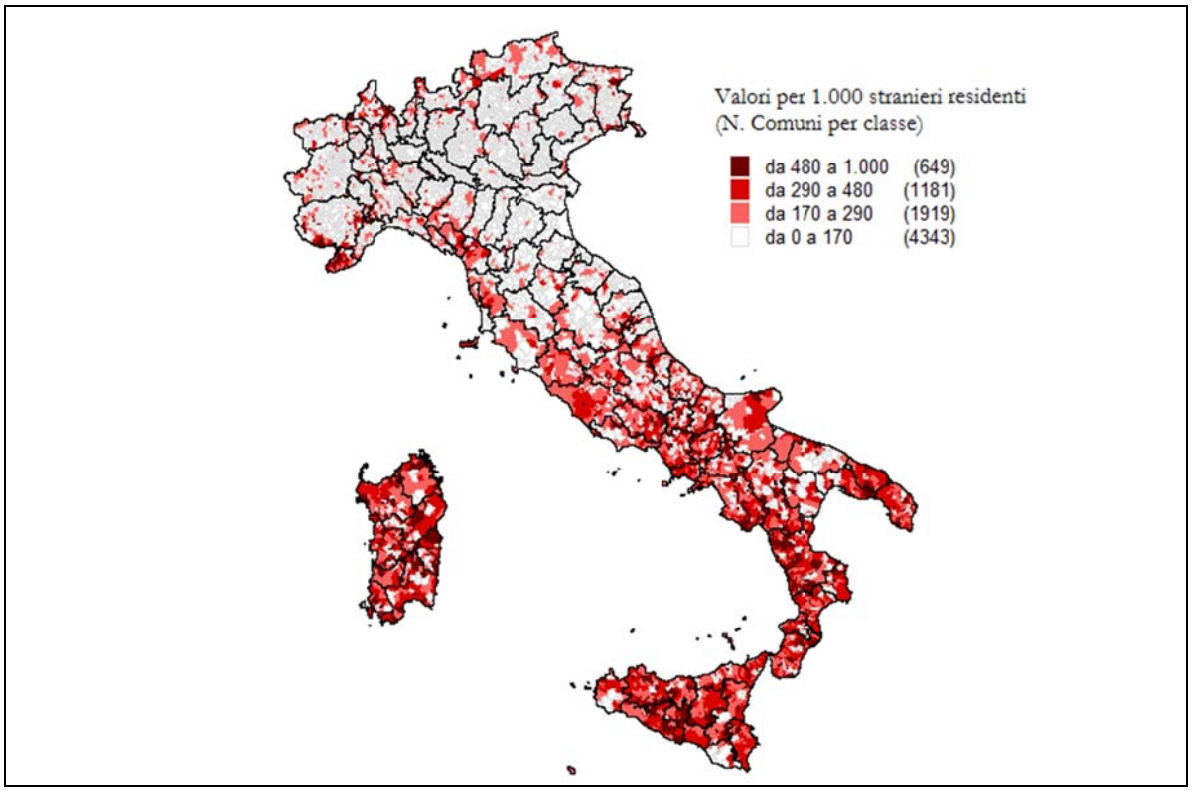
Invece, i cittadini stranieri, oltre ad una significativa concentrazione nei comuni delle isole e nelle aree centro-meridionali (dal Lazio alla Calabria), presentano una distribuzione territoriale più diffusa degli italiani nei comuni del Nord-ovest e, soprattutto, lungo i comuni costieri che vanno dalla Toscana alla Liguria (Figura 3).

Figura 2 – Gli italiani con dimora abituale dubbia



Fonte: nostra elaborazione su dati Istat

Figura 3 – Gli Stranieri con dimora abituale dubbia



Fonte: nostra elaborazione su dati Istat 2016

Quest'ultimo fenomeno è da mettere in relazione con i pensionati stranieri dell'Unione Europea, prevalentemente britannici e tedeschi, che ormai da decenni hanno scelto come dimora abituale i comuni della Toscana (Warner et al., 1999).

Osservando la struttura per età in base al Paese di cittadinanza, la comunità cinese mostra la composizione per età più giovane (l'età mediana è di soli 13 anni) sia rispetto agli italiani che agli stranieri. Tra questi ultimi, gli albanesi sono la comunità con l'età mediana più elevata (quasi 49 anni).

Per gli stranieri, se il rapporto tra i sessi (uomini su 100 donne) riflette la composizione di genere delle comunità straniere in Italia, la distribuzione per Paese di cittadinanza di questa sottopopolazione non riproduce esattamente la geografia della presenza straniera in Italia. Ad esempio, gli immigrati provenienti dalla Moldavia, dall'India e dal Perù non risultano nella graduatoria delle prime dieci cittadinanze residenti in Italia (Tavola 6).

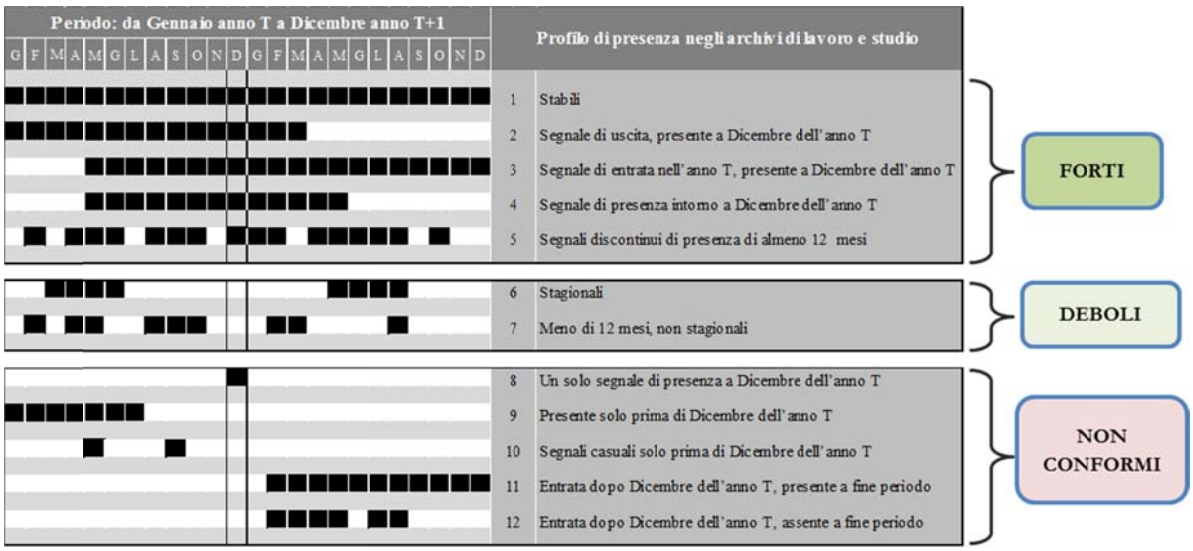
In conclusione, questo gruppo rappresenta, ai fini del Censimento permanente, una sottopopolazione critica di notevole importanza. Il suo ammontare, l'assenza di segnali nelle fonti amministrative di studio e di lavoro e la forte concentrazione nei comuni del Mezzogiorno rappresentano forti elementi di rischio ai fini del conteggio, senza escludere, per altro, che questa sottopopolazione potrebbe alimentare, a seguito della crisi economica tuttora in corso, consistenti flussi migratori verso l'estero che, in caso di mancate notifiche presso gli Uffici di anagrafe, possono generare una sovracopertura sostanziale dei registri di popolazione.

3.10.2 Gli individui non presenti negli archivi anagrafici con segnali di lavoro e studio

Il gruppo C è composto da 1,1 milioni di persone ed esprime la sottocopertura potenziale delle anagrafi comunali. Infatti, questa sottopopolazione si compone degli individui non iscritti in anagrafe ma con segnali di presenza provenienti dagli archivi di lavoro e di studio considerati. Dal momento che i segnali possono essere riferiti a presenze occasionali oppure a persone che hanno la dimora abituale fuori dall'Italia, è opportuno suddividere ulteriormente questo gruppo in base al tipo e alla continuità del segnale di attività in Italia, in modo da far emergere gli individui per i quali la dimora abituale in Italia risulta più probabile.

A questo scopo, consideriamo un raggruppamento dei singoli profili di continuità dei segnali, già proposti nel Prospetto 3 del paragrafo 3.8, in base alla maggiore o minore corrispondenza con i requisiti minimi per la dimora abituale in Italia stabiliti dalla definizione internazionale. In tal senso, è possibile individuare (Prospetto 5) dei segnali:

Prospetto 5 - Profili di durata dei segnali classificati in base alla definizione internazionale di dimora abituale



- "forti", descritti nei profili dall' 1 al 5 che sono i più rispondenti alla definizione ufficiale di dimora abituale, dal momento che attestano 12 mesi di presenza, compreso il mese di dicembre 2012.
- "deboli" descritti nei profili 6 e 7 dello schema mensile che mostrano la presenza di segnali stagionali e intermittenti (non stagionali) per una durata complessiva inferiore ai dodici mesi.
- "non conformi" nei profili dall' 8 al 12, che descrivono una durata, inferiore ai 12 mesi, non rispondente ai requisiti stabiliti dalla definizione internazionale, oppure perché presenti solo in mesi precedenti o soltanto in mesi successivi alla data di riferimento.

Se si suddivide l'insieme di tutti gli individui non presenti nei registri anagrafici ma con segnali provenienti da altri archivi sulla base di quest' ultima classificazione relativa alla "forza" dei segnali, si individuano i sottogruppi evidenziati nella Tabella 7.

Tavola 7 - Individui con segnali di lavoro e studio non presenti negli archivi anagrafici (Gruppo C) secondo il tipo di fonte e di segnale

Gruppi di Sottopopolazione	Fonte e tipo di segnale	Valori assoluti
C1	Lavoratori con segnali stabili	318.159
C2	Studenti universitari iscritti alla data di riferimento	32.671
C3	Studenti scuole primarie e secondarie	58.327
C4	Individui con segnali deboli	266.763
C5	Individui senza segnale temporale	61.648
C6	Individui con segnali non conformi	348.594
TOTALE		**1.086.162**

Fonte: nostra elaborazione su dati Istat

I segnali "forti" sono associati a 409.157 individui che verosimilmente rappresentano la sottocopertura del registro anagrafico. Il 78% di essi, pari a 318.159 individui (C1), è costituito da lavoratori (almeno una delle fonti di origine del segnale è una fonte relativa al lavoro); il restante 22% è composto da studenti delle scuole primarie o secondarie (58.327 individui, gruppo C3) e da studenti universitari (32.671 individui, gruppo C2). Circa il 90% di tali individui è composto da stranieri.

I segnali "deboli" sono associati a 266.763 individui del sottogruppo C4: sebbene la durata complessiva di tali segnali sia inferiore al periodo minimo richiesto per la dimora abituale in Italia, questa sottopopolazione necessita di ulteriori approfondimenti dal momento che i segnali stagionali o intermittenti potrebbero essere espressione di specifici tipi di attività di lavoro o di studio e non essere quindi necessariamente legate alla permanenza sul territorio.

Infine, vi sono 61.648 individui senza alcuna informazione sulla presenza mensile (C5) e altri 348.594 che presentano segnali "non conformi" alla definizione internazionale: per tali gruppi sono necessarie maggiori informazioni, da desumere dall'analisi longitudinale su un periodo di tempo superiore ai 24 mesi, da ulteriori indagini o da specifici modelli statistici.

Per tracciare i profili di questi sottogruppi e identificare eventuali *clusters* specifici al loro interno sono state effettuate una serie di analisi di approfondimento, utilizzando congiuntamente le variabili relative ai segnali (fonte, durata) e le caratteristiche demografiche degli individui: classe d'età, stato estero di cittadinanza e di nascita, sesso.

Un menzione particolare merita la valutazione dei valori mancati sulle variabili anagrafiche ricavate, come visto nel par. 3.7, dal SIM, che risultano piuttosto frequenti nel caso del gruppo C (Tavola 8 e Tavola 9) proprio perché composto da individui non presenti nei registri anagrafici.

L'informazione sulla cittadinanza è carente in tutte le fonti e i gruppi di questa sottopopolazione. Pertanto, il Paese estero di nascita è l'unica informazione *proxy* del Paese di cittadinanza con la conseguenza che per alcune sottopopolazioni, come i nati in Italia da famiglie straniere, viene erroneamente attribuita la cittadinanza italiana.

Tavola 8 - Percentuale di dati mancanti nel gruppo C per variabile e fonte del segnale

VARIABILI ANAGRAF.	FONTI										
	Autonomi Agricoltura	Autonomi da BDR	Cedolini	Lavoratori Agricoltura	Domestici	Emens	Inpdap	Interinali	Para subordi-nati	Studenti	Universitari
Età	37,3%	34,0%	32,4%	0,2%	0,0%	21,4%	2,0%	76,2%	48,3%	16,9%	14,8%
Sesso	37,3%	33,7%	36,5%	0,2%	0,0%	17,6%	2,1%	76,2%	46,2%	16,9%	14,8%
Cittadinanza	100,0%	96,8%	100,0%	100,0%	100,0%	100,0%	100,0%	88,8%	99,9%	99,0%	95,2%
Paese di Nascita	42,5%	48,3%	42,0%	24,4%	0,5%	24,0%	18,8%	76,6%	49,5%	18,1%	15,6%
Provincia	0,7%	81,0%	28,4%	0,0%	0,0%	1,2%	30,6%	5,0%	11,8%	0,0%	0,0%

Tavola 9 - Percentuale di dati mancanti per variabile e sottopopolazione

VARIABILI ANAGRAF.	C1	C2	C3	C4	C5	C6
Età	6,7%	14,9%	0,1%	15,2%	21,6%	52,50%
Sesso	6,6%	14,9%	0,1%	14,6%	17,3%	52,50%
Cittadinanza	100,0%	95,7%	100,0%	99,9%	99,7%	93,50%
Paese di Nascita	10,7%	16,0%	2,2%	19,7%	32,7%	53,40%
Provincia	6,0%	0,0%	0,0%	0,8%	2,7%	47,20%

Un risultato di particolare rilievo deriva dal confronto di questa sottopopolazione con l'archivio dell'Anagrafe Tributaria (AT) che contiene l'informazione relativa al domicilio fiscale[6]. Il 74% degli individui che risultano sottocoperti negli archivi anagrafici è presente nell'AT con l'informazione sul Comune e l'Indirizzo e, in particolare, nel 47% dei casi il Comune di domicilio fiscale è diverso da quello della fonte di provenienza che rappresenta il luogo dell'attività dell'individuo.

3.10.3 Caratteristiche demografiche dei sottogruppi C

Se si analizzano i sottogruppi in base alle principali caratteristiche demografiche, emergono alcuni aspetti interessanti nella connotazione delle sottopopolazioni e soprattutto questioni da tener presente nella predisposizione della strategia censuaria.

Un primo aspetto interessante riguarda la distribuzione per Stato estero degli studenti non iscritti in anagrafe (Tavola 10). Non essendo disponibile lo Stato estero di cittadinanza, è stato utilizzato lo Stato estero di nascita. Ben il 32% dei non iscritti in anagrafe ma frequentanti la scuola primaria o secondaria sono nati in Italia: si tratta verosimilmente di cittadini di famiglie straniere nati in Italia (questo dato è confermato dall'analisi dei pochi casi in cui si dispone dello stato estero di cittadinanza). La comunità straniera più rappresentata nel caso di scuola primaria o secondaria è quella rumena (18,6%); invece nel caso degli studenti universitari non iscritti in anagrafe si ha una concentrazione di cinesi (14,5 %).

Mentre i sottogruppi C2 e C3 presentano una omogeneità relativamente alla fascia di età e alla fonte di origine del segnale (per definizione, archivi del MIUR), gli altri sottogruppi provengono da varie fonti e hanno una notevole variabilità in termini di età.

[6] Secondo l'art. 58 del DPR 600/73, per le persone fisiche iscritte all'anagrafe dei residenti il domicilio fiscale è nel comune in cui sono iscritti; per le persone fisiche non residenti, tassate sulla base del principio della fonte, il domicilio fiscale è nel comune in cui è prodotta la parte prevalente del reddito.

Tavola 10 – Distribuzioni per Stato estero di nascita degli studenti non iscritti in anagrafe

Paese estero di nascita	C3 – Scuola primaria e secondaria	C2 – Studenti Universitari
Valore mancante	2,2%	16,0%
Italia	31,5%	14,4%
Romania	18,6%	0,7%
Cina	6,2%	14,5%
Nordafrica	5,5%	2,7%
Stati Uniti	1,0%	0,5%
Albania e Paesi ex-Jugo.	5,3%	5,7%
Paesi Est-Europa e Russia	7,1%	3,8%
Filippine	1,3%	0,1%
Germania/Spagna	1,7%	2,3%
Sudamerica	4,5%	2,1%
MedioOriente	0,3%	5,4%
Francia	0,6%	1,7%
Svizzera	0,7%	2,1%
Altri Paesi	13,7%	28,0%

Fonte: nostra elaborazione su dati Istat

Tavola 11 – Caratteristiche demografiche per i gruppi degli individui non iscritti in anagrafe, ma presenti in altri archivi sul lavoro

Variabili anagrafiche e tipo di fonte	C1 - Lavoratori con segnali stabili	C4 - Individui con segnali deboli	C5 - Lavoratori senza segnale temporale	C6 - Individui con segnali non conformi
		Età		
Mancante	7,4%	15,7%	58,0%	22,9%
meno6	0,0%	0,0%	0,0%	1,3%
6_14	0,0%	0,1%	0,2%	5,2%
15-18	0,2%	0,3%	2,1%	2,2%
19-35	50,1%	49,1%	32,9%	41,1%
36-65	41,7%	34,5%	6,8%	27,0%
Oltre 65	0,6%	0,3%	0,0%	0,3%
		Stato estero di nascita		
Mancante	11,8%	20,3%	59,0%	34,7%
Romania	17,8%	**32,9%**	1,9%	23,6%
Cina	11,4%	5,1%	5,3%	3,5%
Nord-Africa	14,7%	9,5%	3,0%	9,5%
Stati Uniti	0,2%	0,2%	1,2%	0,4%
Albania e ex Jugoslavia	4,1%	5,7%	2,8%	4,1%
Altri Est Europa	14,4%	13,4%	3,2%	7,5%
Filippine	2,4%	0,3%	0,1%	1,2%
Spagna e Germania	1,5%	1,0%	1,9%	1,5%
Sudamerica	1,7%	0,6%	1,7%	1,4%
Medioriente	0,3%	0,2%	3,1%	0,3%
Francia	0,5%	0,4%	1,3%	0,5%
Svizzera	0,3%	0,1%	0,6%	0,2%
Altri paesi	18,9%	10,5%	15,1%	11,6%
		Sesso		
Mancante	7,3%	15,1%	58,0%	18,4%
Donne	38,4%	30,8%	20,3%	33,6%
Uomini	54,3%	54,2%	21,7%	48,1%
		Fonti sul lavoro		
Mancante	0,0%	0,0%	0,0%	0,0%
Autonomi dell'agricoltura	0,6%	0,0%	0,0%	0,0%
Autonomi	5,1%	0,0%	50,9%	1,4%
Cedolini	0,1%	0,0%	0,0%	0,1%
Dipendenti dell'agricoltura	9,1%	**34,2**%	0,0%	22,0%
Domestici	**33,9%**	7,1%	0,0%	14,7%
Emens (lavoratori dipendenti)	47,9%	53,3%	0,0%	44,9%
Inpdap	0,1%	0,0%	0,6%	0,3%
Lavoratori Interinali	0,1%	0,9%	0,4%	0,7%
Parasubordinati	3,2%	2,5%	1,9%	4,5%
Studenti	0,0%	2,0%	13,5%	11,5%
Universitari	0,0%	0,0%	32,7%	0,0%

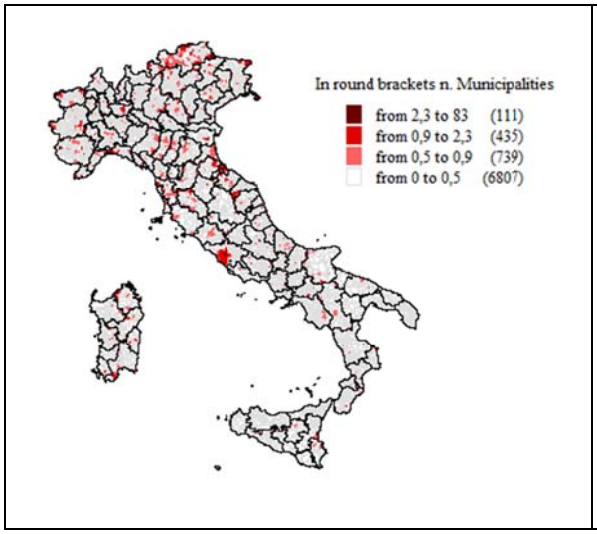
Fonte: nostra elaborazione su dati Istat

La Tavola 11 mostra le distribuzioni univariate di alcune variabili per i sottogruppi C1 (lavoratori stabili), C4 (segnali deboli), C5 (assenza di informazione temporale) e C6 (segnali non conformi).

In particolare, si nota una forte concentrazione dei rumeni nei sottogruppi C4 e C6, caratterizzati da profili di continuità temporale non stabili. Inoltre, si osserva una forte concentrazione di segnali provenienti dall'archivio dei lavoratori dipendenti in agricoltura (DMG) nel sottogruppo C4 dei segnali deboli, mentre nel sottogruppo C1 prevalgono i lavoratori domestici (33,9%).

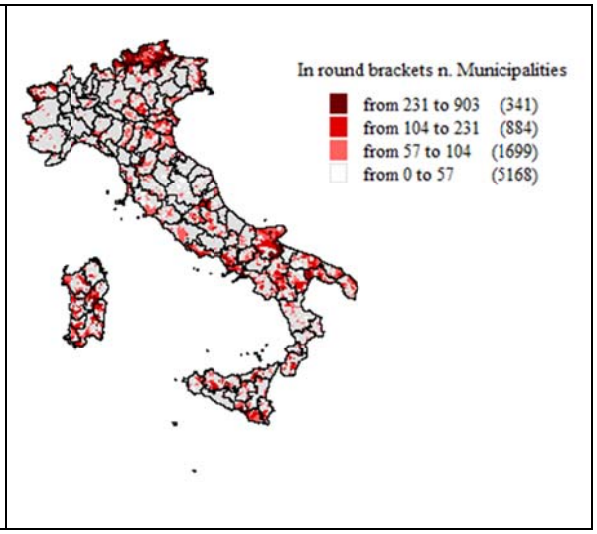
La distribuzione territoriale del sottogruppo C1 mostra una lieve prevalenza di italiani (Figura 5) lungo i confini con gli stati esteri dei comuni del nord, intorno alla città di Roma e alla Repubblica di San Marino; gli stranieri (Figura 6) sono maggiormente presenti in Trentino-Alto Adige e poi nei comuni della pianura Padana, della Puglia, della Campania, della Sicilia e della Sardegna. Questa distribuzione lascia supporre l'esistenza di clusters diversi, uno più legato al fenomeno dei cosiddetti frontalieri e l'altro invece relativo a una presenza straniera irregolare.

Figura 5 – Sottopopolazione C1: Lavoratori e studenti italiani (*Valori per 1.000 cittadini italiani*)



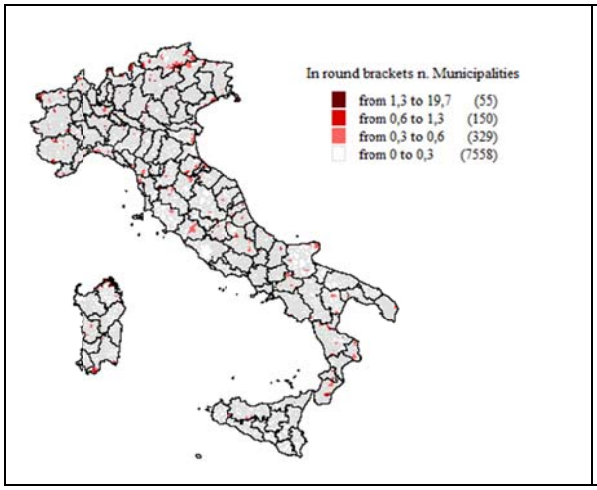
Fonte: nostra elaborazione su dati Istat

Figura 6 - Sottopopolazione C1: Lavoratori e studenti stranieri (*Valori per 1.000 cittadini stranieri*)



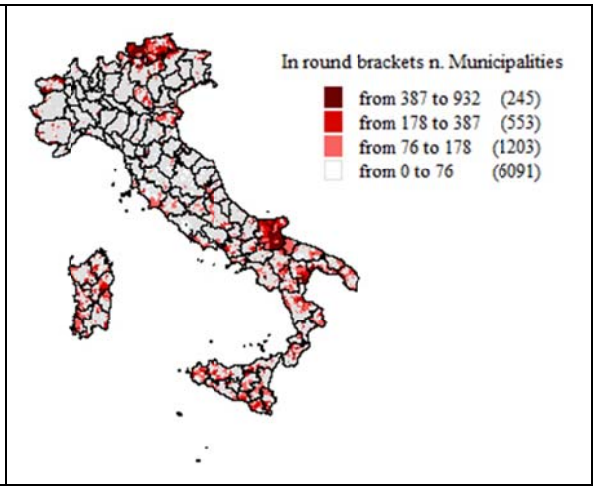
Fonte: nostra elaborazione su dati Istat

Figura 7 - *Sottopopolazione* con segnali deboli (C4) e luogo di nascita Italia (*Valori per 1.000 cittadini italiani*)



Fonte: nostra elaborazione su dati Istat

Figura 8 – Sottopopolazione con segnali deboli (C4) e luogo di nascita Estero (*Valori per 1.000 cittadini stranieri*)



Fonte: nostra elaborazione su dati Istat

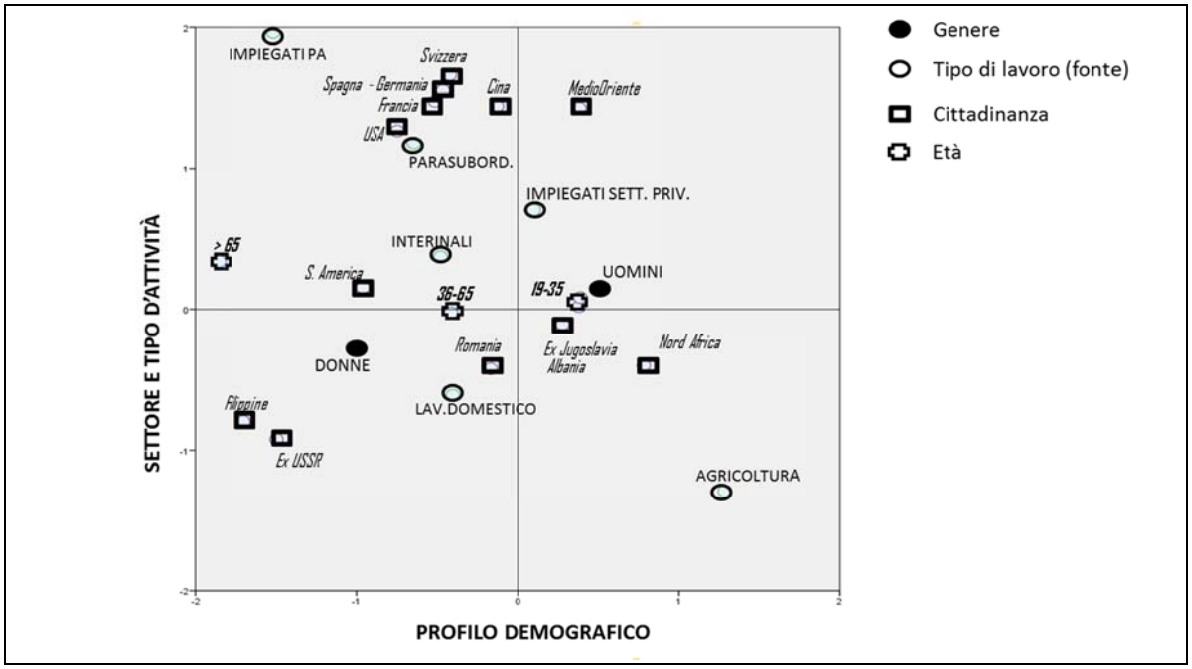
La distribuzione territoriale degli italiani con segnali deboli (C4) riflette di gran lunga quella che caratterizzava il gruppo con segnali forti (Figura 7), mentre nel caso degli stranieri è possibile osservare una più forte concentrazione nei comuni della Puglia settentrionale (provincia di Foggia), oltre che nelle isole e nei comuni del Nord collocati lungo il versante alpino (Figura 8).

3.10.4 Analisi multivariata delle caratteristiche dei sottogruppi e individuazione dei clusters specifici

Per spingersi oltre nella individuazione dei clusters si è ritenuto opportuno procedere ad una analisi multivariata delle caratteristiche demografiche e dei segnali. È stata impiegata l'analisi delle corrispondenze multiple (ACM): le variabili considerate sono sesso, età, cittadinanza, Provincia e fonte amministrativa del segnale.

Per iniziare, è stata condotta l'analisi sui cittadini stranieri del sottogruppo C1. Tale analisi ha evidenziato due fattori, come rappresentato nel plot della Figura 9: l'asse orizzontale rappresenta il profilo demografico, con i quadranti a sinistra caratterizzati dalla presenza delle donne e delle fasce di età più alte, mentre i quadranti di destra sono associati a uomini, più giovani; l'asse verticale è legato alla condizione occupazionale e al settore professionale, con gli occupati nel settore agricolo e domestico che si posizionano nella parte inferiore, mentre gli altri lavoratori dipendenti o autonomi del settore pubblico o privato si concentrano nella parte superiore.

Figura 9 – Sottopopolazione C1: lavoratori stranieri con segnali forti e stabili



Fonte: nostra elaborazione su dati Istat

Per ragioni di leggibilità, nel plot non è proiettata la dimensione territoriale, che però ha costituito una variabile attiva dell'analisi rilevante nella determinazione dei clusters come si può evincere dalle descrizioni riportate nelle righe seguenti.

L'analisi dei risultati della ACM permette di identificare alcune specifiche sottopopolazioni, osservabili anche dal grafico:

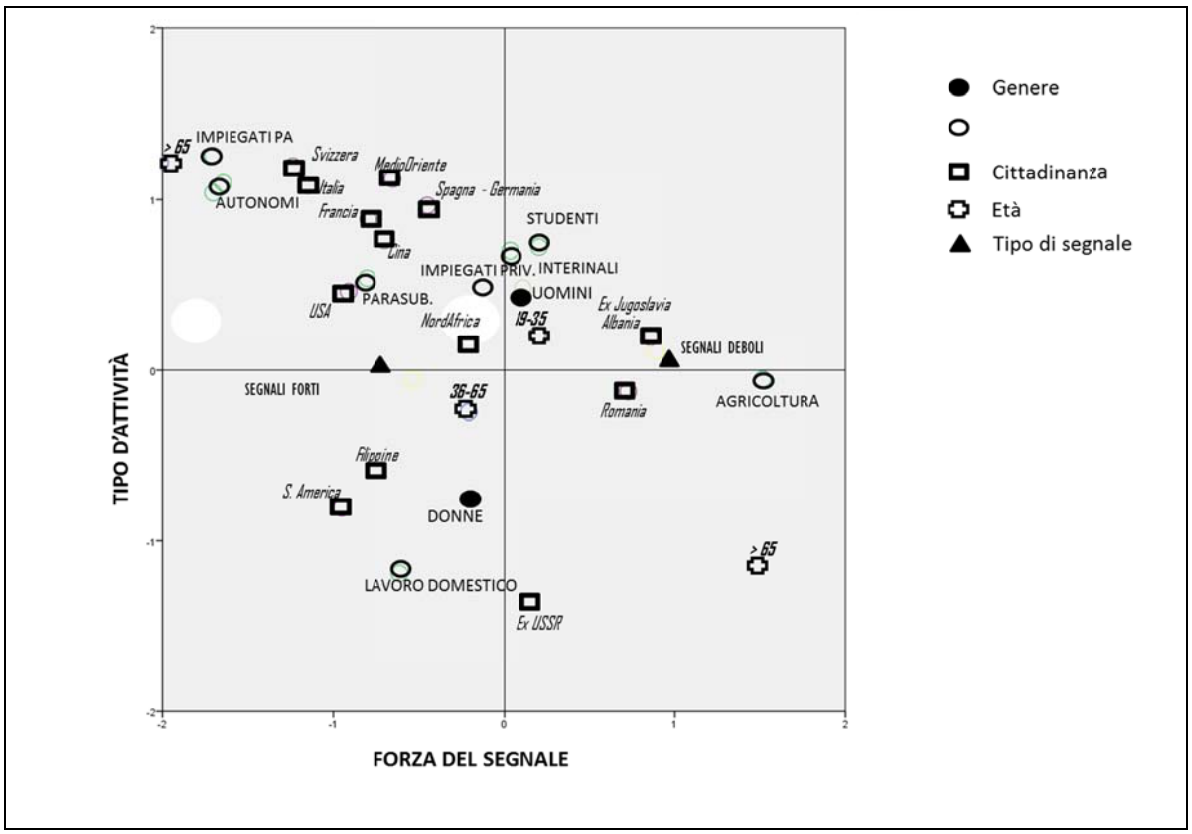
1. Uomini giovani impiegati nel settore agricolo, provenienti principalmente dal nord Africa e in parte dai territori dell'Albania e dell'ex Jugoslavia, i cui segnali si concentrano nelle province del Sud Italia dove l'agricoltura rappresenta il principale settore occupazionale;
2. Donne impiegate come badanti o colf, di età superiore ai 35 anni, con cittadinanza filippina, di Paesi dell'Europa orientale (ex-URSS) o sudamericana, distribuite sul territorio principalmente nelle province dei grandi comuni;

3. Un gruppo consistente di individui proviene dai Paesi dell'Europa occidentale e centrale. Si tratta di lavoratori dipendenti, lavoratori autonomi, consulenti o lavoratori interinali, di anziani con più di 65 anni e concentrati nelle province del centro e del nord Italia;
4. I cinesi si trovano nello stesso quadrante della popolazione dell'Europa occidentale;
5. I romeni non presentano differenze di genere né di età ma si concentrano nei settori agricolo e delle attività domestiche;
6. Gli italiani, i francesi e gli svizzeri si concentrano nelle del Nord di confine. Probabilmente si tratta di lavoratori frontalieri che vivono all'Estero ma lavorano in Italia e pertanto non sono registrati nell'Anagrafe pur mostrano forti segnali di presenza sul territorio legati al lavoro.

A seguire, lo stesso tipo di analisi è stata effettuata sul sottogruppo C4, ovvero le persone con segnali "deboli". Dai risultati è emersa la presenza degli stessi fattori e delle stesse associazioni tra variabili individuate per il sottogruppo C1: si è ritenuto opportuno sviluppare un'analisi congiunta dei due sottogruppi di popolazioni con segnali sia forti che deboli (Figura 10).

I risultati mostrano che tra i fattori discriminanti, oltre al profilo demografico e alla condizione professionale, la stabilità del segnale viene a configurarsi come un'ulteriore dimensione significativa. Difatti, nel plot è possibile osservare come sull'asse orizzontale si contrappongano, su quadranti opposti, i segnali forti da una parte e i segnali deboli dall'altra.

Figura 10 – Sottopopolazioni con segnali forti e deboli (C1 e C4): correlazioni tra variabili



Fonte: nostra elaborazione su dati Istat

I clusters che emergono sono molto simili a quelli evidenziatisi in precedenza per il solo sottogruppo C1 ma ciò che emerge è la forte correlazione tra segnali deboli e specificità connesse al lavoro e al Paese di cittadinanza. In particolare, tale associazione determina il gruppo dei rumeni, degli albanesi e di coloro che provengono da paesi dell'ex-Jugoslavia, di giovane età, impiegati in agricoltura nelle province agricole del sud Italia.

3.10.5 Gli individui con permesso di soggiorno non presenti né negli archivi anagrafici, né negli archivi di lavoro e studio

La popolazione di questo gruppo si compone di due aggregati: a) stranieri con permesso di soggiorno e iscrizione in anagrafe desunta dalla fonte LAC (Gruppo D1), con un ammontare pari a poco più di 100 mila individui; b) stranieri non iscritti in anagrafe con permesso di soggiorno (Gruppo D2), con un ammontare di oltre 351 mila individui. Questo secondo gruppo presenta come unico segnale la titolarità a soggiornare in Italia ma si caratterizza per una durata del permesso che in due casi su tre è pari a 12 mesi; si tratta, quindi, di segnali piuttosto deboli ai fini della dimora abituale in Italia.

Conclusioni

Le attività svolte nell'ambito della sperimentazione sull'uso dei dati amministrativi per il Censimento permanente hanno prodotto alcuni primi importanti risultati, ma hanno fatto emergere anche alcune criticità. È necessario, quindi, continuare nel percorso della sperimentazione sia per trovare soluzioni alle problematiche emerse sia per fornire indicazioni più significative sulla strategia censuaria da adottare.

Un primo risultato ottenuto riguarda l'approccio della Knowledge Discovery da Database che ha mostrato la sua efficacia nella capacità di sfruttare al meglio la ricchezza informativa contenuta nella fonti amministrative. Un ulteriore supporto al processo decisionale dell'Istituto potrebbe consistere nell'integrare il percorso della Knowledge Discovery da Database con l'uso di modelli predittivi.

La tempistica relativa alla disponibilità delle fonti amministrative rappresenta un punto cruciale per il loro impiego a fini censuari. Ad esempio, le fonti delle dichiarazioni fiscali sono disponibili con un ritardo data che va dai 18 ai 20 mesi. Questo elemento potrebbe comportare un forte slittamento in avanti degli output censuari rispetto ai vincoli posti in essere da Eurostat.

La scelta di specifiche fonti amministrative tra tutte quelle disponibili e l'individuazione di un ordine gerarchico degli archivi selezionati rappresenta un elemento imprescindibile: gli archivi di lavoro e di studio si collocano ad un livello superiore rispetto alle altre fonti prese in esame, soprattutto in virtù dell'elevato dettaglio informativo che esse forniscono a livello territoriale e rispetto alla durata dell'attività considerata.

Dalle fonti vengono estratti segnali di presenza sul territorio, che possono efficacemente essere utilizzati per migliorare la qualità dei dati dei registri anagrafici; dal momento che i segnali possono riferirsi anche ad una presenza soltanto occasionale, è necessario un processo di caratterizzazione degli stessi, costruendo variabili derivate che permettano poi di individuare i casi di presenza stabile che meglio corrispondono ai requisiti di *usual residence* previsti dalle norme internazionali.

In riferimento ai segnali diretti forniti dalle fonti amministrative non anagrafiche, la sperimentazione ha mostrato che classificare i segnali in base alla continuità della durata dell'attività in un determinato arco temporale rappresenta un criterio utile per la identificazione di particolari sottopopolazioni relative alla sottocopertura anagrafica. È necessario proseguire in questa direzione: da un lato va migliorato l'algoritmo di classificazione dei segnali, per aumentarne l'accuratezza; dall'altro, i risultati mostrano come i segnali di presenza non continua siano specifici di alcuni settori di attività lavorativa (in particolare, gli stagionali in agricoltura). Pertanto, la classificazione dei segnali deve essere effettuata in base a un criterio più accurato e multidimensionale, in cui il peso del settore lavorativo non provochi distorsioni rispetto alla stima della popolazione dimorante abitualmente.

I segnali "indiretti", forniti dalle relazioni tra individui desumibili dalle fonti amministrative, sono utili per valutare la presenza sul territorio di individui che non mostrano segnali diretti, in quanto non lavorano o studiano e non sono percettori di altro reddito.

Nella sperimentazione effettuata viene utilizzato un *workflow* preliminare per l'uso integrato delle fonti amministrative e dei registri anagrafici ufficiali ai fini del calcolo della popolazione abitualmente dimorante. L'utilizzo di una procedura basata sul *workflow* suddetto con i dati integrati

di fonte amministrativa e archivi anagrafici consente di definire un insieme di individui eleggibili ad essere inclusi nella popolazione abitualmente dimorante in Italia ad una determinata data di riferimento. Questo insieme di possibili dimoranti abitualmente" ammonta a 62,6 milioni di individui nel 2012 (alla data del 31 dicembre) e a 62,4 milioni nel 2013. A partire da questo insieme di individui, è possibile identificare tre sottogruppi principali:

1. La sottopopolazione presente negli archivi anagrafici senza segnali da altre fonti (3,0 milioni di individui nel 2012 e 3,9 milioni nel 2013);
2. La sottopopolazione presente negli archivi anagrafici che presenta anche segnali in altre fonti (58,1 milioni di individui nel 2012 e 57,1 nel 2013);
3. La sottopopolazione non presente in ANVIS ma con segnali da altre fonti (1,5 milioni di individui nel 2012 e 1,4 milioni nel 2013).

Le variabili demografiche, in particolare quelle di genere, età e Paese di cittadinanza nonché la localizzazione sul territorio del segnale sono risultate variabili molto significative per la definizione di profili specifici delle sottopopolazioni. Occorre considerare, però, che una delle criticità è rappresentata da alcuni valori mancanti (soprattutto per il Paese di cittadinanza) proprio per gli individui che non risultano negli archivi anagrafici ma hanno altri segnali diretti dalle fonti amministrative.

Una analisi più approfondita sulle caratteristiche delle sottopopolazioni a rischio di sovra e sottocopertura, effettuata nella sperimentazione con riferimento al 2012, ha permesso di evidenziare alcuni elementi rilevanti e specifici *clusters* di individui. La partizione in sottogruppi e *clusters* è di grande interesse: da un lato permette l'identificazione di sottopopolazioni utili ad ulteriori approfondimenti tematici (ad esempio, tipiche collettività straniere che sfuggono alle anagrafi, ma svolgono attività lavorative specifiche; oppure popolazioni che insistono su determinati territori); dall'altro, questi stessi gruppi possono costituire la base per la definizione di una strategia censuaria che si articoli attraverso l'impiego di tecniche miste che combinano indagini specifiche e opportuni modelli statistici.

Con riferimento alla sovracopertura, è possibile restringere l'insieme di individui sui quali investigare il "rischio" di mancanza di dimora abituale sul territorio nazionale attraverso l'esclusione di specifiche fasce di età per le quali i segnali non sono possibili (ad esempio bambini minori di 6 anni) e l'uso dei segnali indiretti. In riferimento ai 3 milioni di individui che compongono la potenziale sovracopertura tre individui su quattro possiedono la cittadinanza italiana, ma gli stranieri sono mediamente di oltre 6 anni più giovani degli italiani. Questo segmento di popolazione è rappresentato prevalentemente da una popolazione in età attiva (15-64 anni) e che, se si esclude il segmento delle casalinghe, per la sua distribuzione territoriale, particolarmente concentrata nei comuni del Mezzogiorno e in alcune aree centrali del Paese, sia da mettere in relazione alla geografia della disoccupazione e all'esclusione dal mercato del lavoro.

Con riferimento alla sottocopertura, l'analisi mostra la presenza di diversi *clusters* specifici. I segnali di continuità che esprimono stabilità della presenza sul territorio sono relativi a poco più di 400 mila persone, che verosimilmente rappresentano una effettiva sottocopertura del registro anagrafico: si tratta prevalentemente di cittadini stranieri. La localizzazione territoriale e la specifica cittadinanza sono essenziali per identificare i frontalieri, per i quali è ammissibile la mancata registrazione negli archivi anagrafici. L'analisi dei segnali deboli (non continui) ha permesso di evidenziare che in certa parte potrebbero essere comunque associati a individui con una presenza stabile sul territorio, motivo per cui si suggerisce una maggiore caratterizzazione dei segnali diretti.

L'analisi della forza del segnale sulla base della sua continuità temporale rappresenta solo il punto di partenza nell'uso dei dati longitudinali che è possibile elaborare a partire dalle fonti amministrative. Obiettivo prioritario delle prossime sperimentazioni deve essere, pertanto, lo studio di modelli longitudinali, articolati su più anni, per produrre stime di sottopopolazioni maggiormente stabili rispetto alle fluttuazioni legate ai fenomeni specifici, soprattutto quelli legati al mondo del lavoro, da cui derivano i segnali.

L'integrazione di fonti non ancora utilizzate (es. schedari consolari) o in acquisizione nel corso del 2016 (richiedenti asilo, consumi di gas ed elettricità, contratti di acquisto e locazione, certificazione unica), permetterà un incremento dei segnali con corrispondente variazione delle sottopopolazioni ad essi associate, sia in termini di quantità sia di profilo. Alcuni aspetti critici relativi alle

fonti devono però essere maggiormente approfonditi, soprattutto nei termini di ritardo informativo e qualità.

In particolare, con riferimento al ritardo, in alcuni casi è di tipo strutturale e dipendente dalla tempistica del fenomeno amministrativo: è questo il caso di tutte le fonti di dichiarazione fiscale, acquisibili con 12/15 mesi di ritardo rispetto all'anno di riferimento. In altri casi il ritardo non è legato alla tempistica amministrativa, ma a quella di rilascio degli enti: è il caso dei dati del MIUR sugli studenti, disponibile al Ministero a 4/6 mesi di ritardo dall'anno di riferimento ma rilasciato all'ISTAT dopo ben 16 mesi.

Riferimenti bibliografici

Argüeso A., Vega, J.L. (2014). A population census based on registers and a "10% survey" methodological challenges and conclusions. Statistical Journal of the IAOS. 30(1): 35-39.

Bonifazi C., Martini C. (2014). The Impact of the Economic Crisis on Foreigners in the Italian Labour Market. *Journal of Ethnic and Migration Studies*. 40(3).

Cibella N., Gallo G., Pezone A., Tuoto T. (2015). The integration between the 2011 Census Post Eneumeration Survey Data and Administrative Data. The Analysis on Hard-To-Count Population. *Paper presented to the Population Days Conference*. Palermo: 4-6 Febbraio.

Citro, Constance F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*. 40(2): 137-161.

De Angelis S., Mastroluca S., Sasso A. (2015). Neet generation: amount and features of a growing phenomenon. *Paper presented to the 2015 Italian Statistical Conference*. Treviso, 9-11 September.

Crescenzi, F.,Sindoni, G. (2015). The Combined Use of Multiple Data Sources in the Population Census. *Paper preseted to the Unece Group of Experts on population and Housing Censuses*. Geneva: 30 September – 2 October.

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

EMN European Migration Network (2012). Practical responses to irregular migration: the Italian case. Edited by EMN National Contact Point, Idos, Rome, 2012. See: http://ec.europa.eu/dgs/home-affairs/index_en.htm.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy*. Eds. AAAI/MIT Press: Cambridge.

Gallo G., Paluzzi E., Benassi F. (2014). The 2011 Italian experience towards supported-Census for measuring migration. *Paper presented to the Unece Wok Session on Migration Statistics*. Chişinău Republic of Moldova: 10-12 September.

ISTAT, 2014. La misurazione della qualità del 15° Censimento generale della popolazione e delle abitazioni: i risultati dell'indagine di copertura (PES). Seminario del 27 giugno, Roma, http://www.istat.it/it/archivio/126014.

Jensen, P. (1983). Towards a register based statistical system- some Danish experience. *Statistical Journal*. 1(3): 341-365.

Lanzieri, G. (2013). On a New Population Definition for Statistical Purposes. *Paper presented to the Fifteenth Meeting of Group of Experts on Population and Housing Censuses.* Geneva: 30 September – 3 October 2013.

ONS Office for National Statistics (2003). Census Strategic Development Review— Alternatives to a Census. *Review of International Approaches. Information Paper*. London: United Kingdom. Office for National Statistics.

Poulain M., Herm A. (2013). Le register de population centralisé, source de statistiques démographiques en Europe. *Population. 68*(2): 215-247.

Statistics Canada (2011). Preliminary Report on Methodology Options for the 2016 Census. http://www12.statcan.gc.ca/strat/Preliminary%20Report%20on%20Methodology%20Options%20for%20the%202016%20Census.pdf (accessed April 17, 2016).

Statistics Denmark (1995). Statistics on persons in Denmark – a register-based statistical system. Eurostat: Luxembourg.

Statistics Finland (2004). Use of registers and administrative data sources for statistical purposes– best practices in Statistics Finland. Handbook 45. Statistics Finland: Helsinki.

UNECE (2007). Register-based statistics in the Nordic countries: review of best practices with focus on population and social statistics. United Nations Publication, ISBN 978-92-1-116963-8.

UNECE (2011). Using Administrative and Secondary Sources for Official Statistics - A Handbook of Principles and Practices. United Nations Publication.

UNECE (2013). Population Definitions at the 2010 Censuses Round in the Countries of the UNECE Region*. Paper presented to the Fifteenth Meeting of Group of Experts on Population and Housing Censuses*. Geneva: 30 September – 3 October 2013.

Wallgren A., Wallgren B. (2011). To understand the Possibilities of Administrative Data you must change your Statistical Paradigm! *Proceedings of the Survey Research Methods Section. American Statistical Association, Invited Papers.*

Warners A.M., King R., William A.M. and Patterson G. (1999). The well-being of British expatriates retirees in southern Europe. *Ageing and Society*. 19(6): 717-740.

Zhang, Li-Chun (2012). Topics of statistical theory for register-based statistics and data Integration. *Statistica Neerlandica*66, (1), 41-63.