Il processo di produzione dei dati di popolazione legale

1. Il processo di produzione della popolazione legale

In occasione del Censimento 2011, insieme ai dati relativi alla popolazione legale, pubblicati sulla Gazzetta Ufficiale (totale popolazione residente per comune), vengono diffusi i dati relativi alla distribuzione della popolazione residente per sesso, età e cittadinanza (italiana/straniera) per ciascuno degli 8.092 comuni.

La determinazione della popolazione legale dei comuni non può essere effettuata solo sulla base dell'esito del conteggio della popolazione da parte degli Uffici Comunali di Censimento (UCC), necessitando di controllo e validazione centralizzati che garantiscano l'accuratezza e la coerenza dei dati.

Tale attività viene effettuata dall'Istat, a partire dai dati inseriti dagli UCC nel Sistema informatizzato di Gestione della Rilevazione (SGR) per lo svolgimento del confronto censimento-anagrafe, che ha costituito l'ultima fase della rilevazione censuaria. Essa si articola in una serie di fasi, che si susseguono nel cosiddetto "processo di produzione della popolazione legale".

Per garantire la qualità del processo di produzione, è stato realizzato un sistema composto da un ambiente di controllo e correzione, all'interno del quale sono state eseguite tutte le procedure di controllo della coerenza dei dati rilevati, e da un ambiente di interrogazione, che ha permesso di effettuare le verifiche di qualità sui dati.

Il processo di controllo e validazione è riassunto nello schema allegato (Appendice 1 – Diagramma di flusso della produzione della popolazione legale) ed è stato effettuato per lotti comunali. Il diagramma mostra il passaggio tra i diversi stati che un lotto comunale ha assunto a seguito di procedure di correzione, attività di validazione o scambi tra diversi ambienti (correzione => interrogazione => diffusione).

Le fasi di controllo e correzione sono state:

- chiusura di SGR, con controlli sugli esiti del confronto censimento-anagrafe;
- controllo e correzione delle variabili Sesso, Età e Cittadinanza;
- individuazione ed eliminazione dei duplicati tra le persone censite.

Le fasi di validazione hanno compreso sia attività eseguite attraverso procedure statistiche, sia un'attività interattiva basata sulla visualizzazione di report di controllo:

- definizione dei dati per i confronti;
- esame delle anomalie e calcolo degli indicatori di qualità comunali di supporto ai controlli;
- pre-validazione: identificazione dei comuni da inviare in validazione interattiva;
- validazione interattiva;
- controllo tavole di simil-diffusione.

Una volta validati, i dati da diffondere sono stati caricati su I.Stat (sistema di diffusione) effettuando ulteriori verifiche per assicurare l'assenza di errori tecnici.

Nei paragrafi successivi vengono descritte in dettaglio le fasi suddette. In particolare:

- nel paragrafo 2 si descrivono i controlli di coerenza sulle operazioni di confronto censimento-anagrafe e i controlli quantitativi sui relativi bilanci prodotti in SGR e inviati all'Istat da ciascun responsabile di UCC;

- nel paragrafo 3 si descrivono i controlli effettuati sulla presenza e coerenza delle informazioni relative al sesso, alla data di nascita e alla cittadinanza delle persone censite;

- la fase relativa alla de-duplicazione delle persone censite è oggetto del paragrafo 4, nel quale si descrivono le procedure di identificazione e trattamento degli individui e delle famiglie rilevati più volte nel territorio di uno stesso comune o in quello di comuni diversi;

- nei paragrafi 5 e 6 si descrivono le attività di preparazione dei dati da utilizzare per i controlli, sia interattivi che statistici. In particolare nel paragrafo 5 sono descritti i principali aggregati utilizzati come termine di paragone per validare i dati provenienti dalla chiusura di SGR; nel paragrafo 6 si descrivono le modalità di identificazione degli indicatori di qualità delle operazioni censuarie dei comuni e quelle di determinazione delle soglie di "rischio";

- nel paragrafo 7 viene illustrata la "fase di pre-validazione", consistente in una procedura di validazione statistica dei dati comunali, basata sul confronto (in termini di totale e di distribuzione per sesso, età e cittadinanza) tra la popolazione censita di ciascun comune e la corrispondente popolazione risultante dai dati di fonte amministrativa (POSAS e STRASA), ai fini dell'individuazione di casi residuali da sottoporre a validazione interattiva;

- il paragrafo 8 descrive la fase della validazione interattiva, nell'ambito della quale sono stati effettuati ulteriori controlli sui comuni individuati nell'ambito dei controlli statistici effettuati nella fase precedente (analisi della dissomiglianza e *outlier*). La validazione interattiva si è avvalsa di un'applicazione web di supporto sviluppata *in house* e di report interattivi appositamente predisposti nell'ambito del data warehouse di lavoro (DW o Data Warehouse Primario);

- infine, sempre nel paragrafo 8, viene descritta l'ultima fase relativa al caricamento dei dati di popolazione legale nelle basi dati per la diffusione e ai relativi controlli.

2. Chiusura di SGR e controlli formali

Dopo la dichiarazione di chiusura dell'attività di confronto censimento-anagrafe effettuata dal responsabile dell'UCC tramite la funzione "chiudi confronto" di SGR, l'Istat ha effettuato per ciascun comune i seguenti controlli sui dati:

1. per ogni questionario restituito in forma cartacea o via web deve essere stato effettuato il confronto censimento-anagrafe e la lavorazione deve risultare chiusa;
2. per ogni questionario relativo a famiglie/convivenze non trovate, tutti gli individui presenti nella Lista Anagrafica Comunale (LAC) devono essere stati dichiarati irreperibili;

3. per ogni questionario compilato d'ufficio deve essere stato effettuato il confronto censimento-anagrafe;
4. per ogni individuo presente in LAC all'8 ottobre deve essere valorizzato uno stato del confronto censimento-anagrafe tra quelli ammissibili: censito residente, censito residente ad altro indirizzo, irreperibile;
5. per ogni individuo presente in Lista A[1] e non presente in LAC deve essere valorizzato uno dei seguenti stati del confronto: censito non residente o eliminato;
6. per nessun individuo cancellato dalla LAC (variazione anagrafica tra LAC al 31.12.2010 e LAC all'8.10.2012) deve essere valorizzato lo stato del confronto censimento-anagrafe;
7. gli individui dichiarati irreperibili al confronto non devono essere presenti in Lista A;
8. ogni individuo dichiarato irreperibile e al contempo censito come residente ad un indirizzo diverso da quello della LAC deve essere presente nella lista dei censiti ad altro indirizzo e non tra gli irreperibili.

Dopo aver effettuato i controlli sui dati individuali, sono state effettuate una serie di verifiche quantitative sui bilanci ad hoc[2] certificati dai comuni, riguardanti la quadratura delle diverse poste di individui e famiglie censite.

3. Controllo e correzione delle variabili SESSO, ETA' e CITTADINANZA

I dati individuali presenti in SGR) hanno rappresentato l'insieme di riferimento per l'implementazione e la realizzazione del processo di controllo e correzione delle principali variabili demografiche individuali. Oltre ai dati in SGR, per le famiglie che hanno compilato il questionario via web sono state utilizzate anche le informazioni presenti nei questionari[3].

Le variabili demografiche sottoposte al processo di controllo e correzione sono: SESSO, CITTADINANZA (italiana/straniera) e GIORNO, MESE e ANNO DI NASCITA, necessari per la determinazione dell'ETÁ IN ANNI COMPIUTI.

[1] La Lista A dei Fogli di famiglia (Lista dei Fogli di convivenza) è l'elenco delle persone abitualmente dimoranti nell'alloggio (convivenza) ovvero l'elenco dei componenti della famiglia (convivenza), per ciascuno dei quali doveva poi essere compilato il questionario individuale. Al momento del confronto censimento-anagrafe, l'UCC doveva inserire (confermare per i censiti già presenti in LAC) in SGR alcuni dati per ciascun componente della Lista A/Lista dei fogli di convivenza.
[2] I bilanci *ad hoc* contengono le informazioni di riepilogo del confronto censimento-anagrafe, ottenute in automatico a chiusura delle operazioni di confronto eseguite dagli UCC attraverso SGR.
[3] Per le famiglie che hanno compilato il questionario cartaceo, poiché la fase di acquisizione dei questionari stessi non è ancora conclusa, ai fini del controllo e della correzione delle variabili sesso, età e cittadinanza sono state invece utilizzate le sole informazioni presenti in SGR.

Nel processo di controllo e validazione dei dati sono stati considerati tutti gli individui:

- censiti e presenti nella Lista Anagrafica Comunale (LAC); per questi individui le informazioni utilizzate nel processo sono state quelle registrate in anagrafe e trasmesse all'Istat con la LAC e, ove presenti, quelle contenute nel questionario web;

- nuovi censiti (ovvero censiti non presenti nella LAC all'8 ottobre); per questi individui si è fatto riferimento alle informazioni inserite in SGR dall'UCC in fase di confronto censimento anagrafe.

E' opportuno sottolineare che fra le informazioni presenti in SGR, quindi nella LAC di ciascun comune, il codice fiscale è di fondamentale importanza per la verifica delle variabili sesso, cittadinanza e data di nascita. Come è noto infatti, l'algoritmo per la determinazione del codice fiscale opera sulla base di: cognome e nome, anno di nascita, mese di nascita, giorno di nascita (inferiore a 32 per i maschi e superiore a 40 per le femmine), luogo di nascita e, come ultimo carattere, una lettera di controllo per la verifica della validità del codice fiscale stesso.

Nello schema 1 sono riportate, per ciascun gruppo di individui (censiti presenti in LAC e nuovi censiti) e per ciascuna variabile sottoposta a controllo, le fonti principali e le variabili ausiliare utilizzate nel processo di controllo e correzione.

Schema 1: Gruppi di individui per fonte e variabili

	Variabile	Fonte per ciascuna variabile			Variabili ausiliarie
Censiti presenti in LAC	SESSO	SGR	CF	WEB	codice famiglia, nome
	CITTADINANZA	SGR		WEB	CF, codice famiglia, età, nome e cognome
	ANNO DI NASCITA	SGR	CF	WEB	codice famiglia, relazione di parentela, cittadinanza + dati demografici di sintesi
Nuovi censiti	SESSO	SGR			codice famiglia, nome
	CITTADINANZA	SGR			codice famiglia, età, nome e cognome
	ANNO DI NASCITA	SGR			codice famiglia, relazione di parentela, cittadinanza

Per ogni individuo è stato analizzato il valore di ciascuna variabile in ciascuna delle fonti presenti, verificandone la concordanza/discordanza. Nei casi in cui un individuo presentava un valore valido della variabile e tale valore era concordante tra più fonti presenti, l'informazione è stata confermata; di contro, quando si è rilevata una discordanza fra le fonti o in una di esse emergeva un valore anomalo, errato o mancante, le regole deterministiche di imputazione del dato sono state individuate analizzando le distribuzioni semplici e congiunte delle variabili di interesse in relazione con le variabili ausiliarie. Le imputazioni di tipo deterministico sono state effettuate in modo da eliminare il rischio di erronea attribuzione del valore della variabile. Per tutti i casi particolarmente anomali e per sottopopolazioni di particolare interesse come i bambini fra 0 e 10 anni, gli ultracentenari e la popolazione straniera, sono stati implementate procedure *ad hoc* per determinare con il massimo di affidabilità le variabili in esame.

Considerando che l'accuratezza della misurazione di una variabile è anche funzione delle modalità che essa assume, il minor carico per il processo di controllo e correzione è stato osservato per la variabile SESSO. Infatti, solo per un numero marginale di casi sono stati riscontrati valori anomali, mancanti o errati, corretti quasi tutti mediante ricorso al codice fiscale formalmente valido. Marginali sono stati anche i casi di incongruenza fra le fonti, risolti mediante il ricorso a variabili ausiliarie.

Per la variabile CITTADINANZA il processo di controllo e correzione ha avuto come obiettivo quello di distinguere la popolazione italiana da quella straniera (ovvero non ha riguardato lo stato estero di cittadinanza per i cittadini stranieri – questa informazione sarà oggetto di un successivo rilascio). Sono state eseguite prima le procedure di imputazione deterministica dei codici di cittadinanza mancanti (a causa di errori sistematici compiuti dagli UCC in fase di invio della LAC) e successivamente i controlli di validità e concordanza fra le fonti. Per questa variabile, il codice fiscale è stato utilizzato come variabile ausiliaria; infatti il luogo di nascita, pur non coincidendo necessariamente con il paese di cittadinanza, può tuttavia esserne considerato una *proxy*, utile come riferimento per la determinazione della cittadinanza di individui con valori mancanti, anomali o errati, anche grazie al confronto con gli altri individui della famiglia. Nei pochi casi in cui la variabile cittadinanza risultava anomala, errata o mancante per tutti gli individui della famiglia, sono state utilizzate le informazioni provenienti dalle variabili ausiliarie (luogo di nascita del codice fiscale e nome e cognome).

Nella tabella che segue è riportata la distribuzione delle imputazioni eseguite durante il processo di controllo e correzione della variabile CITTADINANZA.

Tabella 1 - Distribuzione delle imputazioni della variabile CITTADINANZA

Cittadinanza PRIMA	Cittadinanza DOPO			Totale
	Italiana	Straniera	Straniera (con codice Stato estero da determinare)	
Mancante, errata o anomala	309.981	40.282	4.852	355.115
Italiana	-	997	1	998
Straniera	29.945		-	29.945
Totale	**339.926**	**41.279**	**4.853**	**386.058**

All'ANNO di NASCITA spetta il ruolo di variabile *pivot* per il questionario di censimento, in quanto condiziona direttamente o indirettamente la compilazione di tutte le sottosezioni del questionario. Per questo motivo le regole di controllo e correzione sono state determinate sulla base di vincoli stringenti. Solo nel caso di coincidenza della variabile nelle tre fonti (35,43% degli individui) il valore è stato confermato senza ulteriori controlli. In tutti gli altri casi, le analisi delle distribuzioni semplici e congiunte hanno portato alla conferma del dato in caso di concordanza fra due fonti, purché il valore della variabile anno di nascita fosse compreso fra il 1920 e il 1999. Per i nati tra il 2000 e il 2011 e per la popolazione ultracentenaria, sussistendo rischi di errata attribuzione derivanti dalla struttura del codice fiscale[4], le procedure di verifica hanno fatto uso di tutte le informazioni provenienti dalle variabili ausiliarie. Per ridurre il rischio di errata attribuzione dell'anno di

[4] Per indicare l'anno di nascita vengono utilizzate solo 2 cifre, quindi i nati nel 1900 e i nati nel 2000 sono identificati dalle stesse cifre. Lo stesso dicasi per i nati nel 1901 e nel 2001.

nascita di queste sottopopolazioni, si è fatto ricorso anche al confronto con i dati individuali della "Rilevazione degli iscritti in anagrafe per nascita".

Nei casi in cui l'affidabilità delle procedure deterministiche di correzione dell'anno di nascita si è rivelata insufficiente, gli individui con dato mancante/incoerente sono stati considerati insieme agli altri componenti della famiglia, al fine di procedere alla correzione interattiva del dato.

Nella tabella che segue è riportata la distribuzione delle variazioni relative alla variabile ANNO DI NASCITA, per ampiezza demografica del comune di residenza dei censiti.

Tabella 2 - Distribuzione degli esiti del processo di controllo e correzione della variabile ANNO DI NASCITA per ampiezza demografica dei comuni

Ampiezza demografica del comune	Concordanza tra le fonti			Nuovi censiti	Imputazioni *ad hoc* o manuali	Totale censiti
	Tre fonti presenti e concordanti	Due fonti presenti e concordanti	Altro			
Fino a 1.000	438.586	609.495	5.281	8.273	649	1.062.284
Da 1.001 a 3.000	1.842.835	2.879.955	30.219	35.365	2.654	4.791.028
Da 3.001 a 5.000	1.540.274	2.860.210	36.061	32.012	2.461	4.471.018
Da 5.001 a 10.000	2.694.686	5.609.869	21.002	72.128	2.274	8.399.959
Da 10.001 a 15.000	1.919.550	3.850.406	15.842	55.456	1.526	5.842.780
Da 15.001 a 30.000	2.875.793	5.557.066	20.709	80.713	3.958	8.538.239
Da 30.001 a 50.000	2.170.018	4.080.786	16.016	72.784	2.183	6.341.787
Da 50.001 a 100.000	2.190.654	4.031.287	17.123	77.583	1.579	6.318.226
Da 100.001 a 250.000	2.410.732	2.390.831	12.940	62.557	821	4.877.881
Da 250.001 a 500.000	774.341	1.037.877	4.341	36.016	558	1.853.133
Da 500.001 a 1 milione	937.582	2.075.067	16.732	47.965	765	3.078.111
Più di 1 milione	1.262.814	2.465.590	31.756	97.512	1.626	3.859.298
Totale	**21.057.865**	**37.448.439**	**228.022**	**678.364**	**21.054**	**59.433.744**
% sul totale dei censiti	**35,43**	**63,01**	**0,38**	**1,14**	**0,04**	**100,00**

Per la correzione delle variabili GIORNO e MESE di NASCITA il processo è stato identico a quello seguito per la variabile ANNO, senza però fare ricorso a procedure *ad hoc*.

Al termine della sistemazione delle variabili GIORNO, MESE e ANNO di NASCITA, sono stati identificati e cancellati i record relativi a bambini nati dopo la data di riferimento del censimento (9 ottobre 2011) ed erroneamente censiti.

Infine, si è proceduto al calcolo dell'ETA' in ANNI COMPIUTI al 9 ottobre 2011.

4. Individuazione ed eliminazione dei duplicati

Per duplicazione si intende la rilevazione di uno stesso individuo per più di una volta nell'ambito del Censimento. La presenza di duplicazioni conduce a un errore per eccesso nella determinazione della popolazione residente in Italia. La doppia rilevazione può avvenire nell'ambito di uno stesso comune (duplicazione intra-comunale) oppure tra più comuni (duplicazione inter-comunale). Entrambi i tipi di errore

possono avere conseguenze in termini di distorsione delle distribuzioni di frequenza relative ad altre variabili individuali.

Nelle precedenti tornate censuarie l'Istat non acquisiva i dati anagrafici nominativi dei censiti, quindi non era possibile identificare eventuali individui censiti in due comuni diversi, mentre la gestione delle duplicazioni intra-comunali veniva delegata ai singoli UCC, senza possibilità di controllo da parte dell'Istat.

Le innovazioni di metodo e tecniche introdotte con il censimento del 2011, relative all'uso di liste pre-censuarie e alla pluralità dei canali di consegna, compilazione e restituzione dei questionari, hanno accresciuto rispetto al passato il rischio di duplicazione. Nondimeno, le stesse innovazioni hanno consentito all'Istat di effettuare la ricerca ed eliminazione dei duplicati nell'ambito del processo di controllo, correzione e validazione dei dati della popolazione legale.

Due sono i possibili tipi di duplicazione:

- l'inclusione di un individuo per due o più volte nello stesso questionario familiare riferito al medesimo alloggio;
- l'inclusione di un individuo in due (o più) questionari in due (o più) alloggi differenti.

Nel primo caso, l'identificazione del duplicato è immediata, poiché i due fogli individuali compilati dallo stesso individuo riportano il medesimo codice identificativo di famiglia. Nel secondo caso, la duplicazione deve essere individuata attraverso il confronto fra le variabili identificative degli individui riportate nella "Lista A" di questionari con codici di famiglia differenti.

Oggetto di questa fase del processo di controllo e correzione è il secondo tipo di duplicazioni, essendo il primo già risolto in fase di caricamento dei dati in SGR.

La ricerca dei duplicati è stata effettuata con criteri deterministici utilizzando come chiave di abbinamento il *codice fiscale*, sia completo (ove disponibile) che ricostruito sulla base delle informazioni anagrafiche individuali presenti nella "Lista A" del questionario. A seguito di ulteriori controlli automatici e manuali, tenendo conto anche dell'entità delle differenze tra nominativi individuali completi ma parzialmente discordanti, sono state individuate circa 178.000 duplicazioni (0,3% della popolazione censita). Inoltre, applicando su un campione di province metodi probabilistici per la ricerca dei duplicati, più onerosi ma più accurati di quelli utilizzati come standard, si è stimato che i duplicati non individuati attraverso la procedura standard costituiscono un residuo di dimensioni nell'ordine delle 11.000 duplicazioni.

L'eliminazione dei duplicati ha seguito regole deterministiche, ad esempio: favorire il record rilevato rispetto a quello censito d'ufficio, favorire il record proveniente da *Long Form* rispetto allo *Short Form*, favorire il record censito con altri componenti della famiglia rispetto a quello rilevato singolarmente e quello rilevato in convivenza rispetto a quello, riferito allo stesso individuo, rilevato in famiglia. La gestione delle relazioni tra le regole e la loro non esaustività rispetto alla possibile casistica hanno consentito di eliminare, con un buon grado di affidabilità della decisione, circa 148.000 duplicazioni (circa 30.000 duplicati sono stati invece mantenuti tra i censiti). In conclusione, l'identificazione e l'eliminazione dei duplicati del censimento, i cui

risultati sono sintetizzati nella tabella 3, ha permesso di verificare che il numero di duplicazioni generate dal censimento è stato contenuto, permettendo inoltre di cancellarne la maggior parte.

Tabella 3 - Sintesi dei risultati dell'individuazione ed eliminazione dei duplicati – livello nazionale

Comuni con duplicati (% su tutti i comuni)	Comuni che hanno subito cancellazioni (% su tutti i comuni)	Record censiti nei 7638 comuni (% su tutti i censiti)	Duplicati (coppie di record)			Record cancellati		
			Inter comun.	Intra comun.	Totali	Inter-com (% sui duplicati inter)	Intra-com (% sui duplicati intra)	Totali (% sui duplicati totali)
7.638 (94,0%)	7.172 (88,6%)	59.308.000 (99,5%)	146.193	31.823	178.016	128.071 (87,6%)	20.046 (63,0%)	148.117 (83,2%)

5. Dati utilizzati per i confronti di validazione

Terminate le attività di controllo e correzione comportanti modifiche dei record individuali, si è passati alla validazione, avente come obiettivo le verifiche di coerenza interna ai dati e con fonti ausiliarie. Le fonti utilizzate per la validazione sono state le seguenti:

- dati provenienti da SGR, relativi a popolazione censita e irreperibile;
- dati relativi al Censimento della Popolazione e delle Abitazioni del 2001;
- dati provenienti dalle basi dati POSAS (Rilevazione anagrafica della popolazione residente comunale per sesso, anno di nascita e stato civile) e STRASA (Rilevazione anagrafica della popolazione straniera residente per sesso ed anno di nascita);
- primi risultati del Censimento della Popolazione e delle Abitazioni 2011.

Per ciascuna fonte sono stati calcolati valori aggregati per suddivisione territoriale (regione, provincia, comune), sesso, cittadinanza, età, classi di età (intervalli quinquennali e decennali), così da consentire l'analisi e il confronto tra grandezze congruenti provenienti da più fonti.

Con riferimento alla base dati SGR, oltre ai dati già menzionati, sono state prese in considerazione ulteriori informazioni:

- gli stati del confronto censimento anagrafe risultanti dal bilancio ad hoc (censito residente, censito residente ad altro indirizzo, censito non residente e irreperibile al censimento);
- le variabili relative alla ripartizione territoriale e all'ampiezza demografica di ciascun comune di residenza.

6. Esame delle anomalie e calcolo degli indicatori di qualità

L'obiettivo di questa attività è stata la valutazione del comportamento dei comuni in fase di confronto censimento-anagrafe e della tenuta dell'archivio anagrafico nel periodo intercensuario.

A tal fine, sono stati analizzati 3 indicatori: a) quota di censiti d'ufficio; b) quota di irreperibili; c) quota di duplicati intra-comunali. Sono stati quindi individuati i comuni che presentavano valori anomali in relazione a ciascun indicatore, nonché i comuni che presentavano differenze significative fra la popolazione censita e la corrispondente popolazione risultante dai dati di POSAS e STRASA. Con criteri statistici sono stati definiti i valori-soglia oltre i quali attribuire un livello di attenzione significativo a un determinato comune. L'inclusione di un determinato comune nell'area di criticità rispetto a uno o più indicatori di qualità ha fornito indicazioni di contesto, a disposizione dei supervisori, nell'ambito della fase di validazione interattiva (cfr. paragrafo 8).

Il criterio

I rapporti tra gli individui "censiti d'ufficio" e "duplicati" rispetto al totale degli individui censiti nel comune sono stati assunti come altrettanti indicatori di affidabilità del dato censuario.

Il 95° percentile è stato assunto come valore-soglia della distribuzione dell'indicatore, oltre il quale il comune è stato assegnato all'area di criticità.

Per i casi così individuati (censiti d'ufficio e duplicati), si è verificata la significatività statistica del superamento del valore-soglia tramite due test binomiali in cui il numero di prove è costituito dalla popolazione del comune e il numero di successi è rappresentato rispettivamente dal numero di censiti d'ufficio e da quello dei duplicati. La probabilità di successo relativa ad ogni test è stata posta pari al valore soglia e il livello di errore del test[5] scelto è pari al 10%. In altre parole, tutti i comuni per i quali la differenza tra il tasso dei censiti d'ufficio (o dei duplicati) e il 95° percentile della distribuzione del corrispondente indicatore risultava maggiore di zero in misura statisticamente significativa, sono stati inclusi nell'area di criticità. Mediante l'uso del test binomiale si sono esclusi i comuni di piccole dimensioni che superavano solo di poco la soglia. L'assegnazione delle soglie di attenzione ai comuni è stata di tipo cautelativo, rischiando quindi di includere nell'area di criticità più comuni del necessario, per garantire però che i comuni effettivamente anomali fossero tutti inclusi.

Un criterio analogo è stato utilizzato per la costruzione dell'indicatore relativo alla quota di irreperibili. In questo caso, al denominatore del rapporto è stato posto il totale degli individui inclusi nella LAC, essendo questa la popolazione di origine degli irreperibili al censimento. Inoltre, si è ritenuto utile considerare anche il limite inferiore della distribuzione dei comuni, cioè quelli che presentavano nessun irreperibile o troppo pochi irreperibili (infatti, se una quota elevata di irreperibili può essere considerata sintomo di cattiva tenuta delle anagrafi, anche una quota di irreperibili troppo bassa - nessun irreperibile/pochi irreperibili – può costituire un segnale di allarme in quanto può rappresentare un sintomo di cattiva rilevazione censuaria). A tal fine, si è tenuto conto, oltre che del 95° percentile, anche del 5° percentile, come valore-soglia oltre il quale il comune veniva posto nell'area di criticità. Anche per gli irreperibili è stato adottato un test con criterio binomiale, per evitare di includere nell'area di criticità i piccoli comuni con valori dell'indicatore poco oltre la soglia.

La valutazione delle differenze fra la popolazione censita e quella di POSAS e fra gli stranieri censiti e la popolazione di STRASA è stata effettuata usando per entrambi i confronti la "Differenza relativa simmetrica". Questa è stata misurata come rapporto tra la differenza in valore assoluto (Censiti/POSAS) e la popolazione

[5] Probabilità che la soglia sia stata superata per effetto del caso.

totale censita nel primo caso e rapporto tra la differenza assoluta (Stranieri censiti/STRASA)e la popolazione straniera censita nel secondo caso.

Il 90° percentile è stato scelto come valore soglia oltre il quale il comune ricade nell'area di criticità. Nel caso della popolazione straniera sono stati considerati solo i comuni con oltre 5.000 abitanti, escludendo quelli con dimensioni ridotte per i quali le differenze tra la popolazione censita e la popolazione di STRASA sono scarsamente significative.

Nella tabella 4 è riportata la distribuzione dei comuni per classe di ampiezza demografica e anomalie individuate. La somma dei comuni con e senza anomalie non coincide con il numero totale di comuni perché una parte dei comuni è inclusa in più classi di anomalia.

Tabella 4 - Distribuzione dei comuni per classe di ampiezza demografica e anomalie individuate

Ampiezza Demografica	Numero comuni	Numero comuni senza anomalie	Numero comuni con anomalie				
			Per irreperibili	Per censiti d'ufficio	Per duplicati	Per differenze con Posas e/o Strasa	Popolazione censita
fino a 1.000	1.951	1.691	56	47	13	196	1.062.284
da 1.001 a 3.000	2.602	2.218	92	77	31	258	4.791.028
da 3.001 a 5.000	1.149	972	46	36	17	115	4.471.018
da 5.001 a 10.000	1.187	966	63	35	34	170	8.399.959
da 10.001 a 15.000	479	348	32	23	30	100	5.842.780
da 15.001 a 30.000	417	314	34	29	32	60	8.538.239
da 30.001 a 50.000	166	110	20	14	26	21	6.341.787
da 50.001 a 100.000	95	46	21	11	22	29	6.318.226
da 100.001 a 250.000	34	18	8	5	8	9	4.877.881
da 250.001 a 500.000	6	2	4	1	2	0	1.853.133
da 500.001 a 1 milione	4	0	2	1	4	0	3.078.111
Oltre 1 milione	2	0	2	0	2	2	3.859.298
Totale	**8.092**	**6.685**	**380**	**279**	**221**	**960**	**59.433.744**

7. Identificazione dei comuni da sottoporre a validazione

L'identificazione dei comuni da sottoporre a validazione da parte di operatori è stata guidata dal livello di dissomiglianza percentuale esistente tra le distribuzioni per età, sesso e cittadinanza rilevate al censimento e le corrispondenti distribuzioni rilevate dalle indagini su dati amministrativi sulla "Popolazione residente comunale per sesso, anno di nascita e stato civile" (POSAS) e sulla "Popolazione straniera residente per sesso e anno di nascita" (STRASA).

Per i comuni fino a 5.000 abitanti sono stati eseguiti controlli, oltre che in base alla dissomiglianza tra le distribuzioni per età, anche applicando un modello di regressione lineare tra le frequenze di popolazione (per singolo anno di età, sesso e cittadinanza) di POSAS/STRASA (variabile esplicativa) e le corrispondenti frequenze della popolazione censita (variabile dipendente). L'analisi dei residui del modello ha consentito di identificare e inviare a controllo i comuni con un numero relativamente più elevato di dati *anomali*, cioè aventi scarso adattamento al modello.

Indice di dissomiglianza

L'indice di dissomiglianza utilizzato per i controlli ha assunto la seguente forma

$$D_{j,s,c} = \sum_{i=0}^{100} \left(\frac{f_{j,i,s,c}}{F_{j,s,c}} - \frac{a_{j,i,s,c}}{A_{j,s,c}} \right) * 100$$

con

j, indice di comune

s, sesso (maschio, femmina)

c, cittadinanza (italiana, straniera)

$f_{j,i,s,c}$ frequenza assoluta di censiti nel comune j per profilo di sesso 's', cittadinanza 'c' e i-mo anno d'età

$a_{j,i,s,c}$ frequenza assoluta di individui POSAS/STRASA nel comune j per profilo di sesso 's', cittadinanza 'c' e i-mo anno d'età

$$F_{j,s,c} = \sum_{i=0}^{100} f_{j,i,s,c}$$

$$A_{j,s,c} = \sum_{i=0}^{100} a_{j,i,s,c}$$

e varia da un valore minimo dello 0% nel caso di perfetta uguaglianza tra le due distribuzioni poste a confronto e un massimo del 100% nel caso le due distribuzioni siano allocate in categorie completamente distinte.

Modello di regressione lineare

Il modello di regressione lineare usato per l'identificazione dei valori anomali è stato il seguente:

$$f_{j,i,s,c} = \beta \cdot a_{j,i,s,c} + \varepsilon_{j,i,s,c}$$

con

$$\varepsilon_{j,i,s,c} \sim N(0, \sigma^2), \qquad cov\left(\varepsilon_{j,i,s,c}, \varepsilon_{j',i',s',c'}\right) = 0 .$$

Il modello di regressione ha utilizzato come unità elementari le frequenze assolute dei singoli profili per ETÀ, SESSO e CITTADINANZA osservati nei comuni "*j*".

Il modello di regressione lineare è stato applicato per 108 volte, una per ciascuna provincia, ai 5.699 comuni con meno di 5000 abitanti. In base all'analisi dei residui sono stati definiti *anomali* tutti i profili il cui valore del

residuo "studentizzato"[6] si è rivelato superiore a 4 (complessivamente 18.397 su 2.287.044, pari a circa l'1% delle frequenze elementari).

Identificazione dei comuni

L'identificazione dei comuni da inviare alla validazione dei revisori è stata effettuata separatamente per i comuni con popolazione inferiore a 5.000 abitanti e per quelli con popolazione maggiore o uguale a questa soglia.

Per i comuni con almeno 5000 abitanti l'indice di dissomiglianza percentuale è stato calcolato secondo la formula riportata sopra, con riferimento alle quattro distribuzioni per anno d'età: maschi italiani, femmine italiane, maschi stranieri e femmine straniere. E' stato quindi assunto come valore di soglia[7] quello corrispondente al 90° percentile della distribuzione di ciascuno dei quattro indici di dissomiglianza. In questo modo sono stati selezionati per il controllo 624 comuni eccedenti almeno uno dei valori soglia.

I 26 comuni con popolazione censita maggiore di 150 mila abitanti sono stati inviati tutti a verifica ,anche se 21 di essi non presentavano valori superiori alle soglie.

Per i comuni con meno di 5.000 abitanti, il criterio della dissomiglianza è stato applicato solo in relazione alla distribuzione per età dei maschi italiani[8], considerando separatamente i comuni appartenenti a tre classi di ampiezza di popolazione (<1.000; 1.000-2.999; 3.000-4.999). Successive analisi *ad hoc* sui 1.951 comuni al di sotto dei 1.000 abitanti hanno permesso di escluderli da ulteriori verifiche perché di fatto ininfluenti.

Nella classe dei comuni tra 1.000 e 2.999 abitanti ne sono stati selezionati 267 da sottoporre a verifica, in quanto eccedenti il 90° percentile dell'indice di dissomiglianza; ad essi sono stati aggiunti 61 comuni per i quali più di 10 profili sono risultati *anomali* dal punto di vista del modello di regressione lineare.

Tra i comuni compresi tra 3.000 e 4.999 abitanti ne sono stati selezionati 125 da sottoporre a verifica, ai quali ne sono stati aggiunti 259 che presentavano più di 10 profili *anomali* per il modello di regressione lineare.

8. La validazione interattiva

Una volta identificato l'insieme dei comuni su cui effettuare ulteriori controlli, si è proceduto alla validazione interattiva. I comuni per cui i controlli statistici hanno dato esito positivo sono passati in *batch* allo stato 60 (comune validato in *batch*) mentre i comuni per cui i controlli hanno avuto esito negativo sono stati portati a stato 50 (comune da validare).

[6] Si definisce "studentizzata" la stima del residuo divisa per la stima della propria varianza: $r_{j,i,s,c} = \frac{\hat{\varepsilon}_{j,i,s,c}}{\widehat{var}(\varepsilon_{j,i,s,c})}$

[7] Maschi italiani (2,4%), femmine italiane (2,1%), maschi stranieri (15,0%), femmine straniere (15,0%)

[8] Analisi preliminari hanno evidenziato una identica distribuzione per l'indice riferito alle femmine italiane e distribuzioni molto erratiche per gli indici riferiti agli stranieri, sia maschi che femmine, data la loro scarsa presenza in comuni così piccoli.

8.1. Il processo di validazione interattiva: organizzazione controlli

L'attività di validazione interattiva era finalizzata a effettuare ulteriori controlli sui comuni che non avevano superato i controlli statistici effettuati in fase di pre-validazione. La validazione interattiva era espletata da un gruppo di revisori, coordinati da un gruppo di supervisori/esperti tematici, e consisteva nell'analisi di una serie di report per ciascun comune. L'obiettivo era quello di indagare le differenze tra le distribuzioni "note" (dati di confronto) e i dati censuari, al fine di "spiegare" le differenze.

I comuni in stato 50 (comune da validare) venivano presi in carico dai revisori (ai quali veniva assegnato un certo numero di province, tramite l'applicazione di supporto alla validazione - cfr. paragrafo 8.2.2.) e, per ciascuno di essi, venivano analizzati i report di controllo appositamente predisposti (cfr. paragrafo 8.2.3.).

Il primo report da consultare presentava le piramidi delle età relative al totale della popolazione residente calcolate sui dati censuari e sui dati POSAS e, per facilitarne il confronto, un grafico con le due piramidi sovrapposte. Un altro report presentava le piramidi delle età relative alla popolazione straniera calcolate sui dati censuari e sui dati STRASA. Altri report riguardavano la distribuzione per sesso ed età dei nuovi censiti e degli irreperibili. Infine, dovevano essere consultati i report sugli esiti del confronto censimento-anagrafe e quelli con indicatori di struttura calcolati sui dati censuari e sulle fonti di confronto.

Se i dati relativi al comune analizzato erano sufficienti a 'spiegare' le differenze tra le distribuzioni 'attese' e le distribuzioni al censimento (ad esempio, in corrispondenza di una differenza negativa/positiva tra Censimento e POSAS per una determinata classe di età corrispondeva un numero di nuovi censiti/irreperibili equivalente), il comune veniva validato dal revisore (stato 60). In caso contrario veniva inviato al supervisore (comune non validato - stato 55).

Per i comuni non validati, il supervisore verificava che le differenze anomale segnalate dal revisore non fossero dovute a correzioni effettuate nelle fasi di controllo e validazione della data di nascita (stato 20) o in quella di individuazione/eliminazione duplicati (stato 30). In caso positivo, il comune veniva riportato allo stato 20 o 30 e la relativa procedura veniva ri-eseguita. In caso negativo, il comune veniva validato dal supervisore (stato 61) e incluso nella lista dei comuni da segnalare alla Direzione Centrale delle Statistiche socio-demografiche (DCSA) per ulteriori approfondimenti. Il supervisore analizzava inoltre le informazioni relative agli indicatori di qualità, che costituivano informazioni di contesto utili ai fini della valutazione complessiva da effettuare per procedere alla validazione.

Nel complesso, il 74,2% dei comuni ha superato positivamente la fase di pre-validazione, mentre il 13,8% dei comuni è stato validato dai revisori (validazione interattiva di primo livello) e l'1% dei comuni è stato validato dai supervisori (validazione interattiva di secondo livello).

8.2. Strumenti informatici di supporto al processo di validazione interattiva

Il processo di validazione ha richiesto la predisposizione di un ambiente di interrogazione che consentisse di effettuare le verifiche di qualità interattive sui dati in lavorazione. E' stato quindi realizzato un data warehouse di lavoro (Data Warehouse Primario, cfr. paragrafo 8.2.1.), a partire dal quale le attività di

validazione interattiva sono state espletate attraverso l'utilizzo di due applicazioni realizzate con tecnologie web:

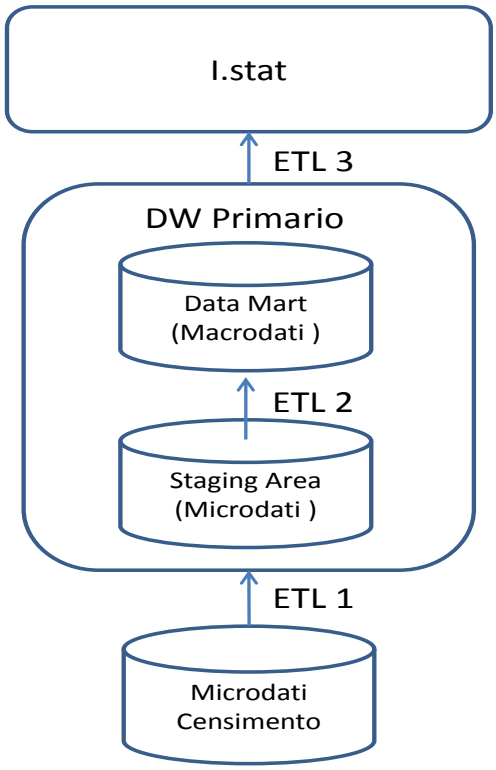
- un'applicazione sviluppata *in house* che garantiva i passaggi di stato e la verifica del flusso (cfr. paragrafo 8.2.2.)
- un'applicazione che ha consentito di accedere in modo interattivo ai dati del Data Warehouse Primario mediante report di controllo appositamente predisposti (cfr. paragrafo 8.2.3.).

8.2.1. Il Data Warehouse Primario

Da un punto di vista funzionale, il Data Warehouse Primario (DWP) ha consentito:

- la rappresentazione multidimensionale della popolazione legale rispetto alle dimensioni finalizzate alla validazione (ad esempio, classe di ampiezza demografica del comune, indicatori di qualità su irreperibili/censiti d'ufficio/duplicati intra-comunali, etc.)
- la rappresentazione multidimensionale della popolazione legale rispetto alle dimensioni finalizzate alla pubblicazione (sesso, età, cittadinanza, territorio)
- la realizzazione di un insieme di misure calcolate finalizzate alla realizzazione della reportistica a supporto della validazione.

Da un punto di vista architetturale, il posizionamento del DWP è illustrato nella figura che segue.



Il DWP si colloca tra il database dei micro-dati censuari e I.stat, che costituisce il sistema di web warehousing corporate adottato dall'Istat. Il collegamento del DWP con questi due livelli, rispettivamente posti a monte e a valle, è stato realizzato mediante procedure automatiche di ETL (Extract, Transform, Load). In generale, le procedure ETL hanno lo scopo di prelevare i dati da una qualsiasi fonte dati, effettuare delle trasformazioni sui dati e procedere al successivo caricamento del risultato delle trasformazioni sulle basi di dati di destinazione. In ambito data warehousing, l'utilizzo di procedure di ETL è previsto per la fase di popolamento del data warehouse. Tuttavia, nell'architettura specifica del DWP le procedure di ETL hanno supportato anche l'estrazione dei dati dal data warehouse per il popolamento di I.stat.

Più in particolare, partendo dal basso della figura sopra riportata, l'alimentazione del DW primario è stata realizzata mediante le procedure identificate con ETL1 in figura, che hanno alimentato una base di dati, denominata Staging Area, in cui sono stati effettuati controlli e riconciliazione delle fonti dati. A partire dalla Staging Area, un ulteriore insieme di procedure, denominate ETL 2, ha permesso il caricamento dei dati finalizzati alla creazione dei cubi multidimensionali (Data Mart[9]). Infine, un terzo insieme di procedure di ETL (ETL3) ha permesso il caricamento dei dati necessari alla pubblicazione sul sistema I.stat. Si noti che gli insiemi di procedure ETL1 e ETL2 sono stati iterati durante il processo di validazione: laddove le analisi effettuate sui report segnalavano la necessità di effettuare delle modifiche sui micro-dati, le procedure di caricamento venivano rilanciate di modo da aggiornare conseguentemente il DWP. Sono state realizzate, in totale, 68 procedure di ETL, ed in particolare 29 ETL1, 36 ETL2 e 3 ETL3.

Il DWP è stato realizzato con la suite di Business Intelligence della Microsoft (Microsoft SQL Server Enterprise Edition vs2), ad eccezione dello strumento di ETL, per il quale si è scelto Pentaho Data Integration - Kettle.

8.2.2. L'applicazione di supporto alla validazione interattiva

A supporto della fase di validazione interattiva dei dati della popolazione legale, è stata sviluppata dall'Istat un'applicazione realizzata con tecnologie Web. Gli utenti dell'applicazione erano costituiti da un nutrito gruppo di revisori dell'Istat, coordinati da alcuni supervisori. A ciascun profilo-utente (revisore e supervisore) erano associate prerogative diverse nell'ambito dell'applicazione, che rispecchiavano i diversi compiti e responsabilità (cfr. paragrafo 8.2.).

Le funzioni principali svolte tramite l'applicazione erano due:

[9] I Data Mart creati modellano i dati censuari secondo una rappresentazione multidimensionale che prevede:
- un'unica misura, rappresentata dal conteggio degli individui.
- un totale di 13 dimensioni (9 finalizzate alla validazione e 4 finalizzate alla pubblicazione).

I dati censuari sono stati confrontati, ai fini della validazione, con un insieme di fonti: POSAS, STRASA, Censimento 2001, Dati Provvisori. Per ciascuna fonte, si è fornita una rappresentazione multidimensionale del conteggio degli individui rispetto alla dimensioni di confronto disponibili.
Inoltre, a partire dai Data Mart creati è stato sviluppato un insieme di "misure", denominate membri calcolati. Questi sono stati utilizzati per la creazione dei report, ad esempio a supporto del calcolo degli indici di dissomiglianza o degli indicatori di differenza con i *data-set* di *benchmark*. La disponibilità dei membri calcolati ha consentito: (i) uno sviluppo più agevole dei report, e (ii) una ottimizzazione nei tempi di risposta connessi alla interrogazione del DWP a partire dai report. Più in particolare, lo sviluppo dei report è stato semplificato in quanto la "logica" di calcolo è stata semplicemente richiamata dai report (invece di essere sviluppata all'interno dei report stessi). Inoltre, la possibilità di pre-computare i membri calcolati (invece di effettuare computazione *on-the-fly*) ha consentito una riduzione dei tempi di accesso ai report in fase di validazione. Il numero totale di membri calcolati che sono stati realizzati è pari a 34.

i) assegnazione dei comuni ai revisori;

ii) validazione del comune.

i) La funzione di assegnazione dei comuni ai revisori, accessibile ai soli supervisori, consentiva di assegnare i comuni da validare ad uno dei revisori tramite un'interfaccia che consentiva di visualizzare elenchi di comuni, eventualmente filtrati con criteri territoriali (selezionando regione e provincia), e di effettuare una selezione singola o multipla. Le assegnazioni effettuate erano modificabili in base alle necessità del processo di validazione.

ii) La funzione di validazione dei comuni consentiva di visualizzare i dati riepilogativi rappresentativi dei comuni in forma tabellare (un comune per riga) e mostrava gli indicatori di qualità utilizzati nel processo di validazione. Era possibile filtrare tale lista tramite parametri territoriali o in base allo stato di elaborazione del comune. Al revisore era consentita la visualizzazione e gestione dei soli comuni a lui precedentemente assegnati, mentre i supervisori potevano agire su tutti i comuni.

Il supervisore o il revisore può procedere alla validazione dei comuni di sua competenza direttamente o nella pagina di dettaglio relativa al singolo comune, a cui era possibile accedere cliccando su un'icona posizionata sulla destra nella riga del comune. La pagina di dettaglio conteneva informazioni di contesto (ad esempio la popolazione residente secondo POSAS), i bottoni di validazione, abitati o disabilitati a seconda dello stato di elaborazione dei comuni selezionati dall'utente, e il bottone 'Inserisci nota', che consentiva ai revisori/supervisori di tenere traccia dell'analisi eseguita per ciascun comune, al fine di documentare il processo di validazione.

Era consentita la validazione contemporanea di più comuni nello stesso stato.

8.2.3. I report per la validazione interattiva

A supporto delle attività di validazione è stata inoltre utilizzata un'applicazione per la visualizzazione interattiva dei report necessari ai revisori e dei supervisori per la validazione. A tal fine, è stato utilizzato un *tool*[10] che ha consentito di pubblicare su WEB i report (configurandone gli accessi riservati ai revisori e ai supervisori), nonché di scaricare i dati in vari formati (PDF, EXCEL, CSV).

I report, utilizzati per la validazione interattiva sono di due tipi:

i) report che hanno consentito di tenere sotto controllo le diverse fasi del processo di produzione e di monitorare lo svolgimento delle attività di validazione (cfr. paragrafo 8.2.1.);

ii) report di validazione ovvero i report utilizzati per l'analisi dei dati comunali e contenenti indici, rappresentazioni grafiche e confronti con altre fonti.

Entrambi i tipi di report fornivano sia una rappresentazione grafica dei dati sia una più dettagliata in forma tabellare.

[10] SQL Server Reporting Services.

i) Report di navigazione e monitoraggio

Per facilitare l'accesso alle informazioni necessarie per la validazione del dato comunale, i singoli report sono stati raggruppati sotto un unico report di navigazione. Ciò ha consentito ai revisori/supervisori di scegliere la provincia e il comune di interesse, di visualizzare l'insieme dei dati e degli indicatori necessari per la validazione e di accedere in modo interattivo a tutti i *Report di Validazione*.

Il report di navigazione fornisce un quadro completo della situazione del comune: le informazioni di sintesi segnalano eventuali situazioni di "allerta" relative al numero di irreperibili, di censiti d'ufficio, di de-duplicati, alla differenza dei censiti rispetto a POSAS, al passaggio di soglia demografica; viene confrontata la popolazione censita con quella POSAS come anche la popolazione straniera con quella STRASA, e gli individui irreperibili con quelli censiti ma non residenti (nuovi censiti); si mostrano informazioni sull'esito del confronto Censimento-Anagrafe e i dati di dettaglio sui censiti d'ufficio; si è inoltre scelto di raffrontare la popolazione censita con altre fonti di confronto quali i dati provvisori, i dati censuari del 2001 ed in particolare i dati POSAS.

Il monitoraggio del sistema di validazione è stato, invece, realizzato attraverso alcuni report che hanno consentito ai supervisori di verificare lo stato della lavorazione dei singoli comuni da parte dei revisori e fornire una visione d'insieme sullo stato dei controlli e della validazione dei comuni di una intera provincia e nel complesso.

ii) Report di validazione

Come detto, nel processo di produzione dei dati della popolazione legale, i report di validazione sono stati utilizzati come supporto informativo per il controllo e la validazione dei comuni che non superavano positivamente la fase di pre-validazione (cfr. paragrafo 7) e necessitavano quindi di ulteriori verifiche.

La struttura e i contenuti dei *Report di validazione* rispondevano quindi alle esigenze di analisi dei revisori, ai quali spettava effettuare verifiche e controlli da realizzare confrontando il dato censuario con i dati di benchmark e gli indicatori.

In particolare ci si è avvalsi di due tipi di report: le piramidi per età, utilizzate per analizzare la struttura della popolazione per classi di età, e i report contenenti indicatori di struttura ritenuti utili all'analisi del dato censuario da validare.

Ad esempio, nel report "Piramide per età Posas-Legale", veniva visualizzata una tabella con la distribuzione per sesso e per classi quinquennali di età della popolazione censita e della corrispettiva popolazione del dato amministrativo di fonte POSAS. Alla tabella si affiancavano le corrispettive piramidi per età e una loro sovrapposizione in forma continua, così da permettere ai revisori di individuare le eventuali classi di età, per maschi e femmine, che evidenziavano visivamente scarti rilevanti tra il dato censuario e quello POSAS.

Nei report dove invece venivano utilizzati indicatori di struttura, si confrontava il dato censuario rilevato sia con i dati POSAS che con il dato censuario del 2001. Come, ad esempio, nel report "Indici Censiti – Posas – Cens2001" dove veniva visualizzato il confronto tra l'età media, il rapporto di mascolinità e gli indici di vecchiaia, di dipendenza e di ricambio e dei censiti con i corrispettivi dati di fonte POSAS e con quelli del Censimento del 2001.

9. Controllo tavole di simil-diffusione e caricamento dati su I.Stat

Le attività connesse alla diffusione dei dati di popolazione legale hanno riguardato tre aspetti:

- la definizione della struttura delle tavole di diffusione
- la predisposizione di report di simil-diffusione[11] dall'ambiente di datawarehouse
- il caricamento dei dati su I.Stat.

Sul data warehouse dell'Istituto i dati consultabili sono organizzati in due gruppi di tavole: "Popolazione residente - Dati definitivi" e "Numero di comuni e popolazione residente - Dati definitivi".

Il primo gruppo è composto da tre tavole:

- "Popolazione legale pubblicata nella G.U. n. XXX del XX/2012" che riporta quanto diffuso in G.U., ovvero la popolazione residente per ciascun comune;
- "Popolazione residente per sesso, singole età e cittadinanza" che visualizza i dati per sesso, singolo anno di età e cittadinanza ai diversi livelli di dettaglio territoriale - comune, provincia, regione, ripartizione e Italia;
- "Popolazione residente per sesso, classi di età e cittadinanza" che contiene le stesse informazioni aggregate per classi di età decennali.

Il secondo gruppo è costituito da tre tavole a livello provinciale che consentono il confronto con i dati dei precedenti censimenti. Esse contengono il numero di comuni e la relativa popolazione residente secondo l'ampiezza demografica degli stessi e secondo la zona altimetrica e il numero dei comuni che rispetto al 14° Censimento della popolazione e delle abitazioni hanno subìto un incremento o un decremento di popolazione, classificati in base alla variazione percentuale della popolazione tra il 2001 e il 2011 (fino al 5 per cento, 5,01-10 per cento, 10,01-15 per cento, 15,01-25 per cento, 25,01 per cento e più).

Le tavole per la diffusione sono state validate usando appositi report, costituiti da tavole di dati aventi la stessa struttura delle tavole da pubblicare su I.Stat. Questi report hanno consentito di visualizzare, attraverso un menù a tendina, dati per regione o provincia o singolo comune. Il confronto tra le tavole di simil-diffusione e le tavole su I.Stat ha consentito da un lato di verificare la coerenza nella costruzione del file prodotto per I.Stat e dall'altro la correttezza del caricamento dei dati sul data warehouse dell'Istituto. Per il controllo dei dati di popolazione legale è stato predisposto un report di simil-diffusione che ha permesso di controllare l'esatta corrispondenza tra quanto pubblicato sulla Gazzetta Ufficiale e quanto presente nel DWH e su I.Stat.

[11] Report con la stessa struttura delle tavole di diffusione da pubblicare su I.Stat.

Appendice 1 - Diagramma di flusso della produzione della popolazione legale