NOTA METODOLOGICA

Elaborazione dei movimenti pendolari a livello sub-comunale per il 15° Censimento generale della popolazione e delle abitazioni 2011

Il 15° Censimento generale della popolazione 2011 contiene una specifica sezione dedicata al pendolarismo (Istat 2015). In questa area tematica sono stati analizzati gli spostamenti delle persone residenti in famiglia o in convivenza che hanno dichiarato di recarsi al luogo abituale di studio o di lavoro, partendo dall'alloggio di dimora abituale e rientrando giornalmente nello stesso.

Come per i precedenti censimenti, oltre alla usuali pubblicazioni dei dati censuari, è stata diffusa la matrice di pendolarismo a livello comunale contenente i dati sul numero di persone residenti che si spostano tra comuni – o all'interno dello stesso comune – classificate per il motivo dello spostamento, per il sesso, per il mezzo di trasporto utilizzato, per la fascia oraria di partenza e la durata del tragitto[1].

Durante il censimento del 2011, i questionari di rilevazione per gli individui residenti sia in famiglia sia in convivenza chiedevano al rispondente di inserire l'indirizzo del luogo abituale di studio o di lavoro. Tale dato veniva scritto come testo libero tramite il canale scelto per la compilazione: per i questionari compilati via Web è stato direttamente disponibile nel sistema informatico a supporto del censimento, mentre, per i questionari compilati in modalità cartacea è stato acquisito attraverso la lettura ottica. L'indirizzo è stato così utilizzato come dato di partenza per lo studio dei movimenti pendolari a livello sub-comunale[2].

La qualità degli indirizzi di studio o lavoro dipende fortemente dal tipo di compilazione del questionario. Nel caso specifico è stata piuttosto alta se proveniente dalla compilazione digitale via Web, mentre molti degli indirizzi del questionario cartaceo, acquisiti otticamente e poi sottoposti ad OCR (riconoscimento ottico dei caratteri), sono stati scartati in quanto non riconoscibili. Inoltre, non tutti i rispondenti hanno fornito questa informazione poiché non era obbligatoria. Il numero totale di indirizzi di studio o lavoro disponibili da fonte censuaria è riassunto nella Tavola 1.

[1] https://www.istat.it/it/archivio/157423

[2] Il presente testo è stato pubblicato in precedenza su Istat, (2016). "Forme, livelli e dinamiche dell'urbanizzazione in Italia" disponibile alla pagina: https://www.istat.it/it/archivio/199520 (Capitolo 9, Sezione 2 e Nota metodologica).

TAVOLA 1 – PENDOLARI (IL CUI LUOGO DI DESTINAZIONE NON È ALL'ESTERO) PER DISPONIBILITÀ DELL'INDIRIZZO DI STUDIO O DI LAVORO - ANNO 2011 (VALORI ASSOLUTI E VALORI PERCENTUALI)

DISPONIBILITÀ DELL'INDIRIZZO	PENDOLARI	
	VALORI ASSOLUTI	VALORI PERCENTUALI
Fonte censuaria	23.424.250	81,3
Senza indirizzo	5.381.190	18,7
- di cui recuperato da altre fonti	*3.786.597*	*13,1*
Totale	**28.805.440**	**100**

Fonte: Istat, 15° Censimento generale della popolazione e delle abitazioni 2011

La mancanza del 18,7 percento degli indirizzi degli individui pendolari dalla fonte censuaria ha reso necessario il loro recupero da altri archivi: è stata così avviata un'articolata attività di analisi, controllo ed integrazione di dati provenienti da più fonti, in modo da ottenere informazioni il più possibile complete sui movimenti pendolari. Tale integrazione ha permesso di recuperare, per più del 70 percento dei pendolari, un indirizzo di studio o lavoro che, da fonte censuaria, era mancante. Come livello minimo di dettaglio sub-comunale è stato definito quello della sezione di censimento: è stato infatti possibile associare la sezione di censimento ad un indirizzo sottoponendo quest'ultimo ad un processo di normalizzazione[3] e geo-codifica[4] tramite opportuno software (cfr. Sezione 2). Il risultato della normalizzazione è legato alla precisione con la quale l'indirizzo è scritto: il *software* potrebbe infatti non riuscire a normalizzare e, di conseguenza, a geo-codificare alcuni indirizzi se la loro qualità è troppo bassa (cfr. Tavola 2) come descritto nella Sezione 3.

TAVOLA 2 – PERCENTUALI DI NORMALIZZAZIONE DEGLI INDIRIZZI ASSOCIATI AI PENDOLARI – ANNO 2011 (VALORI ASSOLUTI E PERCENTUALI)

INDIVIDUI PENDOLARI	VALORE ASSOLUTO	PERCENTUALE DI INDIRIZZI NORMALIZZATI	PERCENTUALE INDIRIZZI NORMALIZZATI CON SEZIONI VALIDE
con indirizzo disponibile (a)	27.210.847	88,46	81,05
Totale	**28.805.440**	**83,56**	**76,57**

Fonte: Istat, 15° Censimento generale della popolazione e delle abitazioni 2011
(a) da fonti censuarie o da altre fonti

[3] La normalizzazione degli indirizzi consiste nell'analisi e correzione delle inesattezze e delle incoerenze riscontrate in un indirizzo, al fine di disporre di un indirizzo nella sua forma universalmente intellegibile.

[4] La geo-codifica consiste nell'attribuire a ciascun indirizzo la corretta localizzazione della sezione di censimento.

È stato così possibile ricavare la sezione di studio o lavoro per gli individui residenti e, dato che si dispone della sezione di residenza, si è riusciti a determinare la sezione di partenza e quella di destinazione di ogni movimento pendolare. Una volta determinati i movimenti pendolari a livello di sezione di censimento, è stato possibile calcolare anche quelli relativi alle suddivisioni sub-comunali di livello superiore a quello della sezione (località, aree di censimento ed aree sub-comunali).

Le attività di elaborazione dei movimenti pendolari sin qui descritte hanno reso possibile, per la prima volta nella storia dell'Istat, il rilascio della matrice di pendolarismo a livello sub-comunale, conteggiando i pendolari che si spostano da una sezione di partenza ad una di arrivo ovunque situate sul territorio nazionale.

1. Il recupero degli indirizzi mancanti tramite integrazione di dati da fonti diverse

La base di calcolo per il recupero degli indirizzi mancanti da fonte censuaria sono stati gli individui pendolari che hanno dichiarato di recarsi per motivi di studio o lavoro in un comune italiano, in quanto l'attività di geo-codifica degli indirizzi è legata al territorio italiano.

Al fine di ottenere un quadro il più possibile completo e preciso sui dati riguardanti il pendolarismo, è stato deciso di recuperare gli indirizzi mancanti (di studio o lavoro) da altre fonti di dati, anch'esse riferite allo stesso periodo temporale, confrontandole e integrandole opportunamente.

Sono stati, pertanto, utilizzati gli archivi amministrativi acquisiti dall'Istituto. Tali archivi sono molto diversi fra loro per formato dei dati, tipologia delle informazioni contenute, riferimento temporale, modalità di ricezione. Ognuno di essi, quindi, richiede uno specifico trattamento e un'appropriata analisi al fine di inserirli in un unico sistema integrato di micro dati. Questo lavoro è estremamente complesso e viene effettuato nel corso dell'intero anno via via che vengono ricevuti gli archivi. Alcuni di essi, inoltre, hanno riferimento annuale, ma molti hanno riferimento temporale inferiore all'anno, per esempio semestrale, trimestrale o mensile. Questo rende più complessa l'integrazione delle informazioni riferite ad una specifica unità di analisi, quale un individuo o una impresa. Una volta integrati, gli archivi costituiscono il Sistema Integrato dei Micro dati (SIM) dell'Istituto. Tale sistema alimenta una serie di complessi processi di elaborazione e stima che portano alla creazione dei registri socio-demografici ed economici diffusi dall'Istituto.

I suddetti processi costituiscono un articolato flusso di lavoro che presenta anch'esso diversi vincoli temporali, scadenze di rilascio, prodotti intermedi e finali, provvisori e definitivi. In questo modo ogni anno, e più volte durante l'anno in taluni casi, viene aggiornato il sistema dei registri dell'Istituto. Queste caratteristiche dei due sistemi, quello degli archivi e quello dei registri, ne rendono molto complesso l'utilizzo per la correzione o l'integrazione dei dati censuari, in quanto questi ultimi sono riferiti ad una data prefissata e sono relativi all'intera popolazione. Il riferimento temporale è fondamentale, dunque, per garantire la persistenza della coerenza di tutte le informazioni rilevate per i singoli individui alla data del censimento.

Per poter completare l'insieme degli indirizzi di destinazione dei pendolari è stato, pertanto, necessario un dettagliato esame delle informazioni presenti nei due sistemi per decidere quali fossero idonee da prendere in considerazione. Trattandosi di studenti e di lavoratori, i principali archivi utilizzati sono stati i seguenti: registro delle imprese, registro delle unità locali, registro integrato degli occupati nelle imprese e nelle unità locali, anagrafe degli studenti, archivio degli studenti universitari completo di indirizzo della facoltà alla quale risultano iscritti, archivio del personale scolastico, archivio dei collaboratori domestici e familiari, archivio dei dipendenti pubblici.

L'algoritmo di ricerca dell'indirizzo per ciascun individuo esaminato è stato strutturato a cascata al fine di prelevare gli indirizzi dagli archivi o registri più idonei. Dapprima sono stati selezionati gli archivi e i registri di interesse sulla base del riferimento temporale e dell'area tematica. Poi è stato definito un insieme di soglie e di regole per decidere se fosse opportuno estrarre un indirizzo da una determinata fonte per uno specifico individuo. Infine è stata sviluppata una procedura di ricerca che, per ogni singolo individuo da elaborare, esaminasse le fonti prestabilite nell'ordine fissato decidendo di volta in volta se selezionare l'indirizzo oppure no sulla base delle soglie e delle regole individuate. Nei casi in cui tale algoritmo non fosse riuscito a trovare un indirizzo valido in nessuna delle fonti, l'individuo è stato lasciato con tale informazione nulla. Per esempio, nel caso in cui l'individuo in esame fosse un pendolare per motivi di lavoro, l'algoritmo ricercava innanzitutto nel registro integrato degli occupati con riferimento alla settimana lavorativa nella quale cadeva la data del censimento. Qualora non fosse stato trovato un indirizzo valido, l'algoritmo identificava l'impresa o l'unità locale nella quale l'individuo risultava aver lavorato per la maggior parte dell'anno 2011 e ne estraeva l'indirizzo. Qualora ancora non fosse stato reperito un indirizzo valido, l'algoritmo esplorava prima negli archivi degli studenti con riferimento temporale il più vicino possibile alla data del censimento, al fine di considerare i casi degli studenti lavoratori e poi esaminava altri archivi di occupati.

La ricerca era guidata anche dal settore di impiego oppure dal tipo di studi seguiti dichiarati nel questionario di censimento. Il recupero degli indirizzi dalle fonti alternative al Censimento 2011 è risultato molto efficace, anche se ha richiesto lunghi tempi di elaborazione; è stato, quindi, stabilito di utilizzarle anche per sostituire gli indirizzi di studio e di lavoro di bassa qualità per i quali non fosse stato possibile ottenere la sezione di studio o lavoro tramite la prima fase di normalizzazione e geo-codifica.

Nella Tavola 3 è presente il risultato finale ottenuto dall'integrazione di fonti censuarie con altre fonti, che ha consentito di poter disporre di un indirizzo di studio o lavoro per il 94 percento degli individui pendolari.

TAVOLA 3 – PENDOLARI PER TIPOLOGIA DI FONTE DEGLI INDIRIZZI DI STUDIO O DI LAVORO.

PENDOLARI	VALORI ASSOLUTI	VALORI PERCENTUALI
con indirizzo presente da fonte censuaria	23.424.250	81,0
con indirizzo recuperato da altre fonti	3.786.597	13,0
con indirizzo non disponibile da nessuna fonte	1.594.593	6,0
Totale	**28.805.440**	**100,0**

Fonte: elaborazioni Istat su dati del 15° Censimento generale della popolazione e delle abitazioni 2011

2. La fase di normalizzazione degli indirizzi

La fase di normalizzazione è stata finalizzata, nel caso in esame, solo ad una corretta geo-codifica dell'indirizzo di studio o di lavoro dei pendolari utilizzando il software di normalizzazione Egon[5]. Mediante lo sviluppo di procedure in linguaggio PL/SQL Oracle sono stati individuati tutti gli indirizzi da sottoporre a normalizzazione.

L'informazione principale sull'esito del processo fornisce indicazioni circa la completa normalizzazione dell'indirizzo o la normalizzazione con la presenza di errori non gravi che non pregiudicano il buon esito del processo, restituendo le informazioni su codice provincia, codice comune e sezione di censimento corrispondenti. Se sono presenti errori gravi gli indirizzi non sono normalizzati e presentano possibili errori sul codice provincia, sul codice comune o sulla sezione di censimento (Tavola 4).

TAVOLA 4 – LIVELLI DI NORMALIZZAZIONE DI UN INDIRIZZO TRAMITE IL SOFTWARE EGON.

INDIRIZZO	LIVELLO DI NORMALIZZAZIONE
normalizzato	0
normalizzato con avvertenze	1
non normalizzato per presenza di errori	2
non normalizzato per presenza di errori gravi	3

Se il livello di normalizzazione assume i valori 0 o 1, il processo di normalizzazione ha dato buon esito: gli indirizzi sono stati correttamente normalizzati ed il software restituisce il codice provincia, il codice comune corretti. Se il livello di normalizzazione assume i valori 2 o 3, il processo di normalizzazione non ha dato buon esito: gli indirizzi non sono stati normalizzati e presentano errori sul codice provincia e/o sul codice comune e/o sulla sezione di censimento.

Solo per i record correttamente normalizzati è possibile, ma non sempre, ottenere la sezione di censimento. In alcuni casi, infatti, il software Egon non fornisce la sezione di censimento,

[5] Software di normalizzazione e geo-codifica: http://www.egon.com/it/.

anche se l'indirizzo è stato correttamente riconosciuto, e ciò dipende dalle basi territoriali di riferimento: gli stradari utilizzati dal software Egon potrebbero infatti non avere lo stesso riferimento temporale di quello dei database dell'Istituto che ospitano le sezioni di censimento. Di conseguenza, qualche indirizzo - soprattutto se di nuova istituzione - potrebbe non essere ancora associato ad una sezione.

3. I risultati del processo di geo-codifica degli indirizzi

La fase iniziale di normalizzazione e geo-codifica, dopo un pretrattamento dei campi con lo scopo di eseguire una prima revisione qualitativa (pulitura delle stringhe degli indirizzi da caratteri anomali quali spazi iniziali o finali, caratteri anomali punteggiature ridondanti, ecc.), ha riguardato i 23.424.250 individui di cui si disponeva dell'indirizzo di studio o lavoro direttamente dal questionario del Censimento del 2011.

Questa prima fase ha portato alla normalizzazione del 63 percento degli indirizzi associati a questi individui restituendo, oltre al codice provincia, al codice comune e all'indirizzo di studio o lavoro, anche la sezione di censimento. Il 3 percento di indirizzi è stato normalizzato senza, però, l'attribuzione di una sezione di censimento, mentre il restante 34 percento non è stato normalizzato.

L'insieme degli individui con indirizzi normalizzati ma sezione non valida, con indirizzi non normalizzati o con indirizzi nulli rappresenta il numero di individui che è stato selezionato per l'integrazione con altre fonti amministrative (come dettagliato nel Prospetto 1).

Tramite apposite procedure in ambiente Oracle, il recupero degli indirizzi mancanti o non normalizzati è stato eseguito dapprima per gli studenti pendolari, accedendo in particolare ad un archivio contenente i dati degli istituti di istruzione pubblici e privati, di ogni grado e livello.

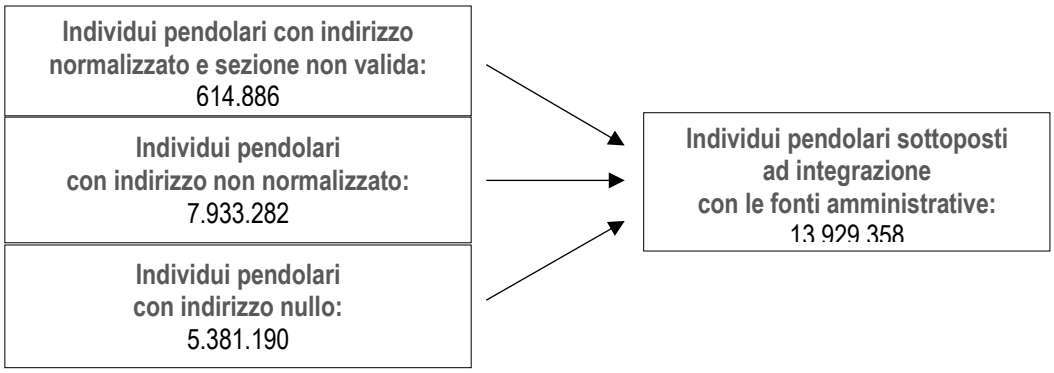
Successivamente si è proceduto con il recupero degli indirizzi dei lavoratori attraverso un metodo più articolato di ricerca in cascata dalle fonti amministrative: è stato effettuato un abbinamento con un primo archivio relativo ai dati sui lavoratori attraverso il quale, utilizzando come chiave di aggancio l'identificativo dell'individuo, la Provincia e il Comune di lavoro, sono stati recuperati il codice dell'impresa e quello dell'unità locale. Tramite il codice dell'individuo, il codice impresa e il codice dell'unità locale così ottenuti si è poi acceduto ad un secondo archivio contenente i dati delle unità locali completi dell'indirizzo.

Nel caso in cui non ci fosse corrispondenza fra l'archivio dei dati censuari e quello delle unità locali, si è acceduto ad un terzo archivio con i dati sulle imprese, tramite le chiavi di aggancio già utilizzate.

E' importante sottolineare che, a fronte di una stessa impresa, possono essere presenti più sedi con diversi indirizzi all'interno di un Comune (ad esempio una banca, un istituto o un ente che ha più sedi sparse nell'ambito di un territorio comunale), dunque, per evitare in questi casi gli inevitabili duplicati di indirizzo di lavoro per uno stesso individuo, è stato seguito il criterio di scelta dell'indirizzo che ha avuto la posizione lavorativa più

rappresentata nel corso dell'anno 2011; in altre parole, è stato selezionato l'indirizzo della sede dove l'individuo ha lavorato per più tempo nell'arco del 2011[6].

PROSPETTO 1 – DETTAGLIO DEGLI INDIRIZZI SOTTOPOSTI A INTEGRAZIONE CON FONTI AMMINISTRATIVE – ANNO 2011



Fonte: elaborazioni Istat su dati del 15° Censimento generale della popolazione e delle abitazioni 2011

Come già effettuato con gli indirizzi provenienti dai dati del Censimento della popolazione, anche per quelli estratti dalle fonti amministrative si è proceduto alla normalizzazione. Infine, nei casi in cui non è stato possibile effettuare un link fra le fonti censuarie e le fonti amministrative, e quindi quando l'indirizzo risultava ancora mancante, sono state elaborate altre procedure Oracle per accedere ad altri archivi dati sui lavoratori domestici e sui dipendenti pubblici, non presi in considerazione precedentemente in quanto molto specialistici. Questi ulteriori accessi hanno permesso di recuperare ulteriori indirizzi ancora mancanti completando la seconda fase di normalizzazione e geo-codifica, i cui risultati sono riportati nella Tavola 5.

TAVOLA 5 – PENDOLARI PER TIPOLOGIA DI NORMALIZZAZIONE DEGLI INDIRIZZI DI STUDIO O LAVORO SUCCESSIVI ALL'INTEGRAZIONE TRA FONTI CENSUARIE E ALTRE FONTI AMMINISTRATIVE.

PENDOLARI	VALORE ASSOLUTO	VALORE PERCENTUALE
con indirizzo normalizzato a cui è stata attribuita la sezione associata	22.055.617	81,0
con indirizzo normalizzato a cui non è stata attribuita la sezione associata	2.015.364	7,4
con indirizzo non normalizzato	3.139.866	11,6
Totale pendolari con indirizzo disponibile sottoposti a normalizzazione	**27.210.847**	**100,0**

Fonte: elaborazioni Istat su dati del 15° Censimento generale della popolazione e delle abitazioni 2011

[6] In questo modo è stato possibile risolvere il problema dei duplicati degli indirizzi per circa 425 mila individui lavoratori.

La percentuale di movimenti pendolari individuabili a livello sub-comunale, in conclusione, è risultata pari al 76,5 percento del totale, quindi piuttosto elevata e tale da consentire di ottenere una soddisfacente matrice di pendolarismo allo stesso livello. Le due tavole successive (Tavola 6 e Tavola 7) mostrano i risultati in maggiore dettaglio suddividendo i totali sulla base della motivazione dello spostamento (studio/lavoro) e della modalità di compilazione del questionario.

TAVOLA 6 – PENDOLARI PER MOTIVO DELLO SPOSTAMENTO, RISULTATO DELLA PROCEDURA DI NORMALIZZAZIONE E GEO-CODIFICA E PROVENIENZA DELL'INDIRIZZO – ANNO 2011 – (VALORI ASSOLUTI)

PENDOLARI	STUDENTI	LAVORATORI	TOTALE
con indirizzo normalizzati	**8.153.908**	**15.917.073**	**24.070.981**
- da Censimento popolazione 2011	*5.133.895*	*10.357.073*	*15.490.968*
- da Archivi SIM	*3.020.013*	*5.560.000*	*8.580.013*
con indirizzo normalizzati con sezione valida	**7.506.017**	**14.549.600**	**22.055.617**
- da Censimento popolazione 2011	*4.986.830*	*9.889.252*	*14.876.082*
- da Archivi SIM	*2.519.187*	*4.660.348*	*7.179.535*
Totale	**9.697.402**	**19.108.038**	**28.805.440**

Fonte: elaborazioni Istat su dati del 15° Censimento generale della popolazione e delle abitazioni 2011

TAVOLA 7 – PENDOLARI PER MOTIVO DELLO SPOSTAMENTO, MODALITÀ DI COMPILAZIONE DEL QUESTIONARIO DEL CENSIMENTO 2011 – ANNO 2011 (VALORI ASSOLUTI)

PENDOLARI	STUDENTI	LAVORATORI	TOTALE
- da compilazione Web	*3.791.004*	*7.133.724*	*10.924.728*
- da compilazione cartacea	*5.906.398*	*11.974.314*	*17.880.712*
TOTALE PENDOLARI	**9.697.402**	**19.108.038**	**28.805.440**
Di cui normalizzati	**8.153.908**	**15.917.073**	**24.070.981**
- da compilazione Web	*3.397.775*	*6.429.420*	*9.827.195*
- da compilazione cartacea	*4.756.133*	*9.487.653*	*14.243.786*
Di cui normalizzati con sezione valida	**7.506.017**	**14.549.600**	**22.055.617**
- da compilazione Web	*3.208.247*	*5.983.665*	*9.191.912*
- da compilazione cartacea	*4.297.770*	*8.565.935*	*12.863.705*

Fonte: elaborazioni Istat su dati del 15° Censimento generale della popolazione e delle abitazioni 2011