



Nota tecnica sulla produzione dei dati del Censimento Permanente: la popolazione residente per genere, età, cittadinanza e grado di istruzione al 31.12.2021

Dalla rilevazione “porta a porta” al Censimento permanente

Da ottobre 2018 l'Istat ha avviato il Censimento permanente della popolazione e delle abitazioni (che sostituisce il Censimento della popolazione decennale), basato sull'integrazione delle informazioni reperibili dalle fonti amministrative con quelle acquisite dalle indagini campionarie effettuate a rotazione su tutti i comuni italiani.

Nel 2020, a causa del Covid-19, non è stato possibile realizzare le rilevazioni sul campo; tuttavia, l'Istat si è posto comunque l'obiettivo di produrre il conteggio della popolazione residente per genere, età, cittadinanza e grado di istruzione. Attraverso una opportuna metodologia, si sono integrate tutte le informazioni amministrative a disposizione; è stato possibile contare anche su alcune nuove fonti dell'INPS, relative ad alcune misure a sostegno di imprese e lavoratori previste proprio nel 2020 per fronteggiare gli effetti economici della crisi indotta dalla pandemia.

Nel 2021 l'Istat ha potuto di nuovo effettuare le rilevazioni censuarie. La ricchezza informativa proveniente dalle fonti amministrative già studiata ed utilizzata per il conteggio della popolazione del 2020 è stata integrata con i risultati delle indagini (da Lista e Areale) attraverso un modello statistico.

L'utilizzo delle fonti amministrative nella strategia del Censimento permanente era stato immaginato subito dopo il censimento del 2011, quando l'Istat ha iniziato a operare la transizione da un Censimento tradizionale “porta a porta” al Censimento permanente. Nel 2011 infatti, pur nell'ambito di una metodologia ancora sostanzialmente convenzionale (rilevazione sul campo esaustiva e simultanea), ci fu un significativo ricorso all'uso degli archivi per guidare la rilevazione: le Liste Anagrafiche Comunali, per la spedizione postale dei questionari alle famiglie iscritte in anagrafe, e la Lista Integrativa da Fonti Ausiliarie, per rilevare la popolazione abitualmente dimorante non ancora iscritta in anagrafe.

Per il conteggio della popolazione nel 2021 è stato possibile avvalersi, inoltre, dei progressi in termini di qualità e tempestività via via conseguiti dai Registri a supporto della produzione statistica ufficiale. In questo contesto il perno del Censimento permanente della popolazione è costituito dal Registro statistico di Base degli Individui, delle famiglie e delle convivenze (d'ora in avanti RBI) che, insieme ai registri tematici (quali quelli sull'occupazione e sull'istruzione) e alla regolare acquisizione di fonti amministrative ha permesso ormai da quasi un decennio di integrare i dati di fonte anagrafica dei Comuni con le informazioni provenienti da altri archivi di INPS, MIUR, Catasto Immobiliare, Casellario dei pensionati, ecc., così come previsto dalla Legge 205/2017.

Il RBI è un ambiente informativo interno all'Istituto a supporto dei processi produttivi statistici; in particolare, è l'infrastruttura di base per la produzione delle statistiche ufficiali sulla popolazione e il riferimento per l'estrazione dei campioni delle indagini previste per il Censimento permanente (cfr. legge censuaria e Piano Generale di Censimento) e di tutte le indagini campionarie sulle famiglie.

In accordo con l'impianto metodologico del Censimento permanente, e nel rispetto delle norme sulla protezione dei dati personali, il RBI e tutte le fonti amministrative non anagrafiche disponibili in Istat contengono dati anonimi “statistici”, ossia derivanti da un processo di elaborazione e validazione statistica e sono riferiti a un numero limitato di variabili funzionali alla rappresentazione delle principali caratteristiche strutturali della popolazione.



Il RBI è “consolidato” con cadenza annuale con riferimento al 31 dicembre di ciascun anno. Questo consolidamento è realizzato mediante l’integrazione dei flussi individuali anagrafici della dinamica demografica (nascite, decessi, trasferimenti di residenza da e per altro comune o da e per l’estero). L’applicazione del modello di contabilità demografica MIDEA (Micro-DEmographic Accounting) consente di sfruttare le potenzialità della base dati micro (flussi+stock) per produrre indicatori più accurati e innovativi sulla dinamica demografica, tenendo conto della sequenza degli eventi demografici sperimentati dagli individui.

I principali output del sistema MIDEA applicati al RBI “consolidato” sono il bilancio demografico della popolazione residente comunale e la struttura della popolazione residente comunale per genere, età, stato civile e cittadinanza. La produzione tempestiva di questi output anticipa i risultati definitivi annuali del Censimento permanente, con l’obiettivo di assicurare la continuità della produzione statistica ufficiale, soddisfacendo al tempo stesso le tempistiche stringenti richieste tanto dagli utilizzatori istituzionali dei dati di popolazione, quanto dal regolamento europeo sulle statistiche demografiche.

Questi stessi output vengono rielaborati e diffusi successivamente come definitivi a seguito della correzione del RBI mediante gli esiti del Censimento permanente.

Il conteggio del Censimento permanente è finalizzato a correggere gli errori di sovra e sotto-copertura del RBI, individuando le persone presenti nel Registro come residenti ma non trovate negli archivi amministrativi (sovra copertura) e quelle che invece sono state trovate nei dati amministrativi come abitualmente dimoranti ma non risultano tali nel Registro (sotto copertura). Questa correzione, anche nel 2021, così come è avvenuta per il 2020, agisce a livello micro, opera infatti attraverso la riclassificazione dei record individuali contenuti nel Registro riconoscendoli come abitualmente dimoranti o meno, sulla base dei segnali di vita amministrativi. Si consolida quindi una innovazione metodologica di rilievo che assicura la corrispondenza in termini di “teste” tra il dato di conteggio e i record di individui abitualmente dimoranti contenuti nel RBI, a differenza di quanto fatto nel 2018 e nel 2019 in cui l’obiettivo della correzione del conteggio veniva conseguito a livello macro, applicando dei pesi (correttori per sovra e sotto copertura) agli individui residenti presenti nel RBI.

Questo impianto garantisce, e lo farà in misura sempre maggiore nei prossimi anni, una completa integrazione con gli output delle statistiche demografiche, contribuendo in tal modo a una maggiore stabilità e coerenza di tutto il sistema dei dati che riguardano la popolazione. In tal senso, la metodologia riproposta nel 2021, già introdotta nel 2020 a causa del Covid-19, rappresenta un investimento per il ciclo censuario post-2021, che vedrà un sempre maggiore utilizzo delle informazioni provenienti dagli archivi amministrativi per stimare il conteggio di popolazione e, laddove possibile, alcune variabili tematiche previste dal Regolamento europeo sui Censimenti della popolazione e delle abitazioni.

I segnali di vita e le fonti utilizzate per il conteggio 2021 della popolazione censita

Ai fini del conteggio della popolazione del 2021 è stato aggiornato l’Archivio Integrato dei Dimoranti Abitualmente in Italia (d’ora in avanti AIDA). Questa base dati raccoglie le informazioni derivanti dalle fonti amministrative non anagrafiche organizzate nel Sistema Integrato di Microdati (SIM). La base dati del SIM costituisce di fatto un *repository* di dati amministrativi che, garantendo l’anonimato degli individui, è stato costruito con lo scopo di sostenere i processi di produzione statistica, sia per le statistiche sociali che per le statistiche economiche. L’assegnazione di un codice univoco e costante permette di identificare ciascun individuo e unità economica all’interno dei diversi archivi e di costruire le relazioni tra le diverse fonti, garantendo al contempo un trattamento di dati privi di identificativi diretti, rispettoso quindi delle norme in materia di riservatezza e protezione dei dati.

Attraverso le indicazioni fornite dagli esperti ricercatori tematici e metodologi di fonti amministrative dell’Istat che usano gli archivi per la produzione di statistiche sulla popolazione, il Servizio Censimento permanente della popolazione ha effettuato una selezione delle fonti da prendere in considerazione per il conteggio della popolazione dimorante abitualmente e, attraverso un gruppo di esperti tematici e metodologi, ha stabilito un ordine gerarchico delle fonti nel processo di integrazione di AIDA. Ciò ha permesso poi di osservare i “segnali di vita” degli individui in termini di “dimora abituale in Italia”.



I ricercatori tematici del Servizio del Censimento permanente della popolazione sono arrivati a una definizione dei segnali di vita (più noti nella letteratura internazionale come “*Signs of administrative life*”) adottata anche nel Piano Generale di Censimento del 2022.

I “segnali di vita amministrativi” si riferiscono ad attività svolte dagli individui desumibili dagli archivi amministrativi. Queste attività permettono di identificare chiaramente un periodo di tempo durevole (ad esempio, un anno) e un luogo (un Comune) in cui si realizzano. Svolgere un lavoro autonomo o lavorare per un’impresa, essere un dipendente pubblico, avere un regolare contratto d’affitto annuale per una abitazione, frequentare una scuola o l’università sono esempi di segnali di vita amministrativi diretti. Invece, si definiscono segnali di vita indiretti quelle situazioni, sempre desumibili dagli archivi amministrativi, che identificano uno status o una condizione, ad esempio essere percettori di reddito di cittadinanza o di una pensione di vecchiaia, oppure essere familiari a carico per i quali il dichiarante del reddito indica di avere a suo carico il coniuge, i figli o altro parente.

Il database AIDA ha integrato, per il conteggio della popolazione 2021, le informazioni anagrafiche individuali di RBI con quelle del Registro Tematico del Lavoro (RTL) dell’Istat, degli archivi dell’istruzione, delle dichiarazioni fiscali e degli archivi della previdenza sociale, nonché del catasto immobiliare dell’Agenzia delle Entrate con un riferimento temporale che va dal primo gennaio 2020 al 31 dicembre 2021.

Nell’ambito delle fonti utilizzate, il Prospetto 1 mostra, per i due anni (2020 e 2021), la denominazione dell’archivio, l’ente titolare fornitore, il numero di record trattati per ciascun anno e la classificazione per tipo di segnale di vita che, secondo la logica gerarchica assegnata dagli esperti del Censimento permanente della popolazione, consente di classificare gli individui ai fini della dimora abituale in Italia.

Complessivamente, i record trattati in AIDA per ciascun anno sono diverse centinaia di milioni considerando, però, che ciascun individuo può dare luogo in tempi diversi a più di una occorrenza nello stesso archivio, e che lo stesso individuo può apparire allo stesso istante in diversi archivi.

Nel costruire la base dati integrata di AIDA si è tenuto conto della periodicità dei dati e della rilevanza rispetto alla definizione di dimora abituale secondo quanto è previsto dal regolamento europeo (regolamento CE n. 1260/2013). Dal Prospetto 1 è possibile osservare che le fonti del MIUR sull’istruzione, le fonti assicurative dell’INPS relative agli individui che partecipano al mercato del lavoro sono in gran parte disponibili sia per il 2020 che per il 2021. Invece, relative al Casellario dei pensionati e assistenziali, quali il reddito di cittadinanza, hanno un riferimento data che copre solo il 2020; inoltre, ci sono alcune fonti dell’INPS relative ai sostegni sociali e assistenziali, quali il reddito di emergenza, i benefici per le indennità Covid-19 e i bonus *baby-sitting*, che sono segnali di vita messi a disposizione dell’Istat per la prima volta per l’anno 2020 e solo alcune (indennità Covid-19) sono confermate anche per il 2021.



PROSPETTO 1 - ARCHIVI AMMINISTRATIVI E REGISTRI TEMATICI CHE FORNISCONO SEGNALI DI VITA DIRETTI E INDIRETTI RISPETTO ALLA DIMORA ABITUALE IN ITALIA AL 31.12.2021

SEGNALI DI VITA PER ORDINE GERARCHICO	ENTE TITOLARE	DENOMINAZIONE ARCHIVIO	N. record	
			2020	2021
1) Segnali diretti di lavoro	INPS	UNIEMens	169.936.659	175.659.666
		DMAG (dichiarazione sulla manodopera agricola)	8.192.097	8.187.281
		Lavoratori autonomi in agricoltura	444.425	441.854
		Archivio dei parasubordinati collaboratori	6.928.020	7.638.878
		Rapporti di lavoro domestico	1.190.744	1.249.355
		GDP (ex INPDAP) - Posizioni degli assicurati iscritti	40.020.267	
		Lavoro occasionale (Libretto Famiglia e Prestazioni Occasionali)	12.919.000	2.423.979
	MIUR	Archivio del personale universitario	357.510	
		Archivio del personale delle scuole statali	1.144.571	
	INAIL	Lavoratori interinali	1.252.876	1.545.488
MEF	NOIPA - Anagrafica (Cedolini Stipendiali dipendenti PA)	21.719.865		
ISTAT	Registro Tematico del Lavoro (RTL)	26.108.524	26.720.662	
2) Segnali diretti di studio	MIUR	Anagrafe degli studenti delle scuole	17.217.058	17.059.016
		Archivio degli iscritti e delle iscrizioni universitarie	1.859.479	1.873.321
		Archivio delle lauree e dei laureati	459.941	436.802
		Anagrafe Nazionale Studenti dei corsi Post laurea (chiusure)	78.255	65.974
		Anagrafe Nazionale Studenti dei corsi Post laurea (iscrizioni)	162.573	132.234
3) Segnali diretti di assistenza sociale e contratti di affitto ad uso abitativo	INPS	Beneficiari CIG a pagamento diretto	19.042.557	8.302.442
		Trattamenti non pensionistici	10.295.000	
		Casellario dei Pensionati e dei trattamenti pensionistici	22.717.034	
		Possessori della Social Card	425.730	
		Reddito d'inclusione (REI)	30.936	
		Reddito di cittadinanza (RDC)	1.575.129	
		Indennità Covid-19 (Bonus 2400 euro Decreto sostegni 2021)		1.289.806
		Indennità Covid-19 per Lavoratori Domestici	275.075	
		Bonus baby sitting e centri estivi (emergenza Covid-19)	1.304.903	
	Indennità 600/1000 euro (emergenza Covid-19)	8.835.765		
MEF	Reddito di emergenza (REM)	628.407		
4) Segnali indiretti	MEF	Banca dati statistica reddituale	41.180.056	
		Agenzia delle Entrate - Contratti di Locazione	4.113.422	4.477.448
		Agenzia delle Entrate - Dichiarazioni Modello Unico Persone	14.646.368	
		Agenzia delle Entrate Dichiarazioni 730	36.289.318	
		Agenzia delle Entrate - Certificazione Unica	50.927.175	
		Agenzia delle Entrate - Catasto delle Unità Immobiliari	111.042.174	
	Agenzia delle Entrate - Catasto delle Unità Immobiliari –	863.641		
ACI	Agenzia delle Entrate - Catasto delle Unità Immobiliari – Trento	1.512.268		
Altre fonti utilizzate	ACI	ACI - Parco veicoli circolanti	50.553.845	50.872.976
	ISTAT	Registro Base degli INDIVIDUI (RBI)		100.985.564
	MIN. INTERNO	Permessi di soggiorno	3.480.379	3.720.669
		Acquisizione e reiezione della cittadinanza italiana	54.069	
MAE	MAE - Ministero Affari Esteri - Archivio Italiani all'Estero	8.998.460	9.332.437	
MEF	Agenzia delle Entrate - Anagrafe delle Persone Fisiche (AT)	101.518.237	103.004.529	

Fonte: Istat 2022



Ritardi rispetto alla data di riferimento si registrano per alcune fonti delle dichiarazioni fiscali, in particolare quelle della Banca dati reddituale e le informazioni dei modelli UNICO/730, per i quali si osservano tempi di utilizzazione delle informazioni con un ritardo data di circa 20 mesi. Ciò è dovuto alla natura delle fonti fiscali, che solitamente si riferiscono alle dichiarazioni dei redditi dell'anno precedente a quello della data di riferimento.

Nel corso degli ultimi anni il sistema integrato di AIDA ha conosciuto una notevole evoluzione non solo in riferimento all'impiego di registri tematici, quali RBI e RTL, ma anche dal punto di vista tecnico e informatico consentendo di sfruttare al meglio, e con tempi di lavorazione sempre più rapidi, le informazioni disponibili. Altrettanto significative sono state le attività che hanno portato ad una ingegnerizzazione di processo e di prodotto e ad aggiornare i codici di provincia e comune migliorando la qualità di localizzazione dei segnali delle fonti amministrative.

Dalla definizione dei “segnali di vita” è possibile evincere anche una classificazione gerarchica dei segnali: i segnali di lavoro e studio, insieme ai contratti di affitto, alle dichiarazioni della Banca dati reddituale¹ sono classificabili come segnali di vita diretti rispetto alla dimora abituale in Italia.

Gli archivi dell'INPS relativi al lavoro e quelli del MIUR sulla frequenza di corsi scolastici e universitari sono particolarmente rilevanti ai fini della individuazione della popolazione abitualmente dimorante in Italia. Il Prospetto 2 mostra che queste fonti offrono un notevole dettaglio informativo, legato non solo alla durata dell'attività, ma anche alla localizzazione della stessa (Comune e indirizzo) e ad alcuni attributi specifici dell'attività svolta (tipologia di contratto di lavoro, corso di studi, ecc.) che forniscono elementi rilevanti nel valutare la forza del segnale di vita sul territorio.

PROSPETTO 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 (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

Fonte: Istat, 2022

Invece, tutte le fonti fiscali (modelli UNICO, modelli 730, ecc.), il possesso di automobili presenti nell'archivio ACI e il possesso di immobili presenti nell'archivio catastale forniscono segnali indiretti rispetto alla dimora abituale. Infatti, a partire dal “Quadro Familiari a carico” dei modelli UNICO Persone Fisiche e 730, si possono individuare le relazioni principali tra “coniugi” e tra il genitore dichiarante e “figlio/i”. Nella base dati dei modelli fiscali sono contenuti, inoltre, anche i dati relativi agli “altri familiari” a carico (il coniuge legalmente ed effettivamente separato; i discendenti dei figli; i genitori, i generi e le nuore; il suocero e la suocera; i fratelli e le sorelle; 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. Dato che questi segnali di presenza in Italia sono desunti “indirettamente” dalla dichiarazione di un percettore di reddito, si è deciso di classificarli come segnali indiretti; così anche il possesso di automobile o di un'unità immobiliare non è considerato un segnale diretto e, in quanto tale, viene associato ai segnali indiretti.

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

¹ Si tratta delle dichiarazioni contenute nella Banca dati reddituale. Attraverso apposite campagne ordinarie, l'INPS richiede ai soggetti beneficiari di prestazioni previdenziali e/o assistenziali collegate al reddito (pensionati, disoccupati beneficiari di cassa integrazione, ecc.) l'annuale dichiarazione dei redditi (modello RED) che incidono sulle prestazioni in godimento.



La base dati della popolazione residente con cui vengono effettuati i confronti in termini di coerenza con i segnali di vita in Italia o meno è costituita dagli individui contrassegnati come residenti nel RBI. Il sistema MIDEA (Micro-DEmographic Accounting)/ANVIS (Anagrafe Virtuale Statistica) è il processo statistico che porta con cadenza annuale alla identificazione della popolazione presente nel RBI eleggibile ad essere considerata residente. La base dati ANVIS viene quindi consolidata ogni anno mediante l'aggiornamento a livello micro con i flussi della dinamica demografica (flussi di iscrizione per nascita; cancellazione per decesso; iscrizione/cancellazione da/per l'estero; iscrizione/cancellazione da/per altri Comuni italiani), verificatisi nei periodi dal 9 ottobre 2011 al 31 dicembre di ciascun anno.

Le informazioni anagrafiche 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;
- ✓ Genere;
- ✓ Cittadinanza;
- ✓ Paese di nascita.

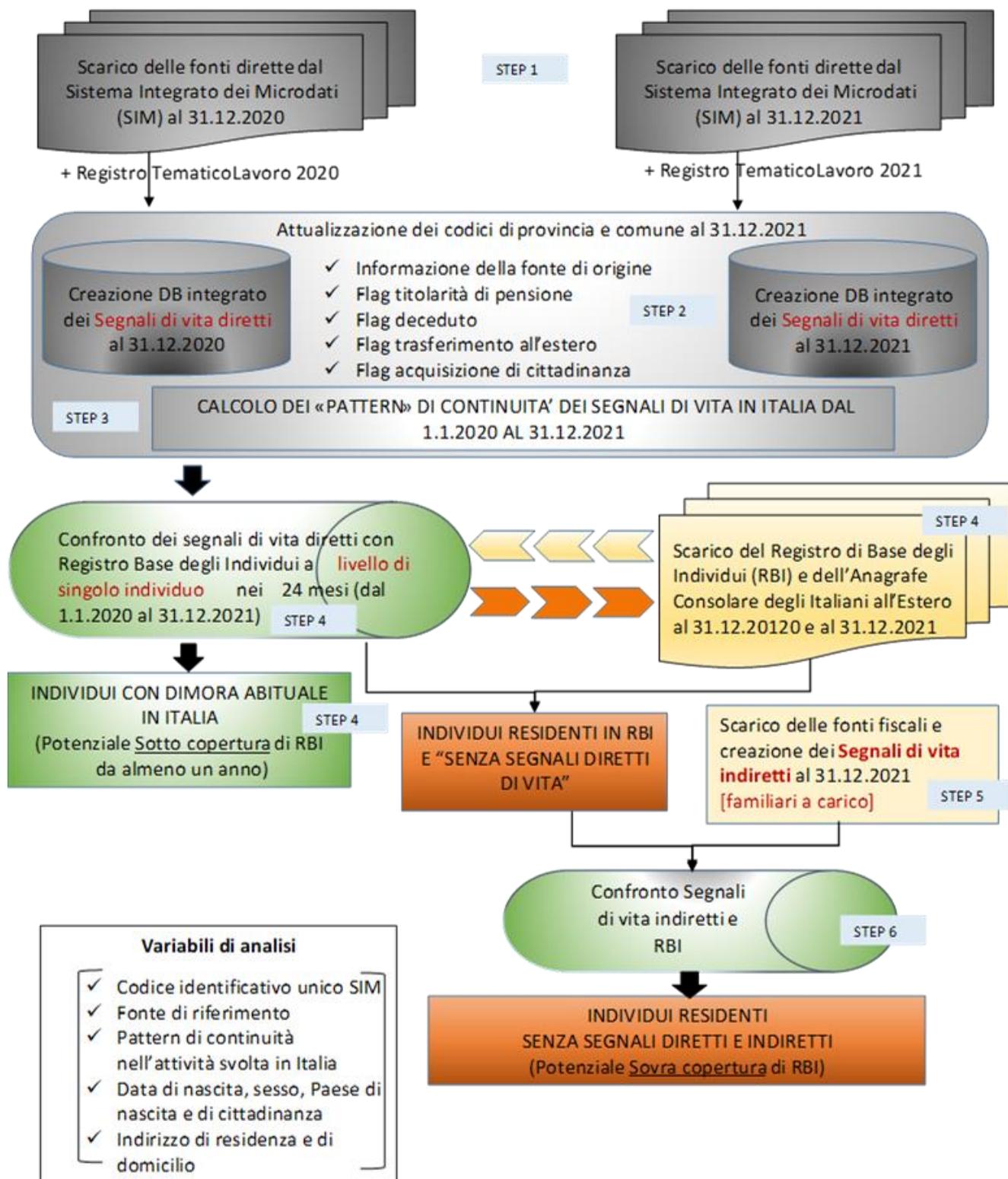
Qualora la Cittadinanza non fosse presente in uno degli archivi considerati (RBI e Permessi di Soggiorno) si è fatto riferimento all'informazione sul Paese di nascita. In presenza di informazioni anagrafiche incongruenti nelle diverse fonti per uno stesso individuo, sono state scelte le informazioni sulla base di una scala gerarchica delle fonti. La fonte dell'Anagrafe Tributaria, insieme a RBI, rappresenta la fonte più completa rispetto alle variabili anagrafiche relative a data di nascita, genere e paese di nascita. Si evidenzia, inoltre, che ai fini dell'analisi della effettiva localizzazione degli individui in termini di dimora abituale, risulta molto utile la presenza in questa fonte anche della variabile "Comune" del domicilio fiscale.

Il flusso di lavoro di AIDA e i profili di continuità per la dimora abituale in Italia

Il Prospetto 3 mostra il flusso di lavoro, e i relativi passi, con cui è stato seguito il processo di integrazione delle fonti amministrative con RBI, dallo scarico delle fonti dal SIM fino alla determinazione della sotto e sovra copertura di RBI. Successivamente allo scarico dal SIM (*Step 1*), i dati derivanti dalle fonti amministrative prescelte sono stati strutturati in maniera adeguata all'analisi della popolazione dimorante abitualmente in Italia.



PROSPETTO 3. FLUSSO DI LAVORO DEI SEGNALI DI VITA PER LA DETERMINAZIONE DEL CONTEGGIO DI POPOLAZIONE DEL CENSIMENTO PERMANENTE DELLA POPOLAZIONE AL 31.12.2021



Fonte: Istat, 2022



In primo luogo, il processo di produzione di AIDA comporta il trattamento di dati provenienti da oltre quaranta archivi amministrativi, ognuno dei quali relativo a più anni, contenenti le informazioni di base sui segnali di vita degli individui. In ciascun archivio sono presenti, per ogni evento amministrativo registrato, le informazioni sulla localizzazione dell'evento stesso, tramite i codici di provincia e comune. Le stesse variabili sono poi utilizzate anche per ulteriori informazioni, quali la residenza, il luogo di nascita o il comune di provenienza. In generale queste informazioni presentano una serie di problematiche:

- ✓ valori validi ma disallineati temporalmente, o comunque non aggiornati alla data di riferimento;
- ✓ valori totalmente o parzialmente mancanti.

Il primo caso è imputabile all'acquisizione degli archivi in momenti diversi e quindi, a seguito delle variazioni amministrative intercorse, uno stesso comune può presentare in differenti annualità codici territoriali diversi. Il fatto poi che il calcolo degli output di AIDA richieda, come vedremo, l'utilizzo di dati relativi a due anni consecutivi rende il processo di produzione ancora più sensibile a questo tipo di problematica.

Il caso dei valori mancanti risulta invece essere rilevante in particolare per alcune fonti che forniscono segnali diretti di presenza sul territorio (ad es. le posizioni degli assicurati iscritti ex INPDAP), oppure l'informazione è disponibile solo in maniera parziale, con una limitazione all'utilizzabilità delle fonti stesse.

Ciò comporta che il Servizio Censimento effettui alcune analisi e delle successive operazioni, quali ad esempio:

- ✓ l'analisi di tipo territoriale, anche relativamente a RBI e RTL;
- ✓ la definizione di una localizzazione prevalente per i segnali di lavoro e studio;
- ✓ aggregazioni di tipo longitudinale basate su variabili geografiche;
- ✓ confronti temporali su diversi archivi;
- ✓ comparabilità con l'informazione proveniente da altre fonti.

Per gestire queste criticità è stata definita e sviluppata una procedura per l'imputazione e l'allineamento temporale delle variabili di localizzazione, facendo uso sia di informazioni presenti negli stessi archivi, sia di informazioni ausiliarie provenienti da fonti esterne, che è stata introdotta nel processo di produzione di AIDA (Step 2). Nella produzione di AIDA, illustrata nel Prospetto 3, sono stati quindi imputati ed aggiornati i codici di provincia e comune di tutte le fonti amministrative adoperate in input dal processo, rendendo così consistenti le informazioni rispetto ad un unico istante di riferimento, che è stato fissato al 31 dicembre 2021. L'imputazione dei valori è avvenuta sfruttando, quando possibile, vincoli per la determinazione univoca dei valori.

Successivamente all'aggiornamento dei codici di provincia e comune, sempre nello Step 2 del processo di AIDA, si procede alla costruzione, sia per il 2020 che per il 2021, della base dati integrata delle fonti di lavoro e studio dell'INPS e del MIUR nonché delle altre fonti con segnali di vita diretti e indiretti (Casellario dei pensionati, Trattamenti non pensionistici, Banca dati reddituale, Contratti di locazione, Cassa Integrazione Guadagni, reddito di cittadinanza, di emergenza e sussidi e indennità Covid-19). Inoltre, si producono anche i *flag* informativi delle fonti originarie relativi all'eventuale decesso, alla titolarità di una pensione, al trasferimento all'estero (desumibile da Banca dati Reddituale e dal Casellario dei Pensionati) e all'acquisizione di cittadinanza italiana.

Un aspetto rilevante per il conteggio della popolazione di AIDA è rappresentato dall'aver tenuto conto della definizione di dimora abituale del Regolamento 1260/2013 dell'Unione Europea. Ai fini della "*usually resident population*", il Regolamento europeo specifica come riferimento temporale la permanenza in un luogo per almeno 12 mesi prima della data di riferimento e, nel caso di un periodo di permanenza inferiore ai 12 mesi l'"intenzione" di vivere stabilmente nello stesso territorio. Sulla base di queste indicazioni e con l'obiettivo di approssimare al meglio l'"intenzione" a stare in Italia, si è scelto di identificare i segnali di vita amministrativi in un periodo più ampio (24 mesi) che va dal primo gennaio 2020 al 31 dicembre 2021 che, per il conteggio di popolazione censita, rappresenta la data di riferimento del Censimento permanente della popolazione 2021.

L'osservazione longitudinale dei segnali diretti su due anni consente di cogliere specifici profili di presenza degli individui sul territorio (o pattern di continuità nella presenza, attività dello Step 3 riportata nel Prospetto 3). Questi profili della presenza in Italia in alcuni casi consentono di identificare chiaramente la dimora abituale in Italia, mentre in altri i "segnali di vita" nel nostro Paese sono di bassa intensità, oppure il segnale che si delinea dal 1° gennaio 2020 al 31 dicembre 2021 è quello dei lavoratori stagionali, ovvero di profili che non possono essere associati alla popolazione dimorante abitualmente.



A questo riguardo, per dare una idea precisa su come vengono classificati i segnali di lavoro e studio rispetto alla dimora abituale in Italia, il Prospetto 4 può risultare un utile schema di rappresentazione dell'utilizzo dei segnali diretti di presenza in Italia. Ad esempio, i profili in blu scuro da 1 a 4 totalizzano almeno 12 mesi di segnali di presenza dal 2020 al 2021, anche se i profili non sono sempre continui per tutto il periodo di osservazione, come nel caso 3 e 4 dello schema. Invece, i segnali dal numero 5 in poi non solo mostrano profili di discontinuità ma complessivamente non totalizzano, nel corso dei 24 mesi di osservazione, "almeno" 12 mesi di segnali diretti di studio e/o di lavoro in Italia. Questi segnali si configurano con una debole intensità della presenza (profili 5 e 6), oppure in alcuni casi come segnali non utili rispetto alla determinazione della dimora abituale in Italia (profili da 7 a 11). Anche se i profili 10 e 11 potrebbero condurre, nella successiva edizione del Censimento permanente, alla individuazione di segnali forti di presenza (più o meno equivalenti ai casi 2 e 4)

PROSPETTO 4 - CONFIGURAZIONE E CLASSIFICAZIONE DEI SEGNALE DIRETTI DI STUDIO E DI LAVORO SECONDO I PROFILI DI CONTINUITÀ DELLA DIMORA ABITUALE IN ITALIA DAL 1.1.2020 AL 31.12.2021

Finestra temporale di osservazione (24 mesi)																								Profili del segnale di lavoro e studio	
G	F	M	A	M	G	L	A	S	O	N	D	G	F	M	A	M	G	L	A	S	O	N	D		
																								1	Continuo nel 2020-2021
																								2	Continuo, prevalentemente nel 2021
																								3	Continuo, oltre dicembre 2020
																								4	Con pause ma almeno un anno
																								5	Stagionale
																								6	Interrotto sia nel 2020 che nel 2021
																								7	Solo a dicembre 2020
																								8	Solo nei primi mesi del 2020
																								9	Interrotto e solo nel 2020
																								10	Solo negli ultimi mesi del 2021
																								11	Interrotto e solo nel 2021

Fonte: Istat, 2022

Ogni segnale è da associare a uno specifico individuo e a una determinata localizzazione territoriale, secondo quanto è mostrato nel Prospetto 5: 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 e di studio. Pertanto, la struttura informativa dell'algoritmo che elabora il segnale diretto di lavoro o studio di AIDA si configura con una stringa che sintetizza, in rosso nella terza colonna, un segnale di lavoro nella prima posizione e un segnale di studio nella posizione 9, mentre la classificazione del profilo di continuità è riportato, sempre con una stringa, nell'ultima colonna del prospetto, dove per il caso specifico, il segnale diretto di presenza nella fonte di lavoro è presente per tutti i mesi.



PROSPETTO 5. ESEMPIO DI UTILIZZO DEI SEGNALI DI VITA DIRETTI DI LAVORO E STUDIO LOCALIZZATI IN UN COMUNE

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}
<i>Esempio</i>	<i>Individuo 18 di Agliè</i>	<i>UniEmens (pos.1)+Università (pos.9)</i>	<i>Tempo indeterminato</i>	<i>Presente tutti i mesi</i>
Dati di esempio	001-001-0000018	1000000010	0---1----	1111111111111111111111

Fonte: Istat, 2022

Riprendendo il flusso di lavoro di AIDA del Prospetto 3, dopo che si è proceduto con la riduzione dei segnali di lavoro e studio, e degli altri segnali diretti sui singoli individui ed è stato determinato il Comune prevalente dove viene svolta l'attività di studio o di lavoro, si prosegue con l'integrazione degli individui con segnali di vita diretti con il Registro di popolazione RBI. In particolare, nello *Step 4* si effettua lo scarico di RBI al 31.12.2020 e al 31.12.2021 e, per singolo individuo, si realizza il confronto con i segnali di vita diretti che consente di determinare, per ciascun Comune, tutti gli individui con dimora abituale in Italia non residenti nel RBI e gli individui residenti nel RBI senza segnali di vita diretti. Il passo successivo (*Step 5*) consiste nello scarico delle fonti fiscali e nel successivo confronto dei segnali di vita indiretti dei "familiari a carico" e dei titolari di una automobile e di un'unità immobiliare con gli individui residenti nel RBI che erano risultati senza segnali di vita diretti dello *Step 4*.

Infine, l'ultimo passaggio consiste quindi nel determinare, nello *Step 6*, la potenziale sovra copertura anagrafica di RBI che è composta dagli individui privi di segnali diretti e indiretti da tutte le fonti considerate che risultano residenti nel RBI. Questo contingente di popolazione in quanto privo di segnali diretti e indiretti delle fonti prese in esame potrebbe essere classificato nell'insieme delle persone individuate come sovra copertura di RBI. Su questo contingente di popolazione si procede, però, con la lavorazione di una ulteriore fase dove viene utilizzata l'informazione sulle famiglie anagrafiche identificate nelle LAC, in modo da riconoscere le persone registrate come "coniugi", all'interno della stessa famiglia, di intestatari della famiglia che hanno segnali di lavoro e studio o altri segnali di vita diretti.

Per il 2021, a differenza del 2020, la disponibilità sia dei dati d'indagine che dei segnali di vita aggiornati (in modo da coprire gli stessi riferimenti temporali) ha permesso di adattare un modello statistico (modello a classi latenti) per migliorare il processo di conteggio della popolazione utilizzando sia il dato amministrativo che i risultati delle indagini censuarie condotte sul campo ad ottobre del 2021.

Ciò ha indotto a non limitare l'attenzione dei segnali di vita solo ed esclusivamente alle fonti amministrative disponibili, ma di valorizzare la ricchezza degli archivi amministrativi, utilizzando contemporaneamente i dati delle indagini censuarie. I ricercatori dell'Istat hanno operato secondo un approccio di tipo *Knowledge Discovery from Databases*; si sono avvalsi, in definitiva, di un processo strutturato e iterativo, in cui una parte delle variabili da analizzare sono costruite in itinere, man mano che si procede con la lavorazione dei segnali amministrativi. Questo tipo di approccio consente, all'interno di una grande mole di dati amministrativi e di indagine, di identificare criteri e profili utili per il trattamento di specifiche sottopopolazioni.

Validazione statistica dei profili di sovra copertura e di sotto copertura del RBI

Un ulteriore passaggio, prima del rilascio definitivo dei dati, ha riguardato la definizione di misure quantitative finalizzate alla misurazione della validità e distorsione dei criteri utilizzati per l'individuazione della sovra e sotto copertura di RBI e all'eventuale segnalazione statistica di nuovi criteri non individuati dagli esperti tematici.

A tale scopo è stato possibile avvalersi delle indagini campionarie progettate per il Censimento Permanente. Il disegno di rilevazione per il Censimento Permanente si articola in due distinte indagini campionarie, una Areale e l'altra da Lista. Entrambe le indagini sono state progettate per rilevare sul campo i dati necessari ad integrare le informazioni mancanti nei registri tematici, in modo da assicurare la completezza e la qualità dei dati censuari.



In teoria, gli esiti di tali indagini potrebbero essere utilizzati come fonte statistica esterna per stimare i tassi di presenza sul territorio degli individui in AIDA per specifici profili di segnali di vita amministrativi. Il confronto tra la stima dei tassi di presenza degli individui sul territorio per profilo e i criteri di sovra e sotto-copertura individuati dagli esperti tematici può essere considerato come un indicatore di validità dei criteri scelti. In pratica, ciò è possibile solo se gli esiti di indagine possono essere considerati una fonte statistica priva di errori, assunzione non valida a causa della mancata risposta totale (MRT) nelle indagini. La mancata osservazione di alcune unità statistiche sul territorio avviene usualmente in tutti quei casi in cui non è stato possibile ottenere alcuna informazione utile relativa alla reperibilità dell'individuo sul territorio e quindi si traduce come un errore di misurazione della variabile target che è la presenza dell'individuo sul territorio.

Per ovviare alle possibili distorsioni nelle stime censuarie si è ritenuto opportuno adottare una strategia di stima basata su modelli a classi latenti, integrando i dati indagine con i segnali amministrativi presenti in AIDA, al fine di definire la probabilità di presenza di un individuo sul territorio. I modelli a classi latenti, utilizzati per la stima dell'errore di misura, si basano sulla disponibilità di misure multiple della stessa variabile categoriale. Tutte le variabili osservate possono essere potenzialmente affette da errori di classificazione, mentre la vera variabile target non è direttamente misurabile ma è una variabile latente. L'associazione tra le variabili osservate è spiegabile dall'esistenza nella popolazione di gruppi (classi latenti) e la distribuzione delle variabili osservate varia tra i gruppi, mentre dentro ciascun gruppo le variabili osservate sono indipendenti. Tali assunzioni permettono di definire la distribuzione congiunta delle variabili osservate, condizionatamente alla classe latente di appartenenza, e di conseguenza di predire per ciascun individuo della popolazione l'appartenenza ad una classe latente date tutte le misure osservate.

Nello specifico gli esiti delle indagini statistiche e i principali segnali di vita presenti in AIDA sono stati trattati come variabili categoriali con due modalità (esistenza/assenza del segnale di presenza), tutte affette da errore di classificazione. È stato quindi stimato un modello con due classi latenti interpretate come la vera presenza/assenza dell'individuo sul territorio.

Per tutte le combinazioni di segnali amministrativi e di indagine, la presenza in RBI e diverse caratteristiche individuali e familiari è stata stimata la probabilità di appartenenza ad entrambe le classi latenti. Un'alta probabilità di appartenere alla classe latente che descrive l'assenza sul territorio permette di individuare i profili probabilmente sovra-coperti in RBI, mentre un'alta probabilità di appartenere alla classe latente che descrive la presenza sul territorio, in concomitanza con l'assenza del segnale in RBI, permette di individuare i profili probabilmente sotto-coperti in RBI. Il confronto tra le probabilità stimate e i criteri deterministici ha permesso sia la conferma sia l'individuazione di nuovi criteri deterministici, definendo il set finale di regole per la valutazione della sovra e sotto copertura di RBI.

La definizione finale degli aggregati di sovra e sotto copertura di RBI al 31.12.2021

Nei paragrafi precedenti si è visto che il processo di AIDA è stato utilizzato per il conteggio del 2021 al fine di identificare due sottogruppi di popolazione che consentono di correggere il RBI al 31.12.2021:

- 1) gli individui dimoranti abitualmente in Italia perché mostrano “segnali di vita diretti” di almeno un anno dagli archivi amministrativi; questi individui, dato che non sono presenti nel RBI come residenti al 31.12.2021, rappresentano la sotto copertura delle anagrafi comunali alla stessa data;
- 2) gli individui residenti nel RBI al 31.12.2021 senza “segnali di vita diretti e indiretti” negli archivi amministrativi, che rappresentano la sovra copertura delle anagrafi comunali.

L'impianto sviluppato e già utilizzato per il conteggio del 2020 si caratterizza per l'uso degli archivi amministrativi in modo “indipendente” rispetto alle informazioni anagrafiche comunali. Questo approccio richiama molto da vicino il carattere di “universalità” del Censimento tradizionale della popolazione: tutti i Comuni sono osservati nello stesso istante di tempo e sono sottoposti a tecniche standardizzate sia in riferimento alla metodologia utilizzata (i profili di continuità) che al trattamento dei dati individuali.

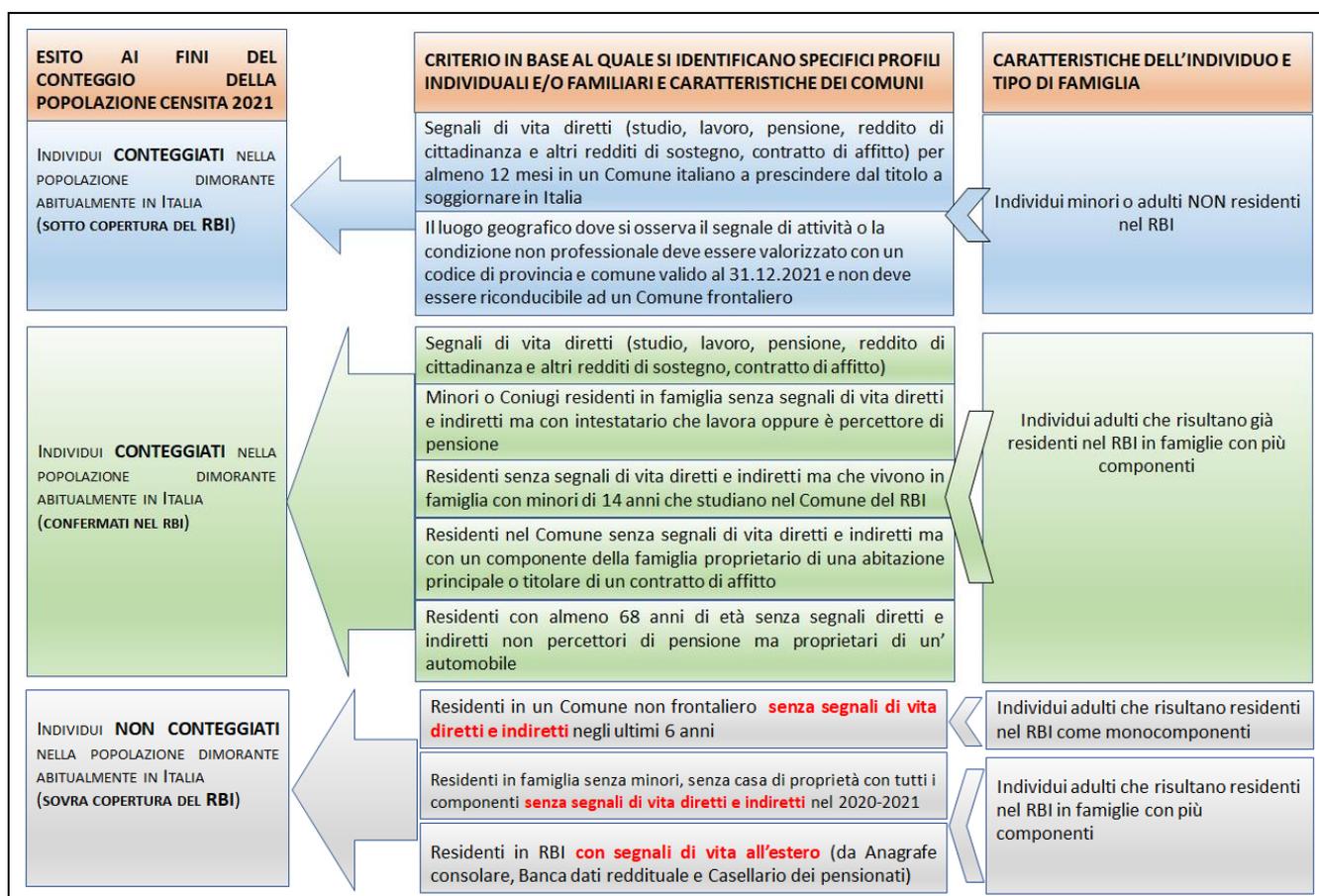
Inoltre, nel 2021, grazie alla disponibilità dei dati di indagine, attraverso i risultati del modello a classi latenti, illustrato nel paragrafo precedente, e ad opportune analisi e valutazione da parte degli esperti tematici dell'Istat si è arrivati a identificare, a valle del processo di AIDA, alcuni criteri in base ai quali è stato possibile considerare anche i fattori di contesto che consentono di migliorare il processo di identificazione della sovra e sotto copertura di RBI.



La definizione finale dei criteri per l'individuazione dei profili di sopra e sotto copertura è stata una combinazione di risultati del modello a classi latenti e la valutazione degli esperti tematici.

Ad esempio, in riferimento alla sotto copertura, il Prospetto 6 mostra che per tutti gli individui di cittadinanza straniera per i quali sono stati identificati i segnali di vita diretti, il permesso di soggiorno non rappresenta un requisito ai fini della identificazione della sotto copertura, mentre se il segnale di vita diretto è localizzato in un Comune frontaliere tutti gli individui stranieri coinvolti non sono considerati dimoranti abitualmente in Italia per le ovvie ragioni connesse ai movimenti frontalieri per lavoro o studio da parte di cittadini dei paesi esteri confinanti con l'Italia. In ogni caso, vale la pena considerare che se la fonte amministrativa di origine non avesse valorizzato il comune del domicilio fiscale, il luogo del contratto di affitto, di lavoro o di studio non sarebbe possibile attribuire un Comune di dimora abituale agli individui classificati in sotto copertura.

PROSPETTO 6. CRITERI PER L'IDENTIFICAZIONE DEGLI INDIVIDUI DIMORANTI ABITUALMENTE IN ITALIA



Fonte: Istat, 2022

Un discorso analogo è stato fatto anche in riferimento agli individui senza segnali diretti e indiretti che identificano la sopra copertura di RBI. In particolare, a valle del processo di AIDA e grazie alle evidenze emerse dall'applicazione del modello sono stati definiti ulteriori criteri o profili di individui che, seguendo l'approccio iterativo e la logica gerarchica, hanno portato alla costruzione di nuove variabili a livello individuale o familiare (Prospetto 6). In base a questo ulteriore passaggio è stato possibile, in alcuni casi, rafforzare l'assenza del segnale di vita sul territorio già risultante dall'output di AIDA e, quindi, avvalorare l'esito di sopra copertura degli individui residenti di RBI senza segnali diretti e indiretti; invece in altri casi la creazione di nuove variabili associate al contesto familiare degli individui, classificati in prima istanza come sopra copertura di RBI, ha reso possibile il loro recupero tra le persone abitualmente dimoranti in Italia (dunque ai fini del conteggio di popolazione è stata confermata la loro residenza nel RBI).



Inoltre, è stata confermata la residenza nel RBI ai fini del conteggio di tutti gli individui dei comuni molto piccoli per i quali gli indicatori longitudinali di tenuta delle anagrafi comunali è risultata molto robusta. Indicazioni in tal senso sono emerse anche da numerose analisi esplorative condotte dai ricercatori dell'Istat nel corso degli ultimi 5 anni.

Sono stati anche confermati come residenti anche tutti gli individui che vivono nelle convivenze anagrafiche. Questo criterio è stato definito a seguito di una attività di verifica effettuata dall'Istat nei primi mesi del 2021, nel corso della quale sono stati validati, presso tutte le anagrafi comunali, sia gli indirizzi sia il dato aggregato di popolazione corrispondente precaricati dall'Istituto sul Sistema di Gestione della Rilevazione. Verifiche molto simili sono state effettuate anche per la popolazione dei grandi anziani (persone con almeno 98 anni di età) per i quali si è proceduto alla loro conferma della dimora abituale nel RBI.

In sintesi, le persone individuate con i segnali di vita diretti del prospetto 6 e quelle individuate con i segnali di vita indiretti (i CONFERMATI del prospetto 6) rappresentano, in considerazione delle conferme e delle aggiunte derivate dall'individuazione della sotto copertura, gli aggregati fondamentali per il conteggio della popolazione.

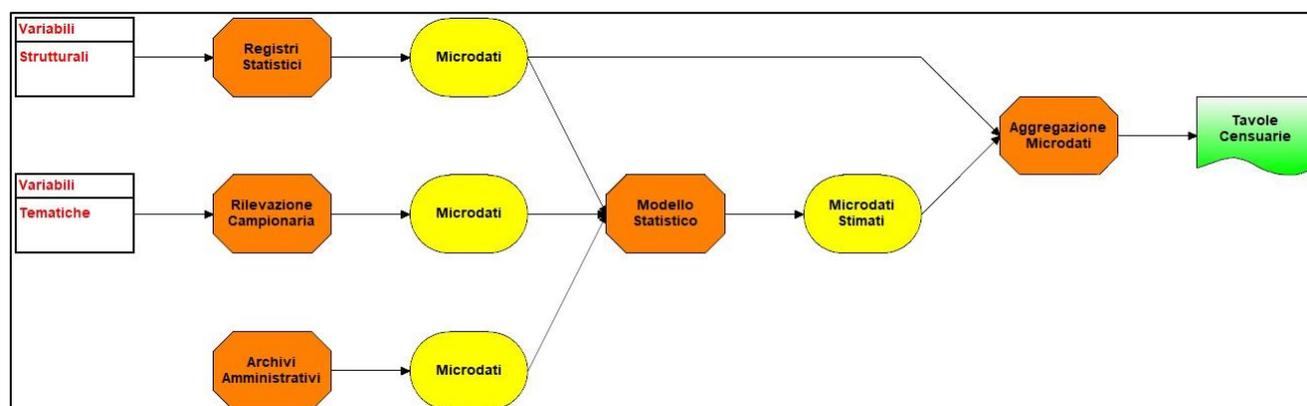
Architettura per la stima del grado di istruzione e degli ipercubi del Censimento permanente

La transizione dal Censimento tradizionale al Censimento permanente ha imposto una totale revisione, sia concettuale che metodologica, del processo di costruzione degli ipercubi censuari. La nuova strategia è finalizzata a massimizzare l'integrazione tra il contributo informativo delle indagini campionarie del Censimento permanente e i dati già inclusi nei registri statistici e negli archivi amministrativi, in modo da assicurare, oltre ad una solida base inferenziale, la completa coerenza del sistema delle statistiche censuarie.

Ciò ha comportato l'abbandono del processo di stima adottato nel 2011 che, prevedendo per alcune variabili una rilevazione esaustiva e per altre campionaria, aveva implicato per le seconde il riporto all'universo delle informazioni campionarie. La scelta fatta per la definizione dell'architettura alla base della costruzione degli ipercubi censuari ricalca, in un certo qual modo, quanto avveniva nelle tornate censuarie precedenti al 2011, nelle quali, essendo prevista la raccolta delle caratteristiche socio-demografiche attraverso un'indagine esaustiva per tutte le variabili, gli ipercubi censuari venivano desunti mediante aggregazione di dati individuali. Precisamente, l'approccio adottato (rappresentato nel Prospetto 7) consiste nella creazione di una base di microdati individuali che, per il grado di istruzione come per tutti gli ipercubi del censimento permanente, utilizza:

- ✓ per le principali variabili demografiche, nello specifico genere, età e cittadinanza, i dati individuali presenti nei registri statistici, che sono denominate *variabili strutturali*;
- ✓ per le restanti variabili, i valori stimati attraverso modelli statistici che usano congiuntamente i dati di indagine e le informazioni presenti nei registri e negli archivi amministrativi. La variabile validata e diffusa per il 2021 che ricade in questa categoria, denominate *variabili tematiche*, è quella relativa al grado di istruzione.

PROSPETTO 7. SCHEMA SINTETICO DELL'ARCHITETTURA PER LA PRODUZIONE DEGLI IPERCUBI DEL CENSIMENTO PERMANENTE DELLA POPOLAZIONE



Fonte: Istat, 2022.



Quindi, gli ipercubi censuari sono calcolati mediante la semplice aggregazione dei valori contenuti nella base dei microdati. Questa architettura è perfettamente consistente con l'informazione contenuta nel sistema dei registri e rispetta, inoltre, i requisiti di coerenza interna, nel senso che una variabile presente in più ipercubi non vede la sua distribuzione statistica cambiare a seconda dell'ipercubo da cui è desunta.

Ipercubo sul grado di istruzione

Il titolo di studio più elevato conseguito da un individuo è stato rilevato sulla popolazione di 9 anni e più. Il corrispondente ipercubo, schematizzato nel prospetto 8, è interamente desunto dalle seguenti fonti di dati:

- il Registro di Base degli Individui al 31.12.2021 corretto per le componenti di sopra e sotto copertura dei segnali amministrativi del Censimento permanente 2021 per quanto riguarda la variabile sesso;
- il Registro Tematico del Titolo di Studio (RTS) per quanto riguarda la variabile grado di istruzione.

Quest'ultimo Registro è il risultato dell'integrazione tra le seguenti fonti di dati:

- la Base Informativa su Istruzione e Titoli di Studio (nella quale sono registrati i dati sui percorsi di studio e sui titoli conseguiti nel territorio italiano dal 2011 in poi);
- il 15° Censimento generale della popolazione e delle abitazioni del 2011;
- l'indagine campionaria del Censimento permanente relativa all' annualità 2021.

PROSPETTO 8. TAVOLA SINOTTICA DELL'IPERCUBO SUL GRADO DI ISTRUZIONE

Popolazione residente di 9 anni e più	Sesso (3) x Grado di istruzione (9)
Sesso (3)	Grado di istruzione (9)
Totale	Totale
Maschi	Analfabeti
Femmine	Alfabeti privi di titolo di studio
	Licenza di scuola elementare
	Licenza di scuola media inferiore o di avviamento professionale
	Diploma di istruzione secondaria di II grado o di qualifica professionale (corso di 3-4 anni) compresi IFTS
	Diploma di tecnico superiore ITS o titolo di studio terziario di primo livello
	Titolo di studio terziario di secondo livello
	Dottorato di ricerca/Diploma accademico di formazione alla ricerca

Fonte: Istat, 2022.