istat
working
papers

N.3
2020

La determinazione della *Family home* e il suo ruolo nella stima della dimora abituale: uno studio sperimentale

Sara Casacci, Davide Di Laurea, Pierpaolo Massoli, Gaia Rocchetti e Maria Carla Runci

Istat
Istituto Nazionale
di Statistica

Istat Working Papers

La determinazione della *Family home* e il suo ruolo nella stima della dimora abituale: uno studio sperimentale

N. 3/2020

ISBN 978-88-458-2012-0

La determinazione della *Family home* e il suo ruolo nella stima della dimora abituale: uno studio sperimentale

Sara Casacci, Davide Di Laurea, Pierpaolo Massoli, Gaia Rocchetti e Maria Carla Runci

Sommario

Il concetto di dimora abituale, il luogo in cui un individuo trascorre normalmente il proprio periodo di riposo (Reg. CE N. 1260/2013), rappresenta uno degli elementi fondamentali nella definizione dei criteri di stima della Popolazione per le statistiche sociali. In Italia si è sempre fatta coincidere la residenza anagrafica con la dimora abituale, intesa come luogo dove l'individuo svolge le proprie consuetudini di vita e le proprie relazioni sociali, anche se si reca a lavorare o a svolgere altre attività altrove, purché vi mantenga le proprie relazioni familiari e sociali. Come tutti i registri amministrativi, anche l'anagrafe della popolazione residente può presentare degli errori. Gli stessi responsabili delle anagrafi comunali utilizzano fonti amministrative per verificare la correttezza delle dichiarazioni rese dai cittadini, circa la loro reale dimora abituale nel Comune.

La sempre più ampia disponibilità di dati amministrativi unitamente alla crescente consapevolezza di come poterli usare a fini statistici permette di verificare se è possibile determinare la dimora abituale di una persona a partire da una integrazione mirata di questi dati. È difatti possibile prendere in considerazione tutti i luoghi in cui gli individui e i loro familiari svolgono le proprie attività di lavoro e/o studio per inferire quale sia il luogo dove trascorrono i propri periodi di riposo. La struttura e la composizione delle famiglie sono inferite esogenamente a partire dalle informazioni sui familiari a carico presenti negli archivi fiscali.

Partendo da questi presupposti, nell'ambito delle attività svolte all'interno del gruppo di lavoro ARCHETIPO (ARCHivi E sisTema di Indagini integrate per il Censimento permanente della Popolazione), sono state condotte alcune sperimentazioni volte a identificare la Family home, ovvero il luogo fisico in cui si realizza verosimilmente l'effettiva convivenza degli individui del "nucleo familiare". Il presente lavoro descrive e mette a confronto due approcci per l'identificazione della Family home, applicati a dati amministrativi integrati. Si tratta di un approccio deterministico, basato sull'assegnazione di punteggi alle residenze dei vari membri della Family home e di un approccio basato su un modello grafico probabilistico, interpretato come un problema di knowledge discovery in un dominio di incertezza.

Vengono svolte alcune considerazioni sulla qualità degli esercizi sperimentali, in termini di completezza e di affidabilità dei risultati. La discussione di vantaggi e svantaggi delle metodiche di stima utilizzate conclude il lavoro. In sintesi, la costruzione di un sistema di relazioni fra individui, a partire da informazioni provenienti da più fonti amministrative integrate, mette in luce alcune differenze fra famiglia (o nucleo familiare) che si può potenzialmente ricostruire utilizzando queste fonti e famiglia anagrafica. Inoltre, al netto di alcune peculiarità, i metodi proposti presentano una forte sovrapposizione nella determinazione della Family home.

Parole chiave: Censimento permanente, relazioni parentali, fonti amministrative, residenza abituale, popolazione abitualmente dimorante, approccio deterministico, approccio grafico, ARCHETIPO.

Abstract

Usual residence, intended as the place where any person spends usually her/his own rest period (Reg. CE N. 1260/2013), is a key concept among the operational criteria adopted in the definition of the Population for social statistics. In Italy, Civil Registers' recorded residence has always been made to coincide with the usual residence, understood as a place where the individual carries out her/his own habits of life and her/his social relations, even if she/he works or carries out other activities elsewhere. Like all administrative registers, Civil Registers may present errors.
The increasing supply of administrative data together with the augmenting knowledge on their effective availability for statistical purposes allows for investigating whether there is any room for integrating available administrative data – both for fiscal or other uses – in order to determine the usual residence of a person. It is possible to take into account all the places where that person, as well as her/his relatives, works and/or studies to infer where she/he spends her/his own rest period. The structure and composition of households are exogenously inferred by tax register data.
The authors, as members of the working group ARCHETIPO (ARCHivi E sisTema di Indagini integrate per il Censimento permanente della Popolazione), have focussed on two approaches for the determination of the Family home, namely the usual residence where it is likely the household unit lives. This paper describes and compares the two approaches: a deterministic approach, based on the assignment of scores to the residences of the various members of the Family home and a probabilistic graphical model, interpreted as a problem of knowledge discovery in a domain of uncertainty.
An assessment on the quality of the experimental exercises is provided, mainly focused on completeness and reliability of integrated data. Pros and cons are also outlined.

Keywords: Permanent Census, family relationships, administrative data, place of usual residence, usually resident population, deterministic approach, graphical approach, ARCHETIPO.

Indice

1. Introduzione[1]

A partire dall'agosto 2015 l'Istat ha costituito il gruppo di lavoro inter-dipartimentale denominato ARCHETIPO (ARCHivi E sisTema di Indagini integrate per il Censimento permanente della Popolazione) con il compito di definire il disegno strategico del Censimento permanente della popolazione e delle abitazioni nella prospettiva di una progressiva integrazione di registri di base e indagini statistiche (Falorsi, 2017).

Tra gli obiettivi del gruppo rientrano: lo studio per il registro base della popolazione abitualmente dimorante, a partire dalle anagrafi comunali, con correzione derivante dai segnali da fonti amministrative integrate nel Sistema Integrato dei Microdati[2] (SIM) (Runci et al., 2016), la produzione delle stime annuali della popolazione di ciascun Comune, l'impiego di metodi statistici per la stima della sovracopertura e della sottocopertura della popolazione abitualmente dimorante e l'individuazione di sottopopolazioni critiche (Borrelli et al., 2016).

A tale fine il gruppo è articolato in team tecnici a cui sono associati dei *Working Package* (*WP*). Tra questi, il *WPC1* ha ricevuto il mandato di "individuazione delle fonti amministrative specifiche per arricchire il SIM con archivi capaci di migliorare i segnali di produttività di dimora abituale e la tracciabilità di popolazioni critiche" (deliberazione DGEN n.78 del 12 agosto 2015).

Il *WPC1* ha lavorato, a partire dall'analisi delle definizioni europee correnti, per individuare i criteri operativi per l'identificazione della dimora abituale. I criteri operativi sono sviluppati da un lato con riferimento alla "presenza effettiva" degli individui all'interno del territorio nazionale (sovracopertura e sottocopertura della residenza anagrafica, accertata da ANVIS - ANagrafe VIrtuale Statistica[3]), dall'altro per l'effettiva identificazione del luogo di dimora, per il momento considerato al solo livello comunale.

Questa analisi concettuale preliminare ha riconosciuto nella frequenza di ritorno nella *Family home* uno dei criteri più rilevanti per la determinazione della *usual residence*; di conseguenza, l'identificazione del luogo di dimora di un individuo non può prescindere dall'analisi dalle sue relazioni di tipo affettivo e familiare.

In questo ambito, un sottogruppo del WPC1[4] è stato impegnato nella definizione dei criteri e delle metodologie atte alla identificazione del luogo di dimora abituale, *home*, di gruppi di individui messi in relazione in una *Family*.

Il punto di partenza è rappresentato da un sistema informativo che ha l'obiettivo di ricostruire le relazioni (parentali e non) fra individui identificabili a partire da più fonti amministrative integrate. Tale sistema ha utilizzato, innanzitutto, i dati provenienti dalle dichiarazioni dei familiari a carico (coniuge, figli, altri familiari) desunte dai modelli per la dichiarazione dei redditi UNICO, 730 e 770 con riferimento all'anno fiscale 2012. A partire da tali dichiarazioni sono stati creati dei cluster fiscali, composti da individui messi (direttamente o indirettamente) in relazione tra di loro sulla base delle informazioni contenute nei modelli fiscali. Si è quindi ipotizzato che le informazioni relative alla famiglia fiscale[5], assieme a quelle provenienti da altre fonti, potessero essere utili all'individuazione di legami tra individui appartenenti a più di una famiglia anagrafica. Questa base di dati è in fase di estensione con le informazioni desumibili da altre fonti: anagrafi, nascite, matrimoni, separazioni legali. L'insieme dei cluster e dei legami identificati sono stati oggetto di analisi per l'identifi-

[1] Il lavoro è frutto della collaborazione tra gli autori, tuttavia i paragrafi 2.2, 2.3 e 4.2 sono da attribuire a Sara Casacci, i paragrafi 2.1 e 3.2 a Davide Di Laurea, i paragrafi 3.1 e 3.3 a Pierpaolo Massoli e i paragrafi 4.1 e 4.3 a Gaia Rocchetti. La base dati su cui è stata svolta la sperimentazione è a cura di Maria Carla Runci. Le opinioni espresse in questo lavoro sono esclusivamente degli autori e non dell'Istat.

[2] Il Sistema Integrato dei Microdati è una infrastruttura di base dell'Istat che realizza l'attribuzione di codici identificativi univoci validi nel tempo e per tutte le fonti a disposizione (Di Bella e Ambroselli, 2014; Ambroselli, 2015).

[3] L'Anagrafe Virtuale Statistica assume come popolazione di riferimento la popolazione residente ed è costruita a partire dai micro-dati della popolazione legale al Censimento della popolazione del 2011 e alimentata con i micro-dati di flusso relativi agli eventi della dinamica demografica (Tucci et al., 2014; Corsetti et al., 2018).

[4] Sara Casacci, Davide Di Laurea, Pierpaolo Massoli, Gaia Rocchetti e Maria Carla Runci.

[5] La famiglia fiscale è il nucleo familiare che è possibile identificare e ricostruire attraverso i dati fiscali. A differenza della famiglia anagrafica, costituita dai soggetti appartenenti a un medesimo nucleo familiare e conviventi nella stessa abitazione, la famiglia fiscale rappresenta la famiglia come definita in base alle norme del Testo Unico delle Imposte sui Redditi, e come individuata sulla base delle informazioni e dei dati delle dichiarazioni dei redditi. Essa risulta pertanto costituita dal contribuente dichiarante, dall'eventuale coniuge, dichiarante o meno, e da tutti i familiari fiscalmente a carico, indipendentemente dalla effettiva convivenza nella medesima dimora (Ministero dell'economia e delle finanze, 2010).

cazione della *Family home* (d'ora in poi *FH*). Si è quindi scelto di non assumere la residenza dichiarata dagli individui, rilevata tramite i dati individuali dell'Anagrafe della popolazione residente, come luogo della *usual residence*, ma di verificarla integrando con i dati di altre fonti.

Il documento è organizzato come segue. Nel paragrafo 2 viene presentato il quadro definitorio generale e alcune caratteristiche della base dati utilizzata. Nel paragrafo 3 si descrivono i due approcci proposti, mentre il paragrafo 4 mostra i risultati della sperimentazione condotta.

2. Quadro generale

2.1 Obiettivo e definizioni

I criteri di stima della popolazione in uno specifico ambito territoriale si basano sul concetto di *usual residence* (d'ora in poi *UR*), come definito dal Regolamento CE N. 1260/2013 sulle statistiche demografiche europee[6], identificata come il luogo in cui un individuo trascorre normalmente il proprio periodo di riposo.

Tale concetto è, in ambito italiano, non dissimile da quello di residenza anagrafica contenuta nelle leggi e nel regolamento anagrafico (Istat, 1992). Alla residenza[7] (intesa come iscrizione anagrafica) sono tuttavia legati, oltre all'esercizio di diritti individuali, anche benefici di natura fiscale; per questa ragione alcuni individui possono mettere in atto comportamenti opportunistici, facendo sì che la residenza non coincida con la effettiva dimora[8]. L'utilizzo dei soli dati delle anagrafi comunali (seppur revisionati sulla base, ad esempio, del classico Censimento decennale) non permette, dunque, la corretta identificazione della popolazione abitualmente dimorante.

La definizione di *UR* è ripresa dal Regolamento CE N. 763/2008 sui Censimenti della Popolazione e degli Edifici. Per l'applicazione del concetto di *UR* a casi specifici si utilizzano i criteri operativi definiti dal Regolamento CE N. 1201/2009[9], in cui il luogo di riposo viene sostanzialmente identificato nelle *Family home* (*FH*).

---

[6] All'articolo 2 punto (d) viene scritto: "*'Usual residence' means the place where a person normally spends the daily period of rest, regardless of temporary absences for purposes of recreation, holidays, visits to friends and relatives, business, medical treatment or religious pilgrimage.*
The following persons alone shall be considered to be usual residents of the geographical area in question:
(i) those who have lived in their place of usual residence for a continuous period of at least 12 months before the reference date; or
(ii) those who arrived in their place of usual residence during the 12 months before the reference date with the intention of staying there for at least one year.
Where the circumstances described in point (i) or (ii) cannot be established, 'usual residence' can be taken to mean the place of legal or registered residence." Questa definizione riproduce pedissequamente quanto già contenuto nel Regolamento CE N. 763/2008 sui Censimenti della Popolazione e degli Edifici.

[7] Secondo la normativa italiana, la residenza non è altro che la registrazione in anagrafe di una situazione di fatto che si determina se sono soddisfatte le seguenti condizioni: 1) la dimora abituale nel territorio comunale; 2) la volontà dell'interessato di stabilirvi liberamente la propria residenza. Si possono, quindi, distinguere due elementi nell'ambito del concetto di residenza: 1) un aspetto oggettivo, costituito dalla stabile permanenza in un luogo (o dimora abituale), rilevabile da consuetudini e relazioni e dove si esplica la vita sociale e familiare; 2) un aspetto soggettivo, rappresentato dalla volontà da parte della persona di risiedere proprio in quel luogo. L'elemento soggettivo non può esprimersi come una pura intenzione ma deve essere dimostrato, con riferimento al punto 1), dalle consuetudini di vita e dallo svolgimento delle normali relazioni sociali (Istat, 2010). Tuttavia, la definizione di dimora abituale che si trova nelle fonti normative non si basa su regole rigidamente definite, lasciando margini interpretativi per ogni singolo caso. In questo ambito, la Corte di Giustizia della Comunità europea (sentenza del 12 luglio 2001, sez. VI) ha fornito una definizione più stretta, individuando, oltre alle manifestazioni di vita sociale e di relazioni dell'individuo, elementi ben precisi: una presenza per almeno 185 giorni all'anno, un'abitazione e ogni altro luogo di collegamento con la comunità di riferimento (Istat, 2015).

[8] È opportuno specificare che il presente lavoro prende in analisi una sola tipologia di presunta non corretta collocazione degli individui nei comuni di iscrizione anagrafica (persone che compongono famiglie fiscali diverse dal punto di vista anagrafico). Esistono tuttavia anche situazioni opposte, ad esempio relative a persone che restano nella stessa home, pur sperimentando luoghi di vita e di riposo diversi.

[9] Alla voce relativa al '*Place of usual residence*' si legge: "*In applying the definition of 'usual residence' given in Article 2(d) of Regulation (EC) No 763/2008, Member States shall treat special cases as follows:*
a) Where a person regularly lives in more than one residence during the year, the residence where he/she spends the majority of the year shall be taken as his/her place of usual residence regardless of whether this is located elsewhere within the country or abroad. However, a person who works away from home during the week and who returns to the family home at weekends shall consider the family home to be his/her place of usual residence regardless of whether his/her place of work is elsewhere in the country or abroad.
b) Primary and secondary school pupils and students who are away from home during the school term shall consider their family home to be their place of usual residence regardless of whether they are pursuing their education elsewhere in the country or abroad.
Tertiary students who are away from home while at college or university shall consider their term-time address to be their place of usual residence regardless of whether this is an institution (such as a boarding school) or a private residence and regardless of whether they are pursuing their

Non esiste invece una definizione ufficiale di *FH*. Sono tuttavia definiti i concetti di '*Family nucleus*' e di '*Household'*:

- "Family nucleus is defined in the narrow sense, that is as two or more persons who belong to the same household and who are related as husband and wife, as partners in a registered partnership, as partners in a consensual union, or as parent and child. Thus a family comprises a couple without children, or a couple with one or more children, or a lone parent with one or more children. This family concept limits relationships between children and adults to direct (first-degree) relationships, that is between parents and children […]"[10].
- "Household is defined in terms of shared residence and common arrangements, as comprising either one person living alone or a group of persons, not necessarily related, living at the same address with common house-keeping […]" (Eurostat, 1999).

La *FH* può essere allora definita come il luogo fisico in cui si realizza la convivenza degli individui facenti parte di un nucleo familiare e in cui, indipendentemente dai luoghi in cui i singoli individui esplicano le loro attività, il nucleo nel suo complesso trascorre, normalmente, il periodo di riposo.

L'individuazione della dimora abituale passa, dunque, per il tramite della identificazione:

- del sottoinsieme di relazioni che legano gli individui fra di loro in quanto nucleo;
- del luogo in cui essi esplicano congiuntamente le relazioni materiali e affettive.

In presenza di una struttura di informazioni disponibili basate esclusivamente su dati integrati di tipo amministrativo (assenza di informazioni da indagini statistiche), l'obiettivo è quello di identificare l'insieme dei legami e dei segnali disponibili per allocare una Family in una Home precisa, in presenza di una pluralità di possibili luoghi di dimora.

2.2 Le basi dati impiegate nella sperimentazione

La ricostruzione dei legami viene effettuata attraverso le informazioni desumibili dai quadri dei familiari a carico delle dichiarazioni fiscali, integrate con informazioni relative alle famiglie anagrafiche e ai Comuni di residenza delle Liste Anagrafiche Comunali[11] (LAC) al 1 gennaio 2013. La struttura informativa creata permette di identificare e classificare i legami, diretti e indiretti, di coppie di individui.

Al momento la *Family* viene identificata dai nuclei familiari formati da:

- coppia di coniugi;
- coppia di coniugi con figli (naturali, adottivi, solo di uno dei due partner);
- un genitore single con i figli (naturali, adottivi).

I dati sulle relazioni, comprensivi di informazioni anagrafiche, sono stati integrati alle informazioni relative a luoghi di lavoro e studio desunti dal sistema informativo *Persons&Places* (Garofalo, 2014) e con le distanze tra luoghi di residenza e di lavoro/studio espresse in minuti di percorrenza in automobile da centro a centro dei Comuni considerati[12]. Inoltre, i Comuni considerati sono stati caratterizzati da ulteriori informazioni relative all'ampiezza demografica, alla litoraneità e montanità[13].

La sperimentazione è stata effettuata a partire da un sottoinsieme di informazioni contenute nella base dati delle relazioni familiari predisposta da SIM (Pagliuca e Balistreri, 2018). Al netto di errori e incongruenze, la base dati considerata contiene informazioni su 35.356.653 individui raggruppati in 12.443.009 cluster fiscali, per un totale di 38.355.213 relazioni.

Con riferimento alle relazioni fiscali presenti nella base dati, queste si distinguono in:

- Relazioni definite: sono le relazioni binarie ricavabili direttamente dai quadri dei familiari

education elsewhere in the country or abroad. Exceptionally, where the place of education is within the country, the place of usual residence may be considered to be the Family home. […]".

Seguono indicazioni operative per altre specifiche sotto-popolazioni di interesse.

[10] Regolamento CE N. 1201/2009, alla voce 'Family Status'.

[11] Le Liste Anagrafiche Comunali (LAC) sono archivi che contengono informazioni relative a tutti gli individui residenti nel territorio di ciascun Comune italiano, distinti per famiglia o convivenza (Ceccarelli et al., 2013).

[12] http://www.istat.it/it/archivio/157423

[13] http://www.istat.it/it/files/2015/04/Classificazioni-statistiche.zip

a carico (circa il 68 per cento del totale delle relazioni);
- Relazioni stimate: sono le relazioni derivate (circa il 32 per cento del totale), ricavate dalla considerazione congiunta di più relazioni definite.

La distribuzione dei tipi di relazioni presenti all'interno dei cluster è riportata in Tabella 2.1[14]. In particolare, si osserva che le relazioni che consentono di identificare il nucleo familiare, come precedentemente definito, costituiscono circa l'85 per cento del totale delle relazioni (rispettivamente si hanno il 29,6 per cento di relazioni tra coniugi e il 55,4 per cento di relazioni genitore-figlio).

Considerando il numero di coppie di coniugi all'interno dei cluster, si ha che per 11,26 milioni di cluster, pari a oltre il 90 per cento, è presente una sola coppia, mentre i cluster con due o più coppie sono circa 36,5 mila, pari allo 0,3 per cento. Ciò significa che nella base dati sono presenti cluster "complessi" sui quali è necessario effettuare delle operazioni di scissione dei nuclei, sebbene il numero sia assai limitato.

Dal momento che le relazioni fiscali tra individui possono instaurarsi anche tra individui appartenenti a famiglie anagrafiche diverse (ad esempio, nel caso in cui due coniugi non abbiano dichiarato la stessa residenza anagrafica), è possibile contare il numero di famiglie anagrafiche a cui appartengono gli individui di uno stesso cluster. Si osserva (cfr. Tabella 2.2) che circa l'11,2 per cento dei cluster è costituito da 2 o più famiglie anagrafiche. A questi cluster corrispondono circa 4,2 milioni di individui.

I cluster con più famiglie anagrafiche residenti in due Comuni differenti sono 717.036 (cfr. Tabella 2.3).

Questo è il sottoinsieme sul quale si è concentrata la sperimentazione dato che, al momento, l'obiettivo prefigurato è l'individuazione di una *FH* tra un set di possibili luoghi di dimora. La selezione dei cluster con famiglie residenti in due Comuni diversi (attualmente sono dunque esclusi i cluster con famiglie residenti in tre Comuni o più) è dovuta alla loro maggiore consistenza numerica e alla necessità di riduzione della complessità computazionale. La scelta (eventualmente modificabile) di legare la *FH* a uno dei luoghi di residenza è conservativa: essa è principalmente dovuta al fatto che non sono disponibili informazioni di qualità che possano legare l'individuo a un ulteriore alloggio (ad esempio contratti di affitto, utenze domestiche, ecc.). Una fase successiva della sperimentazione potrà prevedere l'esecuzione dell'algoritmo anche sui cluster i cui membri risiedono nella stessa famiglia anagrafica.

2.3 La caratterizzazione dei cluster individuati

Il sottoinsieme di cluster di interesse per gli approcci da noi proposti comprende i soli cluster composti da più famiglie anagrafiche, residenti in due Comuni diversi (717.036 cluster). All'interno di questi cluster si considerano, inoltre, le sole relazioni definite, al netto delle relazioni che presentano incongruenze tra tipo di relazione e caratteristiche anagrafiche e i soli cluster con meno di 11 componenti[15]. Ne risultano, nel complesso, 716.881 cluster, per un totale di 2.205.575 individui.

La Tabella 2.4 mostra la distribuzione della tipologia di relazioni prese in considerazione e relative ai 716.881 cluster; da essa si evince che la stragrande maggioranza delle relazioni (il 98,7 per cento) è del tipo P/C (genitori-figli) e COU (coniugi).

Per quanto concerne la presenza di coppie e di eventuali figli nei cluster, si osserva che il 24,1 per cento dei cluster è composto da un nucleo di tipo "monogenitore" (ovvero composto da un genitore con uno o più figli), il 27,2 per cento da una coppia senza figli e il 46,6 per cento da una coppia con uno o più figli (cfr. Tabella 2.5). Le rimanenti tipologie di "nucleo" sono di entità trascurabile.

A completamento del quadro di insieme sulla base dati utilizzata, si presenta in Tabella 2.6 la distribuzione dei cluster in base alle combinazioni di segnali sui luoghi lavoro, studio e università dei componenti del cluster. Si osserva che per 130.040 cluster non sono disponibili segnali sui luoghi di lavoro e di studio per nessuno dei componenti. Ne consegue che per 304.856 individui appartenenti a questi cluster non è possibile al momento desumere alcunché sulla *FH*.

[14] Le tabelle relative al paragrafo 2 sono riportate nell'appendice 1.

[15] Quest'ultima limitazione è semplicemente dovuta a problemi computazionali per l'implementazione dell'approccio grafico. Tuttavia va aggiunto che il numero di cluster non presi in considerazione è irrilevante.

3. Due approcci per la determinazione della *Family home*

3.1 Caratteristiche degli algoritmi secondo gli approcci deterministico e a grafo

Data la definizione adottata di *Family home* come il luogo fisico in cui si realizza la convivenza dei membri del nucleo familiare, sono stati sviluppati algoritmi secondo due approcci distinti che, tramite l'utilizzo di relazioni d'ordine, permettono la classificazione ordinata delle residenze nel caso di più luoghi di residenza associati ai membri del medesimo cluster.

Il set informativo su cui sono basate le relazioni d'ordine, è costituito da:

- i dati anagrafici, con riferimento in particolare alla residenza;
- le relazioni di parentela che legano, a due a due, gli individui del cluster;
- la presenza di luoghi di interesse diversi dalla residenza anagrafica, di lavoro o di studio, per ciascuno dei membri del cluster e la loro distanza (in termini di percorrenza) dai luoghi di residenza[16].

La strategia adottata, comune ai due approcci, parte dall'assunto che la *FH* debba essere individuata tra una delle abitazioni di residenza dei membri della *Family*: avendo a disposizione al momento dati completi sull'indirizzo unicamente per le residenze anagrafiche, si è scelto che siano solo loro a costituire l'insieme delle alternative possibili per l'individuazione della *FH*. L'eventuale disponibilità di ulteriori informazioni, come ad esempio la titolarità di contratti di locazione o per utenze di utilities, consentirebbe di estendere ad altri luoghi di interesse la possibilità di essere designati come *FH*.

Vale la pena di sottolineare che il problema della determinazione della *FH* non è stato affrontato come un classico problema di allocazione ottimale. Ci si è limitati a determinare quale fra le residenze dichiarate sia verosimilmente la *Home* del gruppo di individui del cluster che, senza alcuna perdita di generalità, costituiscono una *Family*.

Dato il contesto sperimentale e la necessità di misurare efficacia e impatto degli algoritmi, gli stessi sono stati applicati ai cluster i cui membri siano residenti in due Comuni distinti. In sostanza, le relazioni d'ordine sono utilizzate come indicatore del grado di plausibilità della *Home*, nel caso di multiple residenze fra i membri di un cluster.

3.2 L'algoritmo deterministico

Per implementare un algoritmo di tipo deterministico, operativamente, vengono presi in considerazione tutti i luoghi di lavoro o studio dei membri del cluster. A ciascuna delle residenze anagrafiche viene attribuito un punteggio iniziale, pari al numero totale di luoghi (N_l) di studio e lavoro associati a tutti i componenti del cluster. Il punteggio iniziale associato a ogni residenza viene, eventualmente, decrementato di una penalità (*PEN*) calcolata sulla base di un criterio di lontananza di ciascun luogo di interesse rispetto alle potenziali *FH*. In particolare, il valore della penalità è nullo quando il luogo di lavoro o studio è considerato "vicino" alla residenza anagrafica. È invece pari a 1 quando il luogo di lavoro/studio è "lontano" dalla residenza anagrafica[17].

Per assegnare valore alle penalità è necessario, in primo luogo, determinare dei valori-soglia che discriminino in maniera dicotomica ciò che è da considerare "vicino" piuttosto che "lontano". Nella sperimentazione, si distingue tra diverse tipi di valori-soglia S_f:

- soglia esogenamente determinata;
- soglie di tipo "endogeno": calcolate come momenti della distribuzione dei tempi di percorrenza tra la provincia di residenza anagrafica e i Comuni degli altri luoghi di interesse.

Più in particolare, per ciascuna provincia di residenza la popolazione di riferimento per il calcolo

[16] Nell'algoritmo secondo l'approccio grafico, si utilizza quale parametro anche l'età anagrafica: in caso di compresenza di luoghi di lavoro e studio per taluni individui, l'età è usata per scegliere quale fra le due attività considerare prevalente.

[17] Il concetto di "vicinanza/lontananza" può essere visto come un concetto multidimensionale che include sia componenti "oggettive" sia "soggettive". Ai fini del presente lavoro, il concetto è stato reso operativo attraverso la distanza tra i luoghi in termini di tempi di percorrenza per motivi legati alla disponibilità di dati e per rendere i risultati dello studio facilmente interpretabili e confrontabili.

delle soglie endogene è costituita da tutti gli individui che hanno un luogo di interesse (lavoro o studio, separatamente considerati) in un Comune non appartenente alla provincia in cui si risiede[18]. La scelta di considerare la provincia di residenza come unità territoriale di riferimento per i tempi di percorrenza, e non già il Comune di residenza, è stata fatta per garantire un maggior numero di osservazioni sulla base delle quali computare le soglie endogene. Rimanendo al livello del Comune di residenza, in un numero cospicuo di casi i valori medi e mediani di percorrenza fra i luoghi di residenza vs lavoro o residenza vs studio sarebbero stati computati da celle con frequenze inferiori alle 20 unità[19].

Per ciascun luogo di interesse individuale e luogo di residenza associato al cluster, si ha dunque:

$$PEN_{Al|Rc} = \begin{cases} 0 \text{ se } t_{RcAl} \leq S_j \\ 1 \text{ se } t_{RcAl} > S_j \end{cases} \quad (1)$$

Dove t_{RcAl} è il tempo di percorrenza fra ogni residenza R_C associata al cluster e ogni altro luogo di interesse A_l dei suoi componenti.

Sommando quindi tutte le penalità riferite alla data residenza anagrafica R_C, al variare dei luoghi di lavoro/studio A_l per ciascuno degli i componenti del cluster si ottiene:

$$PEN_{R_C} = \sum_{i=1}^{n} \sum_{A_l=1}^{k} PEN_{Al|Rc} \quad (2)$$

La penalità totale per ogni residenza (PEN_{Rc}) corrisponde, dunque, al numero dei luoghi di interesse dei membri della *Family* che risultano lontani da R_C.

Il punteggio finale per ogni residenza R_C è pari al punteggio iniziale (pari, a sua volta, al numero totale di altri luoghi di interesse per i membri del cluster, N_l) una volta detratta la relativa penalità totale. Ossia:

$$P_{R_C} = N_l - PEN_{R_C} \quad (3)$$

Nella formulazione appena presentata, il punteggio finale dipende, oltre che dal criterio di vicinanza, anche dal numero di luoghi N_l di studio e lavoro. Reputato necessario depurare da questo secondo fattore[20], si è proceduto a una standardizzazione del punteggio, dividendolo per N_l:

$$p_{R_C} = \frac{P_{R_C}}{N_l} = \frac{N_l - PEN_{R_C}}{N_l} \quad (4)$$

18 Gli individui che lavorano o studiano nel medesimo luogo dove risiedono non incidono, dunque, nella determinazione del criterio di vicinanza/lontananza.

19 Il riferimento è al threshold per le frequenze di cella al di sotto del quale il dato non viene considerato sufficientemente affidabile.

20 È evidente che l'operazione di standardizzazione non è necessaria per stabilire quale fra due o più residenze anagrafiche del medesimo cluster sia la più plausibile come FH, essendo i loro punteggi pienamente comparabili. Specularmente, tale operazione è indispensabile per la confrontabilità dei risultati fra cluster diversi per numero di componenti e/o numero di altri luoghi di interesse a essi associati.

Il *range* di variazione del punteggio standardizzato, p_{R_C}, è dunque [0,1]. Il punteggio assume valore massimo (pari a 1), se la sommatoria di penalità dei luoghi associabili a R_C è nulla, ossia quando tutti gli altri luoghi di interesse, se esistono, sono vicini a R_C. All'opposto, sarà minimo (pari a 0) ove risultassero tutti lontani da R_C.

La procedura individua, infine, per ogni cluster uno dei luoghi di residenza come plausibile *FH*: è la residenza con il massimo valore di p_{R_C}, se la differenza rispetto all'alternativa è superiore a uno scarto arbitrariamente determinato Δ_S. È facile verificare che l'algoritmo non è sempre in grado di indicare una *FH* come più 'plausibile'. Situazioni di indeterminatezza sono quelle in cui alle diverse residenze venga attribuito uno stesso valore di p_{R_C}; oppure, come caso di ordine più generale, quando lo scarto Δ_S risulti inferiore a quello minimo prescelto per operare una scelta.

3.2.1 Esempio di funzionamento dell'algoritmo per soglia singola esogena (Sj = 30 minuti)

Si prenda in esame un cluster in cui i due componenti hanno residenze in due Comuni diversi A e B, ma entrambi lavorano nel Comune C. La distanza (in tempo di percorrenza) tra A e C è pari a 10 minuti, mentre tra B e C è di 45 minuti.

Individuo	Luogo di residenza	Luogo di lavoro	Distanza (minuti)
1	A	C	10
2	B	C	45

Essendo la distanza tra A e C inferiore alla soglia indicata (30 minuti) e la distanza tra B e C superiore, si ottengono i seguenti punteggi standardizzati per i due Comuni di residenza (p_A e p_B):

$$p_{R_C=A} = \frac{N_l - PEN_{R_C=A}}{N_l} = \frac{2-0}{2} = 1$$

$$p_{R_C=B} = \frac{N_l - PEN_{R_C=B}}{N_l} = \frac{2-2}{2} = 0$$

$$\Delta_S = \left| p_{R_C=A} - p_{R_C=B} \right| = \left| 1-0 \right| = 1$$

Nell'esempio scelto, lo scarto Δ_S tra le due penalità standardizzate è massimo, ovvero pari a 1: la residenza nel Comune A esibisce il massimo punteggio ottenibile e al contempo la residenza in B il minimo, data la soglia in uso.

3.3 L'approccio grafico

Il secondo fra gli approcci sviluppati si basa su un modello grafico probabilistico (Jump Koller e Friedman, 2009). Nell'adottarlo, il problema della individuazione della dimora abituale di ciascun individuo, e quindi della *FH*, è stato interpretato come un problema di *knowledge discovery* in un dominio di incertezza (Fayyad et al, 1996).

Come già posto in evidenza, le informazioni a disposizione sono attualmente costituite da dati anagrafici, tempi di percorrenza (o distanze) dei luoghi di lavoro/studio e le relazioni (di parentela) che legano, a due a due, gli individui del cluster.

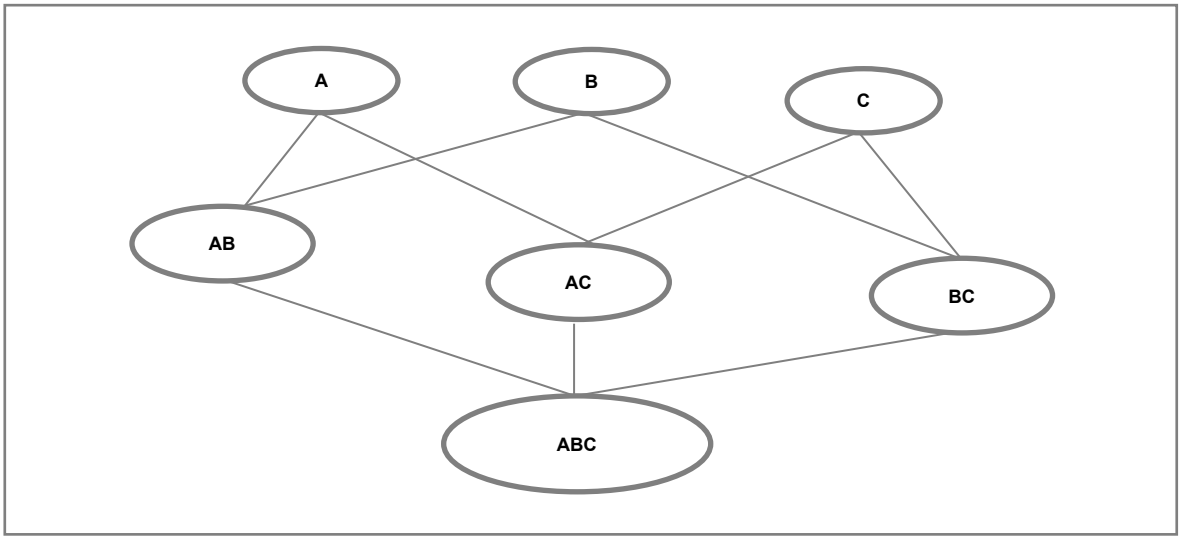
Nell'approccio grafico si costruisce un grafo diretto aciclico (DAG), a partire dagli individui e dalle loro relazioni "a coppia". I nodi del grafo rappresentano delle variabili aleatorie, o eventi, i cui

archi rappresentano la dipendenza causale fra gli eventi stessi.

L'algoritmo sviluppato muove dall'idea che la probabilità che gli individui dimorino assieme nella stessa residenza dipenda, oltre che dalla vicinanza dei luoghi di lavoro/studio e dall'età, anche dal tipo di relazioni che intercorrono fra loro. Il modello grafico è organizzato a livelli. Il primo livello è costituito dagli eventi del tipo: "individuo i-esimo lavora/studia in prossimità della residenza dichiarata R" e si hanno, quindi, tanti nodi quanti sono gli individui. Il secondo livello è costituito da eventi del tipo: "individuo *i-esimo* e *j-esimo* dimorano assieme nella residenza *R*". In questo secondo livello entrano nel modello le relazioni (di parentela) fra gli individui. I livelli successivi al secondo sono costituiti da nodi che caratterizzano gli eventi di combinazioni di più individui dimoranti assieme. Quindi, i nodi del terzo livello sono le combinazioni di tre individui, i nodi del quarto livello sono le combinazioni di quattro individui e così via fino all'ultimo livello il cui unico nodo rappresenta l'evento "tutti gli individui dimorano nella residenza *R*".

In generale dunque, un cluster costituito da *N* individui comporta la costruzione di un grafo costituito da $2^N - 1$ nodi organizzati in *N* livelli. Ogni livello è costituito da un numero di nodi pari a: *N!/(N-k)!k!* nodi. Ogni nodo del livello *k-esimo* ($k \geq 2$) è collegato con i suoi nodi *genitori* del livello *(k-1)-esimo* a seconda della dipendenza causale. Nel modello adottato per le simulazioni, il numero di nodi genitori di ciascun nodo del livello *k-esimo* è pari a *k*.

Figura 3.1 - Modello grafico di una Family; composizione: coppia di coniugi con 1 figlio



Fonte: nostre elaborazioni su archivi e sistemi informativi integrati

In figura 3.1 si riporta il modello grafico di una Family costituita da una coppia di coniugi (*A, B = COU*) con 1 figlio (*A, C = P/C* e *B, C = P/C*).

Il modello di probabilità applicato a tale modello grafico è stato semplificato, rimandando in seguito sviluppi ulteriori. In realtà, si tratta di un modello *ibrido* che comprende variabili continue e discrete. La probabilità che un membro del cluster "preferisca" una residenza rispetto a un'altra è legata alla distanza del proprio luogo di lavoro/studio dalla residenza in esame. Nel caso in cui un individuo studi e lavori, si sceglie l'attività più rilevante basandosi sull'età dell'individuo. Per semplicità si è posto che la probabilità associata ai nodi del primo livello sia data da:

$$p_i = \varphi\left(\frac{d}{S_j}\right) \quad (5)$$

con $i = 1, 2,.., N$, essendo N gli individui del cluster. La funzione $\varphi : \Re \to [0,1]$ è scelta opportunamente in modo che $p = 0{,}5$ quando la distanza d uguaglia la soglia S_j. Si può pensare alla relazione come fosse il grado di appartenenza all'insieme (*fuzzy*) "*stare in prossimità della residenza*". Si è creata dunque una variabile *fuzzy* tanto più vicina a 1 tanto più l'individuo lavora/studia *Vicino* la residenza in esame. Ovviamente, la modalità *Lontano* sarà data dal complemento a 1 della relazione sopra scritta. La probabilità associata ai nodi di secondo livello (combinazioni di due individui) deve tener conto delle relazioni di parentela. Tale probabilità è associata a variabili aleatorie che assumono i valori *Vero* o *Falso* condizionatamente alla vicinanza dei due individui della combinazione alla residenza e alla relazione di parentela che li lega. Con riferimento alla figura 3.1, si può scrivere per la coppia di individui A e B:

$$\begin{aligned} p_{AB} = {} & p\left(AB = Vero \mid A = Vicino, B = Vicino\right) p\left(A = Vicino\right) p\left(B = Vicino\right) \\ & + p\left(AB = Vero \mid A = Vicino, B = \overline{Vicino}\right) p\left(A = Vicino\right) p\left(B = \overline{Vicino}\right) \\ & + p\left(AB = Vero \mid A = \overline{Vicino}, B = Vicino\right) p\left(A = \overline{Vicino}\right) p\left(B = Vicino\right) \\ & + p\left(AB = Vero \mid A = \overline{Vicino}, B = \overline{Vicino}\right) p\left(A = \overline{Vicino}\right) p\left(B = \overline{Vicino}\right) \end{aligned} \quad (6)$$

La probabilità condizionata *p(AB/A,B)* viene stimata da un opportuno *training set* ricavato dai dati di input, differenziando le stime per tipo di relazione di parentela. Sempre con riferimento alla figura 3.1, la probabilità associata all'evento "*tutti gli individui dimorano insieme nella residenza R*" si può calcolare analogamente alla relazione precedente con la seguente relazione:

$$\begin{aligned} p(ABC = Vero) = {} & p\left(AB = Vero\right) p\left(AC = Vero\right) p\left(BC = Vero\right) \\ & + p\left(AB = Falso\right) p\left(AC = Vero\right) p\left(BC = Vero\right) \\ & + p\left(AB = Vero\right) p\left(AC = Falso\right) p\left(BC = Vero\right) \\ & + p\left(AB = Vero\right) p\left(AC = Vero\right) p\left(BC = Falso\right) \end{aligned} \quad (7)$$

Quest'ultima corrisponde alla probabilità che la *Home* della *Family* in esame sia la residenza R dichiarata. Ovviamente, casi di cluster più complessi con più componenti sono deducibili utilizzando questo schema realizzando il modello grafico adeguato.

4. Output e risultati

In questo paragrafo vengono presentati i risultati relativi agli output dei due approcci, "deterministico" e "grafico". Vengono messi a confronto i risultati con le diverse realizzazioni dei valori-soglia, illustrando regolarità e differenze fra le varie istanze rispetto all'incidenza di cluster e di individui per i quali è plausibile la determinazione di una *FH*. L'analisi viene estesa per indagare l'impatto delle procedure sulla differente allocazione territoriale della quota di popolazione la cui *FH* fosse diversa dalla residenza anagrafica. In particolare, ci si focalizza sulle caratteristiche dei Comuni di residenza e di dimora abituale, quando non coincidenti. Si presentano, quindi, i flussi di popolazione, da e per Comuni di dimensione anagrafica diversa, che si osserverebbero qualora gli individui venissero considerati abitualmente dimoranti nel Comune in cui è localizzata la *FH* plausibilmente determinata tramite gli approcci presentati.

Nel dare conto dei risultati, è opportuno ricordare che il sottoinsieme di dati su cui i due algoritmi sono stati applicati è costituito dai cluster composti da due famiglie anagrafiche, residenti in due Comuni diversi e con meno di 11 componenti. Ci si riferisce nel complesso a 716.881 cluster, per un totale di 2.205.575 individui. Vengono inoltre esclusi dal computo totale i 130.040 cluster (pari al 18,14 per cento) per cui non si dispone di alcuna informazione sui luoghi di interesse di almeno uno dei componenti. Al netto di questi ultimi e dei loro componenti, dunque, l'analisi che segue concerne 586.841 cluster e 1.900.719 individui.

4.1 Determinazione della *Family home*

Il primo passo per la valutazione degli approcci consiste nell'analisi del numero di casi (cluster e individui) per cui i punteggi/probabilità attribuiti alle due residenze indicano una prevalenza tra una delle due potenziali *Family home* in esame.

Come già illustrato, entrambi gli algoritmi vengono parametrizzati rispetto a due elementi:

i) il valore-soglia S_j sulla base del quale si modula il criterio di vicinanza/lontananza nel caso deterministico e si determinano le realizzazioni della funzione di probabilità nel modello grafico;

ii) lo scarto tra i punteggi standardizzati (o probabilità nel secondo caso) assegnati a ogni residenza anagrafica, in base al quale si discrimina sulla determinazione o meno di una *FH*.

Per quanto riguarda le soglie di tipo esogeno, i valori presi in considerazione sono stati:

$$S_{j,ESO} = \{15;30;60;90;120\}$$

Le soglie esogene vengono utilizzate esclusivamente nell'algoritmo deterministico.

Le soglie endogene provengono dalla distribuzione del tempo di percorrenza fra Comune di lavoro o studio e provincia di residenza, una volta esclusi chi lavora o studia nel Comune dove risiede. Media e mediana sono i valori di sintesi usati nell'implementazione di entrambi gli approcci.

Infine, per determinare la prevalenza o meno di una residenza anagrafica sull'altra, si è imposto che il valore assoluto della differenza fra punteggi standardizzati sia almeno pari a 0,5[21] ($|\Delta_S| \geq 0,5$).

Per esigenze di brevità, i confronti qui riportati sono stati effettuati considerando come plausibile, e quindi determinata, la *FH* sotto due regimi diversi:

a) regime "medio": la *FH* con punteggio/probabilità più elevato/a è determinata se la differenza tra punteggi/probabilità è almeno pari a 0,5;
b) regime "rigoroso": la *FH* è determinata quando la differenza tra punteggi/probabilità è pari o superiore a 0,9.

[21] Anche in questo caso, si è proceduto a una selezione degli output: vengono riportati nel testo principale solo i risultati relativi a differenze almeno pari a 0,5 e 0,9. In appendice 2, tuttavia, sono state incluse le distribuzioni per tutte le classi di differenze fra punteggi/probabilità.

Tabella 4.1 – Numero e percentuale di cluster e di individui per cui la *FH* risulta determinata o indeterminata, per tipo di soglia e scarto tra i punteggi

Tipo e valori-soglia		Scarto tra punteggi	*FH* determinata				*FH* indeterminata			
			Cluster		Individui		Cluster		Individui	
			Val. ass.	%	Val. ass.	%	Val. ass.	%	Val. ass.	%
APPROCCIO DETERMINISTICO										
Soglie esogene	30 minuti	Δ = 0,5	154.131	26,26	491.388	25,85	432.710	73,74	1.409.331	74,15
		Δ = 0,9	137.631	23,45	425.626	22,39	449.210	76,55	1.475.093	77,61
	60 minuti	Δ = 0,5	97.921	16,69	313.961	16,52	488.920	83,31	1.586.758	83,48
		Δ = 0,9	90.697	15,46	284.364	14,96	496.144	84,54	1.616.355	85,04
Soglie endogene per provincia di residenza	Media	Δ = 0,5	128.610	21,92	415.785	21,88	458.231	78,08	1.484.934	78,12
		Δ = 0,9	110.721	18,87	343.245	18,06	476.120	81,13	1.557.474	81,94
	Mediana	Δ = 0,5	193.830	33,03	622.255	32,74	393.011	66,97	1.278.464	67,26
		Δ = 0,9	160.180	27,30	489.110	25,73	426.661	72,70	1.411.609	74,27
APPROCCIO GRAFICO										
Soglie endogene per provincia di residenza	Media	Δ = 0,5	156.179	26,61	511.436	26,91	430.662	73,39	1.389.283	73,09
		Δ = 0,9	122.222	20,83	400.863	21,09	464.619	79,17	1.499.856	78,91
	Mediana	Δ = 0,5	204.988	34,93	662.010	34,83	381.853	65,07	1.238.709	65,17
		Δ = 0,9	127.958	21,80	420.112	22,10	458.883	78,20	1.480.607	77,90

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati

La percentuale di cluster per i quali gli approcci consentono di individuare una *FH* è variabile al variare della soglia e dello scarto fra punteggi/probabilità (Tabella 4.1). La variabilità osservata è più accentuata nell'approccio deterministico rispetto a quello grafico.

Nel dettaglio, le percentuali più elevate di determinazione della *FH* si hanno utilizzando la soglia endogena mediana: con una differenza almeno pari a 0,5, tramite l'approccio grafico vengono determinate *FH* per il 34,93 per cento dei cluster (204.988 in termini assoluti), mentre è di poco inferiore l'efficacia del deterministico (33,03 per cento, pari a 193.830 cluster).

Di contro le incidenze in assoluto più basse si verificano con la soglia fissa a 60 minuti (rispettivamente 15,46 per cento e 16,69 per cento; nel secondo caso, da regime "medio", è l'unica variante a "determinare" meno del 20 per cento): stando alla poca differenza fra i due regimi, si può concludere che con questa soglia fissa il deterministico discrimina poco ma, quando lo fa, con differenze molto ampie.

Restringendo l'attenzione al solo regime "rigoroso", la maggiore efficacia nel determinare le *FH* è appannaggio del deterministico con soglia media: il 27,30 per cento dei cluster, pari a 160.180 casi. In questo regime, l'approccio grafico "assegna" una *FH* a circa il 21-22 per cento dei casi (a seconda della soglia).

Le considerazioni sin qui svolte rispetto ai cluster restano valide quando si analizzano le percentuali con riferimento all'insieme degli individui. Il numero minimo di persone in cluster con *FH* determinata si ha con il deterministico con soglia fissa a 60 minuti: 284.364 individui in regime di assegnazione "rigoroso". Che aumentano, di poco, fino a 313.961, se si prende in considerazione il regime "medio".

I valori massimi si registrano, invece, usando la soglia mediana e lo scarto da regime "medio": sono 622.255 e 662.010 le persone in cluster con *FH* determinata, rispettivamente con l'algoritmo

deterministico e grafico. Se si passa a una differenza fra i punteggi di 0,9 (sempre con soglia mediana), il numero di individui è stavolta rispettivamente pari a 489.110 e 420.112.

Inoltre, la concordanza tra l'ordine di grandezza della percentuale di cluster e di individui determinati suggerisce che, sebbene la percentuale di unità determinate cambi al variare della soglia, per ogni coppia di metodi confrontati ci sia sempre un gruppo comune di cluster a cui è associabile una *FH*. Tale intuizione è confermata dai dati presentati in appendice 3: in generale, per ogni coppia di metodi si osserva una ampia sovrapposizione tra i cluster a cui è assegnata una *FH*, ma non una inclusione completa (Tabella B in appendice 3). Ad esempio, il metodo a soglia fissa a 30 minuti permette di assegnare la *FH* a 154.131 cluster, quello a soglia fissa a 60 minuti assegna la *FH* a 97.921 cluster. Di questi ultimi, solo 88.381 sono determinati anche da soglia fissa a 30 minuti mentre i rimanenti sono determinati solo da soglia fissa a 60 minuti[22].

Il secondo passo per la valutazione dei due approcci consiste nel conteggio del numero di individui per cui la cui *FH* assegnata coincida o meno con la residenza anagrafica (Tabella 4.2).

Il tipo di approccio e la soglia, anche in questo caso, determinano una certa variabilità sul numero di persone con *FH* diversa dalla propria residenza anagrafica.

Tuttavia, questa variabilità è connessa per lo più alle frequenze assolute. Difatti, l'incidenza relativa di coloro che hanno una *FH* diversa dalla residenza anagrafica, sul totale di casi "determinati", è piuttosto stabile. Con regime medio le percentuali di individui dislocati rispetto alla residenza anagrafica variano tra il 36,80 per cento (deterministico con soglia endogena media) e il 39,40 per cento (approccio grafico con soglia mediana); nel regime rigoroso si osservano percentuali lievemente più elevate: dal 37 per cento (approccio grafico con soglia media) al 39,7 per cento (deterministico con mediana).

Il *range* di variazione è, dunque, ampio solo rispetto ai valori assoluti degli individui con *FH* diversa rispetto al dato anagrafico. Con regime medio, si va dalle 116.879 unità, nel caso del deterministico con soglia fissa a 60 minuti alle 260.978 dell'algoritmo grafico con mediana. Se si restringe il requisito per l'assegnazione della *FH* a una differenza almeno pari a 0,9, il minimo di individui dislocati in un Comune non coincidente con quello di residenza è pari a 108.066 (sempre nel caso del deterministico con soglia uguale a 60 minuti); il massimo è invece da ascrivere al deterministico con soglia mediana: 194.384 le persone dislocate.

[22] Va aggiunto che, quando la *FH* risulta determinata, la prevalente fra le due alternative possibili è la medesima fra i diversi metodi (Tabella C in appendice 3).

Tabella 4.2 – Numero e percentuale di individui per cui la *FH* è diversa o uguale alla residenza anagrafica, per tipo di soglia e scarto tra i punteggi

Tipo e valori-soglia		Scarto tra punteggi	Individui con *FH* determinata			
			FH diversa da residenza		*FH* uguale a residenza	
			Val. ass.	%	Val. ass.	%
			APPROCCIO DETERMINISTICO			
Soglie esogene	30 minuti	Δ = 0,5	184.147	37,50	307.241	62,50
		Δ = 0,9	163.749	38,50	261.877	61,50
	60 minuti	Δ = 0,5	116.879	37,20	197.082	62,80
		Δ = 0,9	108.066	38,00	176.298	62,00
Soglie endogene per provincia di residenza	Media	Δ = 0,5	153.030	36,80	262.755	63,20
		Δ = 0,9	131.316	38,30	211.929	61,70
	Mediana	Δ = 0,5	237.534	38,20	384.721	61,80
		Δ = 0,9	194.384	39,70	294.726	60,30
			APPROCCIO GRAFICO			
Soglie endogene per provincia di residenza	Media	Δ = 0,5	191.990	37,50	319.446	62,50
		Δ = 0,9	148.347	37,00	252.516	63,00
	Mediana	Δ = 0,5	260.978	39,40	401.032	60,60
		Δ = 0,9	162.285	38,60	257.827	61,40

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati

4.2 Differenze tra punteggi/probabilità attribuiti con i due approcci

In tutti i metodi presentati si osservano elevate percentuali di cluster in cui le due residenze ottengono punteggi uguali. Nel dettaglio, relativamente all'approccio deterministico, la percentuale di cluster in cui le due residenze ottengono lo stesso punteggio varia tra il 37,1 per cento ottenuto utilizzando come soglia endogena la mediana provinciale e il 63,2 per cento ottenuto utilizzando la soglia fissa a 60 minuti (un incremento della soglia produce un incremento del numero di cluster in cui le residenze ottengono lo stesso punteggio). Le percentuali relative a differenze tra scarti nulle sono invece inferiori con l'approccio grafico (24,7 per cento con la media provinciale e 6,5 per cento con la mediana provinciale), a causa della parametrizzazione delle relazioni precedentemente illustrata. Tuttavia, l'approccio grafico determina scarti tra le probabilità che, seppur non nulli, sono comunque molto bassi (inferiori a 0,1) per più di un terzo dei cluster in esame. Entrambi gli approcci individuano, quindi, una percentuale analoga di cluster con differenze tra gli scarti molto basse o nulle.

È opportuno segnalare che l'approccio grafico consente, a differenza di quello deterministico, di distinguere i cluster per cui nessuno dei due Comuni di residenza è una verosimile *FH* (le probabilità assegnate alle due residenze sono entrambe nulle): tali casi risultano pari a circa il 3 per cento dei cluster (soglia media provinciale) e al 4,32 per cento (soglia mediana provinciale).

Il numero dei casi in cui la differenza tra i punteggi è molto alta (compresa tra 0,9 e 1) è, nell'approccio deterministico, piuttosto sensibile alla soglia utilizzata (all'aumentare della soglia diminuisce la percentuale di cluster con differenza alta); l'approccio grafico, invece, risulta più stabile e individua circa un 17 per cento di cluster in cui una delle due residenze ha una probabilità prossima o uguale a 1 e l'altra nulla (o pressoché nulla).

È interessante osservare come i diversi approcci e le diverse soglie producano, una volta identificato uno dei due Comuni di residenza come plausibile *FH*, risultati coerenti tra loro: al variare del

metodo e della soglia la *FH* selezionata è, nella maggior parte dei casi, la stessa (cfr. appendice 3).

La Tabella A dell'appendice 2 riporta in dettaglio le distribuzioni degli scarti assoluti tra i punteggi/probabilità assegnati a ognuna delle due residenze anagrafiche relative ai membri dei cluster per entrambi gli approcci, al variare delle soglie selezionate.

Si noti che, come descritto nella parte 2 del documento, sebbene l'idea di fondo dei due approcci sia comune, i due approcci si differenziano rispetto al principio secondo il quale si attribuiscono tali "misure" di plausibilità. Nell'approccio "deterministico" si parte da un punteggio massimo uguale per le due residenze a cui si sottraggono delle penalità per ogni luogo di lavoro/studio "lontano" da queste. Nell'approccio "grafico" si valutano le probabilità di tutte le combinazioni di componenti del cluster vivano insieme nei Comuni di residenza anagrafica dichiarati, a partire dalle probabilità che ogni individuo sia vicino alla residenza in esame perché lavora /studia vicino alla residenza. Questo implica una diversa interpretazione del valore 0 riportato nella Tabella A. Nell'approccio deterministico lo 0 equivale al caso in cui le due residenze ricevono lo stesso punteggio. Questo può accadere nei seguenti casi: a) tutti i luoghi di lavoro e studio sono lontani da entrambe le residenze; b) tutti i luoghi sono vicini alle residenze; c) per le due residenze vi sono combinazioni di luoghi vicini e lontani tali che vengono assegnati uguali punteggi alle due residenze (si vedrà poi come questo cambi nel caso delle tre soglie endogene). Nell'approccio grafico lo 0 indica la situazione in cui le probabilità associate alle residenze sono tra loro uguali. Questo può determinarsi per i motivi a) - c) prima elencati. A differenza del deterministico, inoltre, l'approccio grafico consente di identificare quei casi in cui le probabilità che i membri del cluster vivano insieme in una delle residenze anagrafiche siano nulle per entrambe le residenze. Sono questi i casi, indicati con NONE, per cui nessuno dei due Comuni risulterebbe plausibile come *FH*.

Relativamente alle distribuzioni delle differenze tra punteggi/probabilità, si osserva innanzitutto che per il 18,14 per cento dei cluster considerati, per un totale di 130.040, i due approcci non sono al momento applicabili per mancanza di segnali sui luoghi di lavoro/studio. In tutti i metodi presentati si osservano elevate percentuali di cluster per i quali si ottengono eguali punteggi per le due residenze. Nel dettaglio, tali percentuali sono più elevate per il metodo deterministico rispetto al probabilistico per l'effetto della parametrizzazione delle relazioni.

Nell'ambito dell'approccio deterministico, la soglia fissa a 60 minuti e la soglia endogena basata sulla media provinciale mostrano la maggiore percentuale di differenze assolute nulle.

La Tabella A dell'appendice 2 mostra, inoltre, che la distribuzione degli scarti per l'approccio grafico presenta frequenze più elevate per la classe (0, 0.1) – rispettivamente 32,18 per cento dei cluster quando la soglia è la media provinciale e 36,18 per cento quando la soglia è la mediana provinciale – segnalando che per più di un terzo dei cluster in esame l'approccio attribuisce valori alle probabilità molto vicini tra loro. Considerando tali casi insieme a quelli in cui i punteggi del deterministico basato sulle soglie media e mediana provinciale sono nulli, si osserva che entrambi gli approcci non attribuiscono punteggi/probabilità sostanzialmente diverse alle due residenze per oltre il 50 per cento dei cluster.

I confronti tra le frequenze della parte alta della distribuzione - (0.9, 1] – mostrano che l'approccio deterministico basato sulla mediana provinciale ha la più alta percentuale di cluster in cui una delle due residenze riceve il punteggio massimo a discapito dell'altra che avrà punteggio minimo (22,34 per cento). Frequenze elevate si riscontrano anche per l'approccio deterministico a soglia fissa (19,2. Si osserva che, nell'ambito dell'approccio deterministico, la frequenza degli scarti (0.9, 1] è molto sensibile al variare della soglia, mentre risulta più stabile nell'approccio a grafo.

In sintesi, dall'analisi della distribuzione degli scarti è possibile concludere che:

1) Entrambi i metodi attribuiscono punteggi uguali o molto vicini tra loro per oltre il 50 per cento dei cluster.
2) La distribuzione degli scarti tra i punteggi/probabilità, e di conseguenza l'attribuzione dei punteggi a ogni residenza nei cluster, è sensibile alla soglia prescelta.
3) In particolare, la frequenza dei cluster per i quali una delle due residenze ha un punteggio prossimo o uguale al massimo e l'altra nullo (o pressoché nullo) è variabile al variare della soglia.

4.3 Comuni di residenza anagrafica e di dimora abituale: un confronto

In questo paragrafo vengono analizzate alcune caratteristiche dei Comuni degli individui per i quali gli algoritmi individuano una *FH* diversa dalla residenza anagrafica. L'obiettivo è valutare se vi sia una concentrazione di casi per cui la *FH* è diversa dalla residenza anagrafica in funzione di alcune caratteristiche dei Comuni.

Sulla base delle informazioni desunte dall'*Archivio degli elenchi dei Comuni, codici statistici e denominazioni*[23], i Comuni in esame sono stati classificati come Comune capoluogo o meno; litoraneo o meno; Comune montano, parzialmente montano o non montano.

I Comuni sono stati, inoltre, classificati rispetto all'ampiezza demografica distinguendo tra "micro Comuni" (fino a 5.000 abitanti), "piccoli centri" (tra 5.001 e 15.000 abitanti), "Comuni di cintura" (da 15.001 a 50.000 abitanti), "medi centri urbani" (da 50.001 a 100.000 abitanti) e "grandi città" (oltre 100.000 abitanti).

La Tabella 4.3 è relativa alla distribuzione degli individui con *FH* diversa dalla residenza nei Comuni di residenza, per caratteristiche del Comune. La maggior parte risiede in Comuni non capoluogo, con percentuali che variano dal 71 a quasi l'80 per cento, a seconda delle varianti. Lo stesso vale con riferimento alla caratteristica della litoraneità: tra il 61 e il 70 per cento dei componenti dei cluster la cui *FH* è in un Comune diverso da quello registrato in anagrafe, risiede in Comuni non litoranei. Infine, risiedono in Comuni non montani con percentuali che variano tra il 52 e il 61 per cento; la restante parte vive in Comuni parzialmente montani, 17-21 per cento, o totalmente montani, tra il 23 e il 27 per cento.

Tabella 4.3 – Individui con *FH* diversa dalla residenza, per caratteristiche del Comune di residenza, tipo di soglia e scarto tra i punteggi/probabilità

Tipo e valori-soglia		Scarto tra punteggi	Comuni capoluogo		Comuni litoranei		Comuni montani			Individui con *FH* diversa da residenza
			No	Sì	No	Sì	No	Parziale	Sì	
APPROCCIO DETERMINISTICO										
Soglie esogene	30 minuti	Δ = 0,5	76,85	23,15	64,35	35,65	54,31	18,65	27,04	184.147
		Δ = 0,9	76,7	23,3	63,98	36,02	54,31	18,72	26,97	163.749
	60 minuti	Δ = 0,5	71,07	28,93	61,46	38,54	51,89	21,35	26,75	116.879
		Δ = 0,9	71,12	28,88	61,26	38,74	52,03	21,4	26,57	108.066
Soglie endogene per provincia di residenza	Media	Δ = 0,5	73,97	26,03	66,72	33,28	54,15	19,03	26,82	153.030
		Δ = 0,9	73,55	26,45	66,37	33,63	53,62	19,32	27,06	131.316
	Mediana	Δ = 0,5	79,24	20,76	68,37	31,63	59,18	16,93	23,89	237.534
		Δ = 0,9	78,66	21,34	67,64	32,36	58,56	17,27	24,17	194.384
APPROCCIO GRAFICO										
Soglie endogene per provincia di residenza	Media	Δ = 0,5	75,73	24,27	69,27	30,73	56,07	18,27	25,67	191.990
		Δ = 0,9	74,24	25,76	67,14	32,86	54,57	19,1	26,33	148.347
	Mediana	Δ = 0,5	79,69	20,31	69,81	30,19	60,79	16,67	22,54	260.978
		Δ = 0,9	77,53	22,47	67,76	32,24	59,07	17,64	23,29	162.285

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati

[23] http://www.istat.it/it/strumenti/definizioni-e-classificazioni.

La Tabella 4.4 è invece relativa alla distribuzione dei medesimi individui per caratteristiche del Comune dove verrebbe allocata la *FH* che, per l'appunto, non coincide con quello di residenza. Anche in questo caso la maggior parte delle persone dimorerebbe prevalentemente in Comuni non capoluogo, sebbene con percentuali significativamente inferiori a quanto mostrato per i Comuni di residenza: si va dal 52 al 59 per cento a seconda della combinazione di tipo di algoritmo, valore-soglia e livello della differenza fra punteggi/probabilità. Sorprendentemente simili invece le incidenze relative per i Comuni non litoranei, che ospiterebbero le dimore abituali di queste persone in percentuali che vanno dal 63 al 70 per cento. Si segnala, tuttavia, che la distribuzione della popolazione con *FH* diversa dalla residenza in funzione della litoraneità del Comune di dimora abituale differisce dalla analoga distribuzione calcolata sul totale della popolazione residente in Italia[24]. In particolare, la popolazione con *FH* diversa dalla residenza risulta maggiormente concentrata nei Comuni litoranei rispetto al totale della popolazione.

I Comuni non montani sarebbero rappresentati per il 61-66 per cento, incidenze superiori rispetto alle residenze. Differenze che si ampliano nei casi restanti, con una inversione della rilevanza fra le due tipologie di Comune montano: le persone la cui *FH* è in Comuni parzialmente montani rappresentano tra il 23 e il 28 per cento del totale dei dislocati; Comuni totalmente montani ospiterebbero il rimanente 11-12 per cento. A differenza di quanto emerso rispetto alla litoraneità, la distribuzione degli individui con *FH* diversa dalla residenza per caratteristiche montane del Comune di dimora risulta simile alla distribuzione relativa al totale della popolazione, indicando un'assenza di concentrazione della popolazione in esame nei Comuni montani rispetto ai Comuni non o parzialmente montani.

Tabella 4.4 – Individui con *FH* diversa da residenza, per caratteristiche del Comune di dimora abituale, tipo di soglia e scarto tra i punteggi/probabilità

Tipo e valori-soglia		Scarto tra punteggi	Comuni capoluogo		Comuni litoranei		Comuni montani			Individui con *FH* diversa da residenza
			No	Sì	No	Sì	No	Parziale	Sì	
APPROCCIO DETERMINISTICO										
Soglie esogene	30 minuti	Δ = 0,5	53,35	46,65	63,48	36,52	61,08	27,54	11,38	184.147
		Δ = 0,9	52,39	47,61	62,66	37,34	60,79	28,12	11,09	163.749
	60 minuti	Δ = 0,5	55,67	44,33	65,46	34,54	64,01	24,75	11,24	116.879
		Δ = 0,9	55,36	44,64	65,34	34,66	64,01	24,94	11,05	108.066
Soglie endogene per provincia di residenza	Media	Δ = 0,5	56,91	43,09	68,30	31,70	65,42	22,87	11,71	153.030
		Δ = 0,9	56,12	43,88	67,96	32,04	65,21	23,29	11,50	131.316
	Mediana	Δ = 0,5	53,82	46,18	65,67	34,33	63,71	24,85	11,44	237.534
		Δ = 0,9	52,01	47,99	64,24	35,76	62,54	26,04	11,42	194.384
APPROCCIO GRAFICO										
Soglie endogene per provincia di residenza	Media	Δ = 0,5	59,36	40,64	69,68	30,32	65,60	21,99	12,42	191.990
		Δ = 0,9	57,46	42,54	68,27	31,73	64,94	22,85	12,21	148.347
	Mediana	Δ = 0,5	55,46	44,54	66,21	33,79	64,16	24,26	11,58	260.978
		Δ = 0,9	53,24	46,76	64,15	35,85	62,24	25,67	12,09	162.285

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati

Rispetto alla distribuzione dei Comuni per ampiezza demografica (Tabella 4.5), nel caso delle residenze anagrafiche gli individui risiedono, tra il 69 e il 76 per cento, in Comuni con meno di 50.000 abitanti; fra questi, si registra una prevalenza molto lieve – tra il 26 e il 29 per cento– della dimensione micro, rispetto ai centri piccoli o di cintura, questi ultimi ospitando in media il 22-23 per cento della popolazione analizzata.

[24] Elaborazioni degli autori su dati della rilevazione annuale Istat "Movimento e calcolo della popolazione residente".

La quota di coloro che risiedono in Comuni con oltre 50 mila abitanti si attesta tra il 24 e il 31 per cento; questi casi non sono equamente distribuiti fra grandi città e centri urbani di medie dimensioni, essendo il loro rapporto pari a 2-2,5 a 1.

Significativamente diversa la distribuzione relativa delle stesse persone per ampiezza demografica del Comune della *FH* non corrispondente alla residenza: tra il 48 e il 55 per cento starebbero in Comuni piccoli, fino a 50 mila unità di popolazione (Tabella 4.6). La differenza principale rispetto alla distribuzione per residenza è concentrata nel raggruppamento dei micro Comuni, la cui quota compresa tra 11 e 13 per cento è inferiore alla metà di quanto registrato in precedenza. Sostanzialmente uguale la quota di chi vivrebbe nei Comuni di cintura e solo lievemente inferiore quella dei dimoranti nei piccoli centri.

Specularmente, è di gran lunga superiore l'incidenza di chi vive in Comuni medio-grandi. Anche in questo caso la differenza è particolarmente concentrata: la quota di persone in Comuni di medie dimensione gravita intorno al 11 per cento (con variazioni minime intorno al dato medio), 2,5 punti percentuali in più di quanto emerge per i Comuni di residenza; di contro, le persone la cui dimora abituale è localizzata in Comuni grandi sono il 34-42 per cento del totale, mentre si attestano sul 16-22 per cento nel caso delle residenze anagrafiche.

Tabella 4.5 – Individui con *FH* diversa da residenza, per ampiezza demografica del Comune di residenza, tipo di soglia e scarto tra punteggi/probabilità

Tipo e valori-soglia		Scarto tra punteggi	(0, 5.000] - micro-Comuni	(5.000, 15.000] – piccoli centri	(15.000, 50.000] - Comuni di cintura	(0, 50.000]	(50.000, 100.000] - medi centri urbani	100.000 e oltre - grandi città	Oltre 50.000	Individui con *FH* diversa da residenza
APPROCCIO DETERMINISTICO										
Soglie esogene	30 minuti	Δ = 0,5	29,29	22,3	23,12	74,71	7,99	17,29	25,29	184.147
		Δ = 0,9	29,17	22,2	23,15	74,51	8,03	17,46	25,49	163.749
	60 minuti	Δ = 0,5	25,93	21,47	21,34	68,74	9,24	22,02	31,26	116.879
		Δ = 0,9	25,87	21,56	21,33	68,76	9,2	22,04	31,24	108.066
Soglie endogene per provincia di residenza	Media	Δ = 0,5	29,04	22,31	20,33	71,69	8,35	19,96	28,31	153.030
		Δ = 0,9	28,96	22,14	20,13	71,23	8,5	20,27	28,77	131.316
	Mediana	Δ = 0,5	27,2	25,18	23,57	75,95	7,81	16,24	24,05	237.534
		Δ = 0,9	27,09	24,8	23,59	75,49	7,81	16,7	24,51	194.384
APPROCCIO GRAFICO										
Soglie endogene per provincia di residenza	Media	Δ = 0,5	29,13	23,41	21,06	73,61	7,88	18,51	26,39	191.990
		Δ = 0,9	29,04	22,5	20,51	72,05	8,17	19,77	27,95	148.347
	Mediana	Δ = 0,5	26,09	25,58	24,01	75,69	8,37	15,94	24,31	260.978
		Δ = 0,9	26,04	24,34	23,62	74,01	8,3	17,69	25,99	162.285

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati

Tabella 4.6 – Individui con *FH* diversa da residenza, per ampiezza demografica del Comune di dimora abituale, tipo di soglia e scarto tra punteggi/probabilità

Tipo e valori-soglia		Scarto tra punteggi	(0, 5.000] - micro-Comuni	(5.000, 15.000] – piccoli centri	(15.000, 50.000] - Comuni di cintura	(0, 50.000]	(50.000, 100.000] - medi centri urbani	100.000 e oltre - grandi città	Oltre 50.000	Individui con *FH* diversa da residenza
APPROCCIO DETERMINISTICO										
Soglie esogene	30 minuti	Δ = 0,5	11,62	16,77	20,39	48,77	10,92	40,32	51,23	184.147
		Δ = 0,9	11,38	16,31	20,04	47,73	10,74	41,54	52,27	163.749
	60 minuti	Δ = 0,5	11,69	17,33	21,74	50,75	11,46	37,79	49,25	116.879
		Δ = 0,9	11,56	17,19	21,62	50,38	11,32	38,30	49,62	108.066
Soglie endogene per provincia di residenza	Media	Δ = 0,5	12,30	18,48	21,79	52,57	10,50	36,93	47,43	153.030
		Δ = 0,9	12,03	18,12	21,61	51,77	10,33	37,90	48,23	131.316
	Mediana	Δ = 0,5	11,53	17,00	20,99	49,52	11,03	39,45	50,48	237.534
		Δ = 0,9	11,31	16,33	20,21	47,86	10,70	41,45	52,14	194.384
APPROCCIO GRAFICO										
Soglie endogene per provincia di residenza	Media	Δ = 0,5	13,35	19,66	22,30	55,32	10,34	34,34	44,68	191.990
		Δ = 0,9	12,71	18,63	21,90	53,24	10,50	36,27	46,76	148.347
	Mediana	Δ = 0,5	12,08	17,65	21,53	51,26	10,98	37,76	48,74	260.978
		Δ = 0,9	12,06	16,74	20,35	49,15	10,93	39,91	50,85	162.285

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati

La Tabella 4.7 mostra la composizione percentuale dei flussi di individui, per Comuni di diversa ampiezza demografica, che si avrebbero nel caso in cui si dovesse considerare la *FH* e non la residenza anagrafica come dimora abituale dell'individuo. Negli approcci considerati si osservano flussi più consistenti di individui da e verso Comuni fino a 50.000 abitanti (in media 34 per cento degli individui) e da Comuni fino a 50.000 abitanti a quelli più grandi (in media 40 per cento degli individui). Flussi di minore entità si registrerebbero da Comuni "grandi" a quelli con meno di 50.000 abitanti (in media 15 per cento degli individui).

Tabella 4.7 - Composizione di flussi di individui da e verso Comuni di diversa ampiezza demografica

Soglia	Scarto tra punteggi/ probabilità	Da fino 50.000 – fino 50.000		Da fino 50.000 – A oltre 50.000		Da oltre 50.000 – fino 50.000		Da oltre 50.000 – A oltre 50.000		Totale
		Val. ass.	%	Val. ass.	%	Val. ass.	%	Val. ass.	%	
APPROCCIO DETERMINISTICO										
Soglia fissa 30 minuti	Δ = 0,5	63.358	34,40	74.226	40,30	26.445	14,40	20.118	10,90	184.147
	Δ = 0,9	54.713	33,40	67.299	41,10	23.439	14,30	18.298	11,20	163.749
Soglia fissa 60 minuti	Δ = 0,5	39.534	33,80	40.805	34,90	19.786	16,90	16.754	14,30	116.879
	Δ = 0,9	36.286	33,60	38.020	35,20	18.155	16,80	15.605	14,40	108.066
Soglia endogena media provinciale	Δ = 0,5	55.727	36,40	53.981	35,30	24.726	16,20	18.596	12,20	153.030
	Δ = 0,9	46.908	35,70	46.623	35,50	21.073	16,00	16.712	12,70	131.316
Soglia endogena mediana provinciale	Δ = 0,5	85.089	35,80	95.316	40,10	32.530	13,70	24.599	10,40	237.534
	Δ = 0,9	66.223	34,10	80.516	41,40	26.802	13,80	20.843	10,70	194.384
APPROCCIO GRAFICO										
Soglia endogena media provinciale	Δ = 0,5	75.855	40,00	65.464	34,10	30.355	15,80	20.316	10,60	191.990
	Δ = 0,9	54.765	37,00	52.126	35,10	24.208	16,30	17.248	11,60	148.347
Soglia endogena mediana provinciale	Δ = 0,5	97.481	37,00	100.057	38,30	36.300	13,90	27.140	10,40	260.978
	Δ = 0,9	55.744	34,00	64.361	39,70	24.023	14,80	18.157	11,20	162.285

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati

5. Conclusioni

Il lavoro presenta una sintesi degli approcci logici e metodologici sviluppati al fine di identificare la *Family home*, ovvero il luogo fisico in cui si realizza verosimilmente l'effettiva convivenza degli individui del "nucleo familiare", per quella parte di popolazione che mostra segnali contrastanti dall'unione di dati fiscali e dati di residenza anagrafica, e presenta le prime evidenze quantitative emerse da alcuni test condotti su dati amministrativi integrati.

Si tratta di un lavoro di tipo sperimentale, che si presta a ulteriori approfondimenti e perfezionamenti. I possibili sviluppi futuri sono molteplici, sia di tipo incrementale e di affinamento sia tesi all'implementazione di algoritmi complementari.

I limiti finora riscontrati possono essere riassunti nei seguenti aspetti:

- parzialità e incompletezza della base di dati utilizzata;
- mancanza di dati che si reputano con forte impatto informativo sulla possibilità di stimare i centri di gravitazione individuale e familiare sul territorio, come i contratti di affitto e i consumi elettrici e di gas;
- necessità di estendere e allargare la modellistica utilizzata, sia per trattare cluster più complessi, sia per coprire la casistica delle residenze multiple nel medesimo Comune;
- estensione della modellistica per co-determinare composizione della Family e Home: nel *setting* finora utilizzato, la determinazione del perimetro della Family precede sequenzialmente la scelta della *FH*, senza venire messa in discussione a posteriori;
- tempestività e limiti nell'utilizzo dei dati individuali del Censimento permanente della popolazione (in relazione anche alla nuova normativa sulla *privacy*).

Nonostante i limiti descritti alcuni risultati sono degni di nota:

- la costruzione di un sistema di relazioni fra individui, a partire da informazioni provenienti da più fonti amministrative integrate, mette in evidenza una certa differenza fra famiglia (o nucleo familiare) che si può potenzialmente ricostruire utilizzando queste

fonti e famiglia anagrafica.

- la *FH* rappresenta un criterio fondamentale per identificare la dimora abituale;
- al netto di alcune peculiarità, i metodi proposti presentano una forte sovrapposizione nella determinazione della *FH*, a dimostrazione della robustezza dell'approccio logico considerato. In tutti i metodi presentati si osservano elevate percentuali di cluster in cui le due residenze ottengono punteggi uguali; la distribuzione degli scarti tra i punteggi/probabilità ottenuti tramite i due approcci, e di conseguenza l'attribuzione dei punteggi a ogni residenza nei cluster, è sensibile alla soglia prescelta;
- l'approccio grafico consente, a differenza di quello deterministico, di distinguere i cluster per cui nessuno dei due Comuni di residenza è una verosimile *FH*;
- pur con le parzialità della sperimentazione e l'adozione di criteri fortemente restrittivi, la "riallocazione" degli individui in un Comune diverso dalla residenza anagrafica riguarda al massimo il 40 per cento (circa 260 mila persone) della popolazione considerata;
- l'estensione dei dati, il miglioramento dei modelli, l'utilizzo di nuove fonti può determinare una riallocazione territoriale di individui e di nuclei familiari non del tutto irrilevante per le analisi demografiche e socio-economiche e per la configurazione degli interventi dei *policy-maker*.

Riferimenti bibliografici

Ambroselli, S. 2015. I codici identificativi univoci all'interno del SIM (Sistema Integrato di Microdati). *Istat working papers*, n. 5 (https://www.istat.it/it/archivio/156101).

Borrelli, F., A. Chieppa, S. Di Domenico, G. Gallo, S. Rosati, e V. Tomeo. 2016. Primi risultati della sperimentazione condotta su fonti amministrative capaci di valutare i segnali di dimora abituale in Italia e l'individuazione di sottopopolazioni critiche. *Istat working papers*, n. 23 (https://www.istat.it/it/archivio/210582).

Ceccarelli, C., A. Pezone, e S. Rosati. 2013. L'utilizzo delle Liste Anagrafiche Comunali nella statistica ufficiale. *Rivista Italiana di Economia Demografia e Statistica*, Volume LXVII n. 34, pp. 71-78.

Corsetti, G., S. Prati, V. Tomeo, e E. Tucci. 2018. *A micro-based approach to ensure consistency among administrative sources and to improve population statistics*. Relazione presentata al *49th Scientific meeting of the Italian Statistical Society* (SIS 2018), Palermo, 20-22 giugno.

Di Bella, G., e S. Ambroselli. 2014. *Towards a more efficient system of administrative data management and quality evaluation to support statistics production in Istat*. Relazione presentata *all'European Conference on Quality in Official Statistics* (Q2014), Vienna 2-5 giugno.

Eurostat. 1999. *European Community Household Panel (ECHP) - Selected indicators from the 1995 wave*. Luxembourg: Office for Official Publications of the European Communities.

Falorsi, S. 2017. *Scenario attuale e prospettive per il Censimento permanente della popolazione*. Relazione presentata al Forum PA, Roma 23 maggio.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. 1996. From data mining to knowledge discovery: an overview. In Fayyad, U., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining.* Menlo Park, CA, U.S.: American Association for Artificial Intelligence.

Garofalo, G. 2014. Il Progetto ARCHIMEDE. Obiettivi e risultati sperimentali. *Istat working papers*, n. 9 (https://www.istat.it/it/archivio/140232).

Istat. 1992. Anagrafe della popolazione. Legge e regolamento anagrafico. *Metodi e nome*, serie B – n. 29. Roma.

Istat. 2010. Guida alla vigilanza anagrafica. *Metodi e norme*, n. 48. Roma.

Istat. 2015. *Analisi degli aspetti definitori e della loro declinazione operativa per la popolazione abitualmente dimorante al censimento permanente*, Gruppo di lavoro ARCHETIPO.

Jump Koller, D., and N. Friedman. 2009. *Probabilistic Graphical Models*. *Principles and Techniques*. Cambridge, MA, U.S.: The MIT Press.

Ministero dell'economia e delle finanze. 2010. *La famiglia fiscale*. Statistiche Fiscali – Approfondimenti ottobre 2010.

Pagliuca, D., e A. Balistreri. 2018. *SILF v1.0 Sistema dei legami familiari*. Documento Istat.

Runci, M.C., G. Di Bella, e L. Galiè. 2016. Il sistema di integrazione dei dati amministrativi in Istat. *Istat working papers* n. 18. (https://www.istat.it/it/archivio/193056).

Tucci, E., M. Marsili, e V. Terra Abrami. 2014. *Improving quality of international migration outcomes by incorporating the micro-approach in managing current demographic accounting (MIDEA) and statistical population registers (ANVIS)*. Relazione presentata a *Economic Commission for Europe*, *Conference of European Statisticians*, Chisinau, 10-12 settembre.

Appendice 1 – Tavole relative al paragrafo 2

Tabella 2.1 – Tipologia delle relazioni presenti all'interno dei cluster

Relazione	Totale	%
Genitore-figlio	21.245.538	55,39
Coppia di coniugi	11.335.661	29,55
Fratelli	5.082.199	13,25
Altro familiare	634.505	1,65
Nonno-nipote	57.310	0,15
Totale	**38.355.213**	**100,00**

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati.

Tabella 2.2 – Numero di cluster e numero di individui a essi appartenenti per numero di famiglie anagrafiche da essi messe in relazione

Numero famiglie anagrafiche per cluster	Cluster		Individui	
	N	%	N	%
1	11.108.624	89,28	31.151.948	88,11
2	1.279.581	10,28	3.968.654	11,22
3	51.277	0,41	215.524	0,61
4	3.187	0,03	17.790	0,05
5	292	0	2.174	0,01
6	31	0	306	0
7 e più	17	0	257	0
Totale	**12.443.009**	**100**	**35.356.653**	**100**

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati.

Tabella 2.3 – Numero di cluster per numero di famiglie anagrafiche e numero Comuni di residenza (cluster con individui residenti in due o più famiglie anagrafiche)

Numero famiglie anagrafiche per cluster	N comuni distinti					
	1	2	3	4	5	Totale
2	587.355	692.226	0	0	0	1.279.581
3	16.298	23.430	11.549	0	0	51.277
4	841	1.276	821	249	0	3.187
5	76	95	73	32	16	292
6	12	7	7	3	2	31
7 e più	7	2	3	3	0	15
Totale	**604.589**	**717.036**	**12.453**	**287**	**18**	**1.334.383**

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati.

Tabella 2.4 – Tipologia delle sole relazioni definite presenti all'interno dei cluster

Relazione	Totale	%
Genitore-figlio	1.163.962	66,81
Coppia di coniugi	554.930	31,85
Fratelli	4.166	0,24
Altro familiare	17.999	1,03
Nonno-nipote	1.252	0,07
Totale	**1.742.309**	**100,00**

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati.

Tabella 2.5 - Tipologia dei nuclei all'interno dei cluster (solo relazioni definite)

Tipologia di "nuclei" nei cluster	Cluster		Individui	
	N	%	N	%
Monogenitore	172.641	24,08	520.104	23,58
Coppia senza figli	195.291	27,24	392.912	17,81
Coppia con figli	333.911	46,58	1.220.351	55,33
Due o più coppie senza figli	74	0,01	324	0,01
Due o più coppie con figli	12.582	1,76	66.274	3,00
Nessuna coppia, nessun figlio	2.382	0,33	5.610	0,25
Totale	**716.881**	**100,00**	**2.205.575**	**100,00**

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati.

Tabella 2.6 - Numero di cluster e numero di componenti dei relativi cluster per presenza delle varie tipologie di luoghi

Presenza segnali di lavoro	Presenza segnali di studio	Presenza di segnali di studio universitario	Cluster		Componenti dei cluster	
			Val. ass.	%	Val. ass.	%
0	0	0	130.040	18,14	304.856	13,82
0	0	1	10.663	1,49	34.277	1,55
0	1	0	19.642	2,74	71.388	3,24
0	1	1	1.996	0,28	8.538	0,39
1	0	0	273.974	38,22	751.607	34,08
1	0	1	58.093	8,10	201.709	9,15
1	1	0	198.511	27,69	729.352	33,07
1	1	1	23.962	3,34	103.848	4,71
Totale			**716.881**	**100,00**	**2.205.575**	**100,00**

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati.
Legenda: 0= nessun segnale; 1= almeno un segnale.

Appendice 2 – Tavole relative al paragrafo 4.2

Tabella A – Distribuzione delle differenze assolute tra i punteggi/probabilità ottenute con i vari approcci

	APPROCCIO DETERMINISTICO			
Differenza tra i punteggi	Soglia fissa 30 minuti		Soglia fissa 60 minuti	
	Val. ass.	%	Val. ass.	%
Assenza di segnali di lavoro/studio	130.040	18,14	130.040	18,14
0	365.388	50,97	453.028	63,19
(0, 0.1]	0	0	0	0
(0.1, 0.2]	1.480	0,21	979	0,14
(0.2, 0.3]	6.003	0,84	3.000	0,42
(0.3, 0.4]	25.195	3,51	15.302	2,13
(0.4, 0.5]	34.644	4,83	16.611	2,32
(0.5, 0.6]	1.037	0,14	715	0,1
(0.6, 0.7]	10.786	1,5	4.461	0,62
(0.7, 0.8]	4.603	0,64	2.010	0,28
(0.8, 0.9]	74	0,01	38	0,01
(0.9, 1]	137.631	19,2	90.697	12,65
Differenza tra i punteggi	Soglia endogena media prov.		Soglia endogena mediana prov.	
	Val. ass.	%	Val. ass.	%
Assenza di segnali di lavoro/studio	130.040	18,14	130.040	18,14
0	382.672	53,38	265.995	37,1
(0, 0.1]	0	0	0	0
(0.1, 0.2]	1.619	0,23	2.385	0,33
(0.2, 0.3]	6.610	0,92	11.425	1,59
(0.3, 0.4]	28.855	4,03	43.785	6,11
(0.4, 0.5]	38.475	5,37	69.421	9,68
(0.5, 0.6]	1.262	0,18	1.656	0,23
(0.6, 0.7]	12.132	1,69	23.334	3,25
(0.7, 0.8]	4.423	0,62	8.543	1,19
(0.8, 0.9]	72	0,01	117	0,02
(0.9, 1]	110.721	15,44	160.180	22,34
	APPROCCIO GRAFICO			
Differenza tra le probabilità	Soglia endogena media provinciale		Soglia endogena mediana provinciale	
	Val. ass.	%	Val. ass.	%
Assenza di segnali di lavoro/studio	130.108	18,14	130.040	18,14
0 (*)	177.050	24,7	46.594	6,5
(0, 0.1]	230.679	32,18	259.388	36,18
(0.1, 0.2]	9.120	1,27	34.345	4,79
(0.2, 0.3]	5.628	0,79	17.245	2,41
(0.3, 0.4]	2.972	0,41	10.415	1,45
(0.4, 0.5]	5.277	0,74	14.111	1,97
(0.5, 0.6]	3.181	0,44	13.797	1,92
(0.6, 0.7]	3.079	0,43	11.788	1,64
(0.7, 0.8]	9.174	1,28	21.082	2,94
(0.8, 0.9]	18.523	2,58	30.363	4,24
(0.9, 1]	122.090	17,03	127.713	17,82
(*) NONE (entrambe le residenze hanno prob. nulla di essere *FH*)	21.796	3,04	30.950	4,32
TOT CLUSTER	716.881	100,00	716.881	100,00

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati.

Appendice 3 – Coerenza tra i metodi

Tabella B – Concordanza nella determinazione della *FH* nelle diverse configurazioni

		S2 - Soglia fissa 60			S2 - Deterministico media provinciale		
		Solo S1	comuni	Solo S2	Solo S1	comuni	Solo S2
S1 - Soglia fissa 30	Δ = 0,5	65.750	88.381	9.540	41.618	112.513	16.097
	Δ = 0,9	57.461	80.170	10.527	39.098	98.553	12.188
S1 - Soglia fissa 60	Δ = 0,5				8.682	89.239	39.371
	Δ = 0,9				9.500	81.197	29.524

		S2 - Deterministico mediana provinciale			S2 - Grafico media provinciale			S2 - Grafico mediana provinciale		
		Solo S1	comuni	Solo S2	Solo S1	comuni	Solo S2	Solo S1	comuni	Solo S2
S1 - Soglia fissa 30	Δ = 0,5	25.330	128.801	65.029	42.007	112.124	44.055	38.710	115.421	89.567
	Δ = 0,9	26.337	111.294	48.886	43.082	94.549	27.673	54.030	83.601	44.357
S1 - Soglia fissa 60	Δ = 0,5	20.145	77.776	116.054	13.926	83.995	72.184	28.854	69.067	135.921
	Δ = 0,9	22.547	68.150	92.030	15.950	74.747	47.475	39.028	51.669	76.289
S1 - Deterministico media provinciale	Δ = 0,5	24.358	104.252	89.578	10.274	118.336	37.843	37.241	91.369	113.619
	Δ = 0,9	23.873	86.848	73.332	12.872	97.849	24.373	45.630	65.091	62.867
S1 - Deterministico mediana provinciale	Δ = 0,5				85.705	108.125	48.054	26.307	167.523	37.465
	Δ = 0,9				76.089	84.091	38.131	47.787	112.393	15.565
S1 - Grafico media provinciale	Δ = 0,5							50.264	105.915	99.073
	Δ = 0,9							49.223	72.999	54.959

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati.

Tabella C – Numero di cluster con *FH* assegnata da entrambi i metodi e numero di cluster con *FH* non coincidenti, per coppia di metodi

		Soglia fissa 60		Deterministico media provinciale		Deterministico mediana provinciale		Grafico media provinciale		Grafico mediana provinciale	
		Tot. Cluster comuni	Cluster *FH* diversa	Tot. Cluster comuni	Cluster *FH* diversa	Tot. Cluster comuni	Cluster *FH* diversa	Tot. Cluster comuni	Cluster *FH* diversa	Tot. Cluster comuni	Cluster *FH* diversa
Soglia fissa 30	Δ = 0,5	88.381	0	112.513	20	128.801	62	112.124	106	115.421	67
	Δ = 0,9	80.170	0	98.533	18	111.294	52	94.549	9	83.601	28
Soglia fissa 60	Δ = 0,5			89.239	18	77.776	37	83.995	32	69.067	41
	Δ = 0,9			81.197	14	68.150	33	74.747	8	51.669	24
Deterministico media provinciale	Δ = 0,5					104.252	1	118.336	0	91.369	8
	Δ = 0,9					86.848	1	97.849	0	65.091	1
Deterministico mediana provinciale	Δ = 0,5							108.125	304	167.523	0
	Δ = 0,9							84.091	0	112.393	0
Grafico media provinciale	Δ = 0,5									105.915	38
	Δ = 0,9									72.999	0

Fonte: nostre elaborazioni su archivi e sistemi informativi integrati.

Informazioni per le autrici e per gli autori

La collana è aperta alle autrici e agli autori dell'Istat e del Sistema statistico nazionale e ad altri studiosi che abbiano partecipato ad attività promosse dall'Istat, dal Sistan, da altri Enti di ricerca e dalle Università (convegni, seminari, gruppi di lavoro, ecc.).

Coloro che desiderano pubblicare su questa collana devono sottoporre il proprio contributo al Comitato di redazione degli *Istat working papers*, inviandolo per posta elettronica all'indirizzo: iwp@istat.it.

Il saggio deve essere redatto seguendo gli standard editoriali previsti (disponibili sul sito dell'Istat), corredato di un sommario in Italiano e in Inglese e accompagnato da una dichiarazione di paternità dell'opera.

Per le autrici e gli autori dell'Istat, la sottomissione dei lavori deve essere accompagnata da un'e-mail della/del propria/o referente (Direttora/e, Responsabile di Servizio, etc.), che ne assicura la presa visione.

Per le autrici e gli autori degli altri Enti del Sistan la trasmissione avviene attraverso la/il responsabile dell'Ufficio di statistica, che ne prende visione. Per tutte le altre autrici e gli altri autori, esterni all'Istat e al Sistan, non è necessaria alcuna presa visione.

Per la stesura del testo occorre seguire le indicazioni presenti nel foglio di stile, con le citazioni e i riferimenti bibliografici redatti secondo il protocollo internazionale 'Autore-Data' del *Chicago Manual of Style*.

Attraverso il Comitato di redazione, tutti i lavori saranno sottoposti a un processo di valutazione doppio e anonimo che determinerà la significatività del lavoro per il progresso dell'attività statistica istituzionale.

La pubblicazione sarà disponibile su formato digitale e sarà consultabile on line gratuitamente.

Gli articoli pubblicati impegnano esclusivamente le autrici e gli autori e le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.