



SAPIENZA
UNIVERSITÀ DI ROMA

Facoltà di Ingegneria
Dipartimento di Ingegneria Meccanica e Aerospaziale

Natura e cause del *digital divide* in Italia: uno studio basato sui microdati

Dottorato in Ingegneria Industriale e Gestionale
XXX ciclo

Cosimo Dolente

Ottobre 2017

Ringraziamenti

Ho iniziato questo percorso di dottorato a un'età inusuale, in cui l'avrei dovuto finire, al massimo. Questo mi ha fornito però la possibilità di portarlo avanti in maniera più consapevole, a partire dalla mia esperienza lavorativa e dalla relativa saggezza del trentenne. Ciò nondimeno, non si è trattato di un percorso facile.

Per avermi aiutato a portarlo fino in fondo, i miei più profondi ringraziamenti vanno al Professor Claudio Leporelli, il quale, come già avvenuto durante il percorso delle tesi di laurea triennale e specialistica, mi ha contagiato con la sua inesauribile curiosità e la sua infinita passione per la Ricerca. È ancora mia profonda convinzione, e sempre più dopo questi anni di frequentazione, che è solo grazie al lavoro e alla passione di persone come lui che l'università italiana riesce ancora nel suo importantissimo ruolo di creazione e diffusione di conoscenza, nonostante essa sia così fortemente bistrattata dalla società e dalla classe dirigente del nostro paese.

Ringrazio poi i tanti amici, che mi sono stati vicino e mi hanno aiutato, ognuno a suo modo, durante questo lungo percorso universitario, che pare ora volgere al termine con il suo ultimo grado. Tra di loro annovero anche i miei colleghi di lavoro, che mi hanno sostenuto nelle più difficili giornate di studio e lavoro, sciogliendone la tensione con pazienza e serenità.

Infine, ringrazio ancora una volta e al di sopra di tutti, la mia famiglia. Le parole non bastano a spiegare il suo costante sostegno e il suo immenso affetto, che durano da tutta la vita, senza le quali non avrei mai potuto neanche immaginare di raggiungere traguardi così importanti.

Indice

| | |
|---|----|
| Introduzione ed executive summary | 4 |
| 1 Il ritardo italiano nella domanda di Internet nel confronto internazionale..... | 22 |
| 2 Le determinanti della domanda Internet in letteratura..... | 29 |
| 2.1 Adozione della connessione e uso individuale..... | 29 |
| 2.2 Attività online e differenze tra utenti | 33 |
| 3 Stato della diffusione di Internet in Italia..... | 36 |
| 3.1 Fonti di dati della ricerca | 36 |
| 3.2 Struttura demografica della popolazione italiana..... | 40 |
| 3.3 L'uso di Internet da parte degli individui..... | 43 |
| 3.4 La diffusione delle connessioni Internet tra le famiglie..... | 51 |
| 3.5 Le motivazioni del non uso | 57 |
| 4 Metodologie per l'analisi dei microdati..... | 65 |
| 4.1 Regressione logistica | 66 |
| 4.1.1 Scelta delle variabili indipendenti..... | 71 |
| 4.1.2 Misure della bontà della calibrazione | 74 |
| 4.1.3 Interpretazione del modello calibrato | 75 |
| 4.2 Tecniche di classificazione LAD (Logical Analysis of Data) | 80 |
| 4.2.1 Binarizzazione delle variabili e generazione dei pattern | 82 |
| 4.2.2 Identificazione dei pattern più interessanti per la caratterizzazione di comportamenti socio-economici probabilistici..... | 85 |
| 4.3 Alberi di classificazione | 92 |
| 4.3.1 CHAID..... | 92 |
| 4.3.2 CART..... | 93 |
| 4.3.3 QUEST..... | 97 |

| | | |
|-------|--|-----|
| 4.4 | Analisi per classi latenti (LCA) | 99 |
| 4.4.1 | Formulazione matematica dell'LCA..... | 99 |
| 4.4.2 | Selezione dello specifico modello..... | 100 |
| 5 | Un nuovo modello di digital divide | 102 |
| 5.1 | Digital divide infrastrutturale..... | 102 |
| 5.2 | Digital divide nell'adozione e nell'uso..... | 104 |
| 5.3 | I quattro stadi del digital divide..... | 105 |
| 6 | Analisi delle determinanti della domanda di Internet in Italia | 107 |
| 6.1 | Uso individuale di Internet e adozione familiare della banda larga..... | 107 |
| 6.1.1 | Analisi dell'adozione della tecnologia tramite regressioni logistiche | 109 |
| 6.1.2 | Analisi esplorativa con classificazione SLAD | 116 |
| 6.1.3 | Partizione della popolazione italiana attraverso gli alberi di classificazione | 127 |
| 6.2 | Frequenza d'uso e ricchezza d'uso..... | 132 |
| 6.2.1 | Classi latenti di utenza di Internet | 133 |
| 7 | Considerazioni di policy | 142 |
| 7.1 | Classificazione delle famiglie sulla base dell'adozione della tecnologia..... | 142 |
| 7.2 | Classificazione delle famiglie sulla base delle classi latenti di uso | 149 |
| | Conclusioni | 157 |
| | Bibliografia | 162 |
| | Indice delle figure..... | 172 |
| | Indice delle tabelle | 174 |
| | Appendice | 177 |

Introduzione ed *executive summary*

A partire dal secondo dopoguerra, la visione delle determinanti della crescita economica è iniziata a mutare, includendo il cambiamento tecnologico tra i suoi motori principali, al fianco delle tradizionali forza lavoro e capitale (Solow, 1956). Tuttavia, spesso ci si è concentrati sui miglioramenti incrementali dei processi tecnologici, mentre alcuni studiosi hanno proposto la definizione di General Purpose Technology (GPT) per quelle tecnologie che intervengono trasversalmente nei processi produttivi e su tempi lunghi subiscono successive innovazioni dirompenti, generando una discontinuità netta nei modi di produzione e di consumo (Bresnahan e Trajtenberg, 1992; Helpman e Trajtenberg, 1996; Helpman, 1998; Andergassen, Nardini e Ricottilli, 2017). Le tecnologie dell'informazione e della comunicazione (*Information and Communications Technology – ICT*) sono considerate anch'esse delle GPT (Basu e Fernald, 2008).

In particolare, Internet non può che essere considerato l'esempio più avanzato e pervasivo di ICT, e produce un impatto importante sulla crescita economica (LECG, 2009; Van Reenen *et al.*, 2010; Pélissié Du Rausas *et al.*, 2011; Katz, 2012; Minges, 2016). Le diverse tecnologie di connettività che si sono susseguite nel tempo rappresentano infatti un sostrato che innerva tutta l'economia, consentendo servizi sempre più avanzati, soprattutto in termini applicativi e di *cloud*.

Al fianco dei servizi di sostituzione delle modalità di comunicazione e di informazione già preesistenti l'era digitale, quali le comunicazioni per posta e quelle telefoniche o la diffusione di notizie, le opportunità offerte da Internet consentono soprattutto di sviluppare prodotti e servizi del tutto innovativi, che non sarebbero possibili in assenza della Rete. Si è sviluppato dunque nel tempo un insieme molto grande di nuovi modelli di business che poggiano su Internet, e sono perciò fortemente dipendenti dalla sua diffusione.

Inoltre, un altro importante impatto della presenza pervasiva della Rete è la sostanziale disintermediazione di tutte le transazioni economiche, quali la prenotazione di viaggi e soggiorni, l'*e-banking*, il pagamento di tasse e servizi ecc. Ancora, l'*e-commerce*, reso possibile dalla presenza della Rete, ha prodotto una netta crescita della logistica e un'importante riorganizzazione della relativa catena del valore.

La misura dei vantaggi che Internet apporta all'economia è però fortemente vincolata dallo stato di diffusione dell'uso da parte della popolazione, per diversi motivi.

In primo luogo, affinché tutti questi servizi possano effettivamente prosperare, e nuovi modelli di business possano avere successo, è indispensabile che la base di utenti cresca, di modo da consentire di raggiungere il maggior numero possibile di clienti attraverso Internet.

In secondo, della popolazione fa parte anche la forza lavoro del paese. Una popolazione che presenta un alto tasso di utenti maturi della Rete è in grado anche di convogliare nel tessuto produttivo le competenze digitali acquisite, particolarmente richieste all'interno di imprese che fanno un uso intenso dell'ICT (Bresnahan, Brynjolfsson e Hitt, 2002; Falck, Heimisch e Wiederhold, 2016). Inoltre, lavoratori con competenze digitali elevate possono diventare essi stessi il motore dell'innovazione della propria organizzazione di lavoro, sia essa un'impresa o un ente pubblico.

Infine, l'ammodernamento della Pubblica Amministrazione, ed il conseguente efficientamento e risparmio nei costi, non può prescindere da un uso massiccio dell'ICT (Agid e Team Digitale, 2016) e dal passaggio alla fornitura online dei servizi al cittadino. Ma ciò può avvenire solo se, sull'altro versante, i cittadini possiedono effettivamente le competenze digitali per utilizzare tali servizi, pena una costosa duplicazione online e offline degli stessi, o l'emarginazione di consistenti fette della popolazione, che non usano Internet.

L'effettivo uso di Internet da parte della totalità della popolazione, in maniera autonoma o con l'ausilio di familiari o servizi di supporto, ha dunque un impatto rilevante sulla possibilità di un paese di cogliere davvero le opportunità di sviluppo della propria economia derivanti dall'uso di questa tecnologia.

In questo senso, il nostro paese rappresenta uno dei casi più problematici in Europa, come testimonia il confronto internazionale (capitolo 0), che vede l'Italia agli ultimi posti per adozione di connessioni da parte delle famiglie e uso da parte degli individui. Si tratta infatti di un fenomeno che si estrinseca su due livelli, che si influenzano fortemente tra di loro: quello familiare e quello individuale. Da un lato, la disponibilità di una connessione a banda larga a casa è condizione per un utilizzo completo della Rete da parte dei singoli individui, e può spingere altri membri di una famiglia a diventare utenti. Dall'altro lo stesso uso da parte dei membri, anche in contesti differenti da quello familiare (ad esempio nel luogo di lavoro o di studio, o tramite dispositivi personali *handheld* quali i telefoni cellulari) porta alla costituzione di quella "massa critica" di utenti che spinge la famiglia all'adozione di una connessione a casa (Castelli e Leporelli, 1993).

La distanza tra la quota totale di utenti nella popolazione e gli obiettivi posti dall'Agenda Digitale Europea (Commissione Europea, 2010) risulta poi essere molto elevata. L'agenda digitale fissava già per il 2015 l'obiettivo di un uso regolare di Internet da parte del 75% della popolazione, superiore di ben 22 punti a quello effettivamente stimato per il nostro paese a fine 2014.

L'ipotesi è dunque che esistano effettivamente delle specificità riguardanti il nostro paese, dall'analisi delle quali è necessario partire quando si voglia mettere a punto delle politiche, pubbliche ma anche private (in termini di marketing), atte a favorire l'uso di Internet da parte degli individui e la diffusione delle connessioni a banda larga tra le famiglie, compatibile con gli obiettivi dell'Agenda Digitale Europea.

Il progetto di ricerca descritto in questo documento si propone di studiare la natura di queste specificità, a partire da una rassegna della letteratura, che ha solo marginalmente analizzato il caso italiano (capitolo 2).

La principale fonte di dati del progetto è rappresentata dai microdati (le risposte elementari) dell'indagine "Aspetti della Vita Quotidiana" dell'Istat (nel seguito indicata anche come AVQ), che coinvolge ogni anno circa 50.000 individui, raggruppati in circa 20.000 famiglie. Tali individui sono selezionati in modo da rappresentare la popolazione italiana, attraverso l'uso di specifici pesi, a livello nazionale e regionale. Le risposte raccolte spaziano su diversi campi della quotidianità, consentendo analisi approfondite e multidimensionali (paragrafo 3.1).

Procedendo dunque all'analisi della domanda di Internet in Italia, a partire da tale fonte di dati, emerge un quadro, già dalle statistiche descrittive, di profonde spaccature nell'uso di Internet rispetto a differenti variabili strutturali, quali l'età, il titolo di studio, la condizione professionale e la posizione all'interno della professione, attuale o passata (paragrafo 3.3). Le classi più giovani, soprattutto quelle degli adolescenti e dei giovani adulti, grossomodo fino ai 34 anni, mostrano tassi di utilizzo elevati, che poi degradano sempre più velocemente al crescere dell'età. Un divario simile si riscontra anche rispetto al titolo di studio, dove, come era lecito attendersi, i livelli di istruzione più elevati mostrano tassi di uso molto migliori di quelli base. Situazioni simili sono riscontrabili anche per condizione e posizione

professionale, dove gli attivi e coloro che ricoprono posizioni di intelletto mostrano tassi di utenza molto più elevati.

Riguardo alle famiglie, l'analisi della diffusione della banda larga sulla base delle caratteristiche della Persona di Riferimento della famiglia (figura del database Istat assimilabile nella maggioranza dei casi al più noto "capofamiglia"), mostra come tale fenomeno sia influenzato dalle variabili strutturali in modo del tutto analogo a quanto accade per l'uso individuale (paragrafo 3.4). Ciò pare confermare la stretta connessione tra i due diversi aspetti della domanda di Internet.

L'analisi descrittiva del capitolo 3 si chiude poi con una disamina dei motivi della non adozione dichiarati dalle famiglie, e del non uso dichiarati dagli individui, contenuti anch'essi all'interno dell'indagine AVQ di Istat (paragrafo 3.5). Emerge così una prima indicazione circa la preponderanza di questioni culturali, legate alla conoscenza e alla percezione di utilità di Internet, nonché alle capacità di utilizzo. Viene comunque confermato un certo ruolo anche degli ostacoli di natura economica, almeno per una porzione, consistente, di famiglie e individui.

Le statistiche descrittive consentono dunque di inquadrare il problema del ritardo della domanda Internet in Italia, e realizzare una prima valutazione delle sue determinanti. Tuttavia, al fine di comprenderne a fondo le dinamiche, è opportuno utilizzare metodologie più complesse, in grado di stabilire l'impatto relativo di ciascuna variabile al netto delle altre, e di prendere in considerazione un numero maggiore di variabili. Una descrizione delle metodologie individuate è fornita nel capitolo 4.

Prima di procedere a questo tipo di indagini, contenute nel capitolo 6, viene proposto un nuovo modello di digital divide (capitolo 5). Questo è infatti un concetto ampio, che abbraccia una varietà di questioni legate a diversi tassi di diffusione di Internet in diversi segmenti della popolazione.

Per molto tempo, negli anni della prima diffusione delle reti fisse a banda larga in tutta Europa, il termine *digital divide* faceva riferimento sostanzialmente al solo divario

infrastrutturale, ovvero alla mancanza di copertura di servizi di connettività a banda larga in alcune aree del paese e per la relativa popolazione.

Tuttavia, come mostrato all'interno del capitolo, negli ultimi anni, la copertura a banda larga (con velocità di almeno 2 Mbit/s), sia fissa che mobile, si è fatta capillare. Le questioni infrastrutturali potrebbero comunque tornare rilevanti qualora nuovi servizi Internet rendessero decisiva la disponibilità di connessioni più veloci. A questo proposito, la Strategia per la Banda Ultralarga del governo italiano punta alla realizzazione di nuove infrastrutture per velocità oltre i 30 Mbit/s, e la copertura a 2 Mbit/s di tutto il paese (Presidenza del Consiglio dei Ministri, 2015).

L'attenzione si è nel frattempo spostata sulle dinamiche di adozione della tecnologia, catturata dai due diversi fenomeni fortemente interrelati di cui si è parlato finora, ovvero la sottoscrizione di un servizio ad Internet da parte della famiglia e la decisione di uso individuale.

In molti paesi, tra cui il nostro, in effetti il disinteresse e la mancanza di *skill* digitali, ancor più che problemi economici, fanno sì che una quota rilevante della popolazione non sottoscriva servizi di connessione o comunque non abbia mai utilizzato Internet (cfr. 3.5).

Nei casi in cui questa prima barriera ad un uso maturo di Internet viene superata, si manifestano però nuove forme di digital divide. Queste riguardano le modalità di utilizzo della Rete, in particolar modo la frequenza d'uso e le specifiche attività svolte online, che definiscono la ricchezza effettiva dell'uso (Van Deursen e Van Dijk, 2013, 2015; Van Deursen, Van Dijk e Ten Klooster, 2015).

Le analisi dimostrano che questi due aspetti sono correlati: gli utenti che accedono con più frequenza svolgono anche un numero maggiore di attività online, di crescente complessità e rilevanza per la loro vita economica e sociale. Internet diviene così uno strumento indispensabile non solo per lo svago, ma anche per la fruizione di servizi rilevanti per l'organizzazione della vita quotidiana. Queste analisi consentono infine di evidenziare contesti nei quali non è la maturità dell'utente finale, ma la carenza e la scarsa qualità dei servizi offerti, a limitarne la diffusione. In Italia ciò è evidente soprattutto per i servizi offerti dalle pubbliche amministrazioni.

Riassumendo il quadro appena descritto, si giunge dunque alla citata proposta di una nuova definizione di digital divide in quattro stadi rispetto a:

- 1) Infrastruttura;
- 2) Adozione della tecnologia, a livello individuale e familiare;
- 3) Frequenza d'uso;
- 4) Ricchezza d'uso, in termini di attività svolte online.

Come si diceva, al momento le caratteristiche dell'infrastruttura non sembrano svolgere un ruolo nel limitare l'adozione: i non utenti non sono in grado di apprezzare le eventuali carenze qualitative dei servizi di connettività, e per molti degli utenti le tipologie d'uso non richiedono prestazioni particolarmente elevate. È possibile invece che gli utenti più maturi abbiano imparato ad apprezzare l'utilità di una maggiore velocità di connessione, per una fruizione più ricca di servizi video e *cloud* (Savage e Waldman, 2009).

La prima parte del capitolo 6 si concentra quindi sul secondo stadio del processo sopra schematizzato, attraverso l'analisi delle determinanti dell'adozione, in termini di uso da parte degli individui e di sottoscrizione di una connessione da parte delle famiglie (paragrafo 6.1).

Essa è svolta dapprima attraverso un modello di regressione logistica a più equazioni, tecnica tra le più utilizzate in letteratura, laddove si ha la necessità di verificare e isolare l'impatto di una batteria (ristretta) di variabili indipendenti su di un output dicotomico o categoriale (paragrafo 6.1.1). Il modello, calibrato sul dataset contenente le rilevazioni Istat AVQ dal 2005 al 2012, consente di ottenere una prima stima dell'impatto delle diverse variabili indipendenti, selezionate a partire dall'analisi della letteratura.

Risultano sostanzialmente confermate le intuizioni derivanti dalle analisi descrittive. A livello individuale, l'età risulta essere il fattore che determina le maggiori differenze tra individui nell'uso di Internet. Dopo di essa, i fattori con la maggiore influenza sono il titolo di studio e la disponibilità di connessione a casa. Sono meno pronunciate, ma comunque significative ed in linea con le attese le influenze esercitate da condizione e posizione professionale, ripartizione geografica e dimensione del comune di residenza.

A livello familiare, dall'analisi dei risultati emerge l'importanza della variabile economica. Tra le variabili più importanti per l'adozione di banda larga sembrano esserci anche quelle legate all'età, nelle sue varie declinazioni, in ordine di importanza: età minima, età della Persona di Riferimento della famiglia, età media. L'effetto del numero dei componenti della famiglia è positivo, confermando la necessità di una "massa critica" di utenti.

La dimensione del comune ha un significativo effetto sul processo di adozione del DSL, in cui i comuni centro di area metropolitana sono favoriti e i comuni molto piccoli, fino a 2.000 abitanti, fortemente sfavoriti. Risultato particolarmente interessante, che conferma l'ipotesi per l'assenza di copertura di linea fissa sia supplita, quando possibile, dalle connessioni mobili e *fixed-wireless*, è che per quanto riguarda l'adozione di questi tipi di connessione, i comuni più favoriti, dopo il centro delle aree metropolitane, sono proprio i comuni più piccoli. L'effetto della ripartizione territoriale può invece essere considerato come trascurabile.

Come ci si poteva attendere, l'effetto del passare degli anni, e quindi dell'avanzare del "naturale" processo di diffusione tra la popolazione, è positivo sull'adozione della banda larga.

Infine, l'effetto dell'interazione tra somma delle probabilità di utilizzo dei componenti, derivata dalla prima regressione, la condizione professionale e lo status economico della famiglia è importante per tutte e tre le tipologie di connessione.

Il modello appena esposto ha il principale svantaggio di contenere la circolarità tra uso di Internet e adozione della connessione da parte della famiglia, propria del fenomeno. In statistica, questa circolarità per cui, per l'appunto, alcune variabili (in questo caso due) sono inserite a volte come dipendenti e a volte come indipendenti all'interno del modello, è nota come endogeneità. Ciò introduce un problema metodologico, perché potrebbe portare a inconsistenze nelle stime (Heckman, 1978; Nerlove, 1978; Bascle, 2008). C'è da osservare che il ruolo preponderante nella stima dell'utilizzo è svolto dalle variabili socio-economiche e demografiche, chiaramente esogene, mentre il possesso di una connessione a casa modifica solo marginalmente la propensione all'uso da parte degli utenti meno motivati. D'altra parte, dai dati emerge che, anche se la famiglia dispone di una connessione ad Internet, i componenti svantaggiati, culturalmente o per età, spesso non utilizzano ugualmente il servizio.

La complessità del modello logistico di uso di Internet da parte degli individui ha motivato un'analisi alternativa basata su tecniche di classificazione logica dei dati (paragrafo 6.1.2).

Infatti, sebbene il modello di regressione logistica a più equazioni sia soddisfacente in termini di bontà della calibrazione e interpretazione del fenomeno, un'analisi ulteriore dei dati può aiutare a stabilire se:

- 1) sia possibile ottenere spiegazioni semplici e differenti per diversi segmenti della popolazione;
- 2) esistano regole di classificazione alternative che possano fornire caratterizzazioni equivalenti dei dati osservati;
- 3) variabili aggiuntive rispetto a quelle considerate nel modello di regressione abbiano un qualche ruolo nella spiegazione delle variabili endogene;
- 4) variabili aggiuntive che non abbiano un effetto causale possano comunque migliorare la classificazione e la spiegazione dei fenomeni. Questo può accadere quando un input è associato a una variabile endogena per effetto di fattori latenti che influenzano entrambi.

Si tratta di un'analisi innovativa, svolta insieme agli autori dell'algoritmo SLAD (Bruni e Bianchi, 2015), basata su adattamenti a tale algoritmo, necessari per affrontare l'analisi di comportamenti socio-economici caratterizzati da scelte discrete, intrinsecamente aleatorie.

Il database utilizzato per l'analisi è ancora quello contenente i microdati dell'indagine "Aspetti della Vita Quotidiana" di Istat, relativi all'anno 2012. Da esso, è stato estratto un numero di variabili superiore a quello utilizzato nel modello precedente (dove per spiegare l'uso individuale di Internet si utilizzano 9 variabili), per poter perseguire al meglio gli obiettivi appena esposti. Sono considerate, infatti, 39 variabili, che descrivono una pluralità di possibili determinanti socio-economiche e culturali dell'uso di Internet.

I risultati dell'analisi rivelano come la tecnica innovativa proposta sia in grado di fornire importanti informazioni aggiuntive sul fenomeno in questione. Tali informazioni sono codificate nella forma di pattern, formule logiche di congiunzione di più condizioni sulle variabili di input, che identificano comportamenti omogenei in termini di probabilità di scelta da parte di sottoinsiemi della popolazione definiti con un numero ridotto di variabili.

Sebbene queste procedure automatiche non forniscano un modello interpretativo vero e proprio, dato che non sono ideate per questo scopo, i pattern selezionati supportano letture alternative del fenomeno, attraverso il quale è possibile ottenere importanti spunti conoscitivi sulla segmentazione della popolazione. In particolare, essa consente di selezionare 19 variabili che hanno effettivamente impatto sull'uso di Internet. Tali variabili, opportunamente dicotomizzate sulla base delle informazioni provenienti dai pattern, in termini di impatto

sull'uso di Internet, vengono poi suggerite a un algoritmo di costruzione di un albero di classificazione (paragrafo 6.1.3), che di fatto seleziona solo 7 di esse.

Da questa operazione si ottiene una partizione della popolazione italiana in 8 cluster, identificati con le lettere da A ad H. In particolare, il cluster H è al contempo il più ampio (33,7% della popolazione) e quello con meno utenti di Internet (12,3% del cluster). Esso è composto di individui maturi, con 55 anni o più, che non hanno mai lavorato, o sono stati *white collar* ma non lavorano più, oppure sono o sono stati *blue collar*. In definitiva, si tratta di anziani che non svolgono al momento mansioni impiegatizie, dirigenziali, imprenditoriali o professionali, e dunque sembra essere evidente la loro marginalità sociale e culturale, rispetto al mondo del lavoro. Il loro tasso di utilizzo di Internet è di gran lunga inferiore a quello di qualunque altro cluster della popolazione.

A partire dalle caratteristiche strutturali dei diversi segmenti di popolazione individuati, è possibile dunque studiare politiche di sostegno alla domanda Internet differenziate per cluster, e quindi più mirate a superare gli elementi ostativi che possono essere anche di natura molto diversa (culturale, di competenze, economica).

Nella seconda parte del capitolo 6 (paragrafo 6.2) vengono poi indagati gli ultimi due stadi proposti del digital divide, la frequenza d'uso e le attività online, *proxy* della ricchezza d'uso. Anche in questo caso, come nel caso del duplice aspetto di uso e connessione, si è di fronte a un'influenza reciproca di questi due fenomeni.

L'ipotesi è: chi accede più spesso a Internet apprezza maggiormente tutte le sue potenzialità, e ne fa dunque un uso più ricco. D'altro canto, una persona interessata a utilizzare una pluralità di diversi servizi online, accederà anche ad Internet con maggiore frequenza.

Entrambi i fenomeni possono dunque essere visti come espressione di una variabile latente (nel senso di non misurata) che potrebbe essere definita come "interesse per Internet". L'interesse può infatti portare ad utilizzare una serie di servizi, conoscerne sempre di più e accedere ogni giorno a Internet.

Inoltre, i risultati dell'analisi di regressione logistica sull'uso suggeriscono una stretta relazione fra le attività svolte online e le attività culturali svolte offline. Insieme ai consumi culturali, dunque, l'interesse per Internet potrebbe essere l'espressione di un'altra variabile latente, la

“cultura” personale, o meglio il capitale culturale di una persona (Bourdieu, 1979, 1986). Essa è intesa come la ricchezza e l'ampiezza degli interessi e del know-how di un individuo, derivate dal percorso formativo, dalle preferenze personali, dalle esperienze lavorative e personali, dalle reti sociali.

La verifica di tali associazioni tra variabili d'uso di Internet e attività culturali è oggetto di un'analisi per classi latenti (paragrafo 6.2.1). I comportamenti osservati su di un insieme di variabili manifeste vengono ricondotti all'appartenenza degli individui a una delle classi in cui è ripartita la popolazione. I valori assunti da altre variabili, dette covariate, determinano la probabilità di appartenenza a ciascuna delle classi, modificata dall'osservazione delle variabili manifeste.

È necessario, in questa sede, svolgere analisi separate per utenti e non utenti, poiché per costruzione queste due categorie presentano differenti variabili manifeste. In particolare, mentre per gli utenti possono essere considerate la frequenza di uso e il tipo di attività svolte online, per i non utenti vengono considerate, insieme alle altre variabili *offline*, i motivi del non uso. Entrambi i tentativi di formulazione rispondono comunque all'esigenza di catturare un'unica variabile di classificazione dell'interesse per Internet e, congiuntamente, di altri comportamenti culturali. Le variabili di input considerate ricalcano poi, in gran parte, quelle utilizzate all'interno della regressione logistica (cfr. 6.1.1), per entrambe le categorie di utenza.

Vengono così ottenute 9 classi totali, 6 di utenti e 3 di non utenti. È possibile analizzare queste classi sulla base delle frequenze sulle rispettive variabili manifeste, e sulle variabili indipendenti di segmentazione.

Per quel che riguarda gli utenti, si ottiene così una descrizione interessante dei diversi profili di uso, dato che i comportamenti osservati differiscono fortemente, tra le diverse classi. Gli utenti con profili d'uso più limitati sono quelli appartenenti alla classe 4 ed alla classe 2: è solo una (l'utilizzo di e-mail) l'attività svolta da almeno il 50% degli utenti assegnati a ciascuna di queste classi. Il numero di attività svolte da almeno il 50% degli utenti cresce fino a raggiungere un massimo di 23 per la classe 1. Considerando il numero medio di attività attese per ciascun utente, è in particolare la classe 4 quella che mostra un uso più povero, seguita dalla 2, mentre la classe 3, con 8 attività, è vicina al valore di 9 della classe 6. Anche analizzando le specifiche tipologie di attività svolte, solo alcune attività di comunicazione,

accesso ad informazioni e partecipazione a *social network* sono svolte da almeno la metà del complesso degli utenti, mentre le attività di fruizione di contenuti (prevalentemente multimediali) sono significativamente presenti diffusamente solo nella prima e sesta classe, quelle di utilizzazione di servizi nella prima e nella quinta e quelle connesse al software e al cloud solo nella prima.

La prima classe emerge quindi come quella cui vengono assegnati gli utenti più sofisticati, sia per attività legate al tempo libero che al lavoro o alla fruizione di servizi. La sesta classe condivide con la prima le attività ricreative (in particolare di *gaming*) ma non quelle di utilizzo di servizi. La quinta condivide con la prima l'utilizzo di servizi utili a persone adulte, ma non le attività ricreative multimediali.

Passando alla composizione sulle variabili strutturali, le classi che vedono la presenza preponderante di giovani sono la 1 (più maturi e professionalmente ben inseriti) e la 6 (con una età media inferiore, preponderanza di studenti e inserimento professionale meno brillante per chi ha completato gli studi); la classe 5 contiene prevalentemente *white collar* maturi, spesso persone di riferimento della famiglia, occupati, con alta incidenza di laureati. Le classi 2, 3 e 4 contengono invece gli utenti con uso limitato, e sono meno distinguibili tra loro di quanto non lo fossero in base alle attività svolte, presentando mediamente situazioni sociali più deboli: maggiore incidenza di pensionati e casalinghe, minore tasso di laureati e *white collar*.

È significativo che la classe 1 sia anche quella con il più alto numero atteso di consumi culturali dichiarati, seguita ancora una volta dalla 5 e dalla 6. All'estremo opposto, le classi con uso limitato, 2 e 4. La classe 3 risulta essere intermedia, anche in questo caso. Quest'ultimo risultato pare confermare la fondatezza dell'ipotesi per cui le attività online e offline siano fortemente connesse, e mostra che la suddivisione in classi latenti cattura effettivamente questa correlazione.

Particolarmente interessante è notare che, sulla base dei parametri calibrati dalla regressione, le variabili di input non sembrano discriminare in maniera sostanzialmente diversa tra le classi 2, 3 e 4. Questo pare essere un risultato fondamentale, poiché indica che i comportamenti nell'uso di Internet non sono stimabili in maniera affidabile sulla base delle sole variabili di segmentazione. Ciò implica, in altri termini, che persone della stessa classe di età, che

condividono lo stesso livello di istruzione, condizione professionale e posizione professionale, mostrano spesso differenti profili d'uso della Rete.

Dopo la verifica della stretta relazione tra attività online e offline, si tratta di un altro passo verso la conferma dell'ipotesi per cui ciò che spinge i differenti usi della Rete sia quella cultura personale di cui si parlava pocanzi, non catturabile attraverso l'uso di variabili oggettive.

Per quanto riguarda i non utenti, le 3 classi stimate, in aggiunta alle 6 appena esposte, divengono le classi 7, 8 e 9 della classificazione complessiva degli individui. La classificazione prende in esame non solo le motivazioni del non uso, ma anche, come si diceva, i consumi culturali, l'utilizzo di servizi finanziari e la soddisfazione per la propria situazione economica. Questo permette di affiancare alle motivazioni del mancato utilizzo una caratterizzazione socio-economica degli individui.

La classe 7 (2,8 milioni di persone, circa il 18% del totale dei non utenti considerati) è composta da individui di età media elevata e in maggioranza inattivi. Tuttavia, tra coloro che hanno lavorato o lavorano ancora, circa la metà è *white collar*, e il 52% possiede almeno un diploma di scuola superiore. In questa classe, il livello di soddisfazione per la propria situazione economica è più elevato rispetto alle altre classi di non utenti, i consumi culturali sono più alti e l'utilizzo di carta di credito è maggiore. Nonostante questa collocazione sociale favorevole, il 60% circa di questi soggetti si dichiara non interessata all'uso di Internet, e il 29% dichiara di non saperlo usare. È significativo notare che in questa classe il 48% degli individui vive in una famiglia che dispone di un accesso ad Internet a larga banda fissa.

La classe 8, di gran lunga la più numerosa riguardo i non utenti (10,5 milioni di persone, circa il 64% del totale), comprende soggetti anch'essi caratterizzati da età elevata, ma anche da significativa marginalità sociale. In questo caso, la quota di individui che vive in famiglie con connessione ad Internet a larga banda fissa scende al 32% circa.

Infine, la terza classe (classe 9), comprendente 2,9 milioni di persone (18% dei non utenti considerati), si differenzia dalle due precedenti per l'età media inferiore e per aver dichiarato nel 60% dei casi di non usare Internet per motivi legati al costo della connessione o degli strumenti necessari. Sembra dunque trattarsi di una classe composta in larga parte di potenziali utenti, che non sono (ancora) tali per motivi economici. Cresce la quota di occupati e studenti (50% circa della classe). Il livello di istruzione è intermedio rispetto alle due classi precedenti, con il 31% circa di diplomati o laureati. In linea con le motivazioni espresse per

il mancato uso, il livello di soddisfazione per la propria situazione economica è il più basso tra le tre classi di non utenti. Anche i consumi culturali ne risentono.

Il successivo capitolo 7 riguarda le considerazioni di policy derivanti dalle analisi effettuate, che hanno come ricaduta pratica l'individuazione dei segmenti di utenza e, soprattutto, di non utenza. A partire da questi, è poi possibile immaginare politiche di sostegno alla domanda di Internet che siano efficaci, perché rispondenti alle diverse esigenze e motivazioni di questi differenti cluster.

Una prima classificazione, utile ad indentificare i cluster di famiglie oggetto di politiche di sostegno, si può ottenere a partire dai risultati dei modelli di regressione logistica (cfr. 6.1.1), identificando alcune variabili chiave nel processo di adozione della banda larga, su cui è possibile agire con politiche di sostegno alla domanda:

- la probabilità di avere una situazione economica ottima o adeguata, attraverso incentivi economici;
- la dimensione del comune, nell'ipotesi i problemi di copertura infrastrutturale più gravi riguardino ormai quasi esclusivamente i comuni più piccoli del paese.

Modificando i valori di tali variabili in alcuni casi significativi, nel paragrafo 7.1 si cerca anche di stimare l'effetto di queste possibili *policy* in termini di adozione addizionale di connessioni a banda larga da parte delle famiglie, all'interno di specifici scenari.

All'interno del modello proposto, l'effetto di politiche economiche di sostegno alla domanda può essere simulato attraverso una modifica della probabilità che la situazione economica sia ottima o adeguata, per diverse categorie di famiglie. In particolare, vengono valutati due scenari di intervento economico:

- incentivo a tutti coloro che vivono situazioni di ristrettezze economiche. L'adozione addizionale derivante da questo scenario è stimato essere del 6,5% delle famiglie italiane, quantificabili in 1,4 milioni di famiglie;
- incentivo a tutte le famiglie in ristrettezze economiche che abbiano un componente tra 14 e 24 anni, poiché in tale fascia d'età si trovano le persone più inclini all'utilizzo

di Internet, e, probabilmente, a “contagiare” i propri familiari. L’adozione addizionale complessiva calcolata per questo scenario di intervento è dell’1,9% delle famiglie italiane, quantificabili in 426.325 famiglie con almeno un componente tra i 14 e i 24 anni.

Per stimare poi quale sarebbe l’adozione incrementale delle famiglie dei centri più piccoli nel caso in cui si riuscisse ad ovviare ai problemi residui di digital divide propri di queste realtà, si è condotta una simulazione *ad-hoc*. L’adozione addizionale complessiva calcolata per questo scenario di intervento è di nuovo di circa l’1,9% delle famiglie italiane, quantificabili questa volta in 424.127 famiglie residenti in comuni sotto i 50.000 abitanti.

Infine, è stata condotta una simulazione che riunisce tutti e tre gli interventi menzionati (scenario 4). L’adozione addizionale complessiva calcolata per l’insieme dei 3 scenari di intervento è di circa il 9,2% delle famiglie italiane, quantificabili in ben 2.025.909 famiglie.

Una seconda classificazione interessante deriva dalla partizione delle famiglie effettuata attraverso gli alberi di classificazione. In questo senso, è possibile classificare le famiglie italiane a seconda dell’appartenenza dei loro componenti ai diversi nodi della partizione operata nel paragrafo 6.1.3. In particolare, si propongono le seguenti categorie:

- 1) Famiglie con componenti appartenenti ai nodi che risultano essere in gran parte utenti (quota di utenti maggiore o uguale al 69% della popolazione), segnatamente i nodi A, C, E, G;
- 2) Famiglie non appartenenti alla classe precedente, ma i cui componenti comprendono dei bambini 0-14 anni, che sono potenzialmente utenti entro pochi anni;
- 3) Famiglie non appartenenti alle due classi precedenti, ma i cui componenti ricadono in tutto o in parte nei nodi B, D ed F, che presentano tassi di utenza intorno al 40%;
- 4) Famiglie non appartenenti alle precedenti, i cui componenti sono dunque interamente ricompresi nel nodo H, che presenta un tasso di utenza molto basso, pari al 12,9%.

Le due classi di gran lunga più consistenti sono la prima, che rappresenta le famiglie con componenti appartenenti ai nodi di utenti A, C, E e G, con quote sempre superiori al 50% delle famiglie italiane, anche se in diminuzione, e l’ultima, che rappresenta le famiglie con i

soli componenti del nodo H, in gran parte non utenti, che si attesta intono al 30% delle famiglie italiane.

E proprio questa ultima categoria rappresenta uno “zoccolo duro” di famiglie in stragrande maggioranza composte di non utenti anziani, che vivono in famiglie composte esclusivamente di persone mature e che non hanno (più) accesso a posizioni lavorative che abilitino all’uso di Internet. Si tratta dunque di famiglie che è difficile raggiungere attraverso policy di incentivazione della domanda di Internet, perché risultano fortemente penalizzate nell’uso per motivi strutturali (età elevata, titolo di studio basso, condizione e posizione professionale sfavorevoli).

Ancora, un’altra classificazione utilizzabile per individuare le famiglie oggetto di policy, è basata sulla composizione delle famiglie rispetto alle classi latenti (paragrafo 7.2):

- 1) Famiglie con almeno un utente appartenente alle classi di utenti più evoluti, prevalentemente giovani, cioè la 1 e la 6;
- 2) Famiglie senza utenti delle classi 1 e 6, ma con utenti esperti, prevalentemente maturi, della classe 5;
- 3) Famiglie senza utenti delle classi di uso più completo (1, 5 e 6), ma di cui fanno parte utenti meno sofisticati (appartenenti alle classi 2, 3 e 4);
- 4) Famiglie senza utenti, ma in cui sono presenti non utenti della classe 9, che segnalano spesso motivi economici per il non uso;
- 5) Famiglie non appartenenti alle classi precedenti, in cui sono presenti bambini (0-14 anni);
- 6) Famiglie in cui sono presenti non utenti delle classi 7 o 8, non appartenenti alle classi precedenti;
- 7) Famiglie composte esclusivamente da persone di 75 anni o più;
- 8) Famiglie esclusivamente composte da individui non classificati, cui non appartengano utenti, non utenti della classe 9 o bambini.

Il paragrafo presenta alcune elaborazioni rispetto alla composizione dei diversi gruppi di famiglie (opportunamente riaggregati) e agli strumenti utilizzati per la connessione (separatamente per fissa e mobile).

L'incentivo economico può avere effetto su un segmento specifico di famiglie, quelle della quarta categoria, in cui sono presenti individui la cui adozione è probabilmente limitata da problemi di natura economica. Si tratta di 1,3 milioni di famiglie, cui appartengono circa 3,5 milioni di persone. Qualora, attraverso un piano mirato, si riuscisse ad incentivare l'adozione da una larga parte di tali famiglie, si otterrebbe dunque un incremento significativo nella diffusione di Internet. Tuttavia, una gran parte dei non utenti non sarebbe incentivata da un tale tipo di intervento, poiché le classi 7 e 8 sono composte da individui che in gran parte non citano le problematiche economiche tra i motivi del non uso.

Valutare politiche adatte a superare i problemi relativi a disinteresse, mancanza di capacità e uso superficiale della Rete non è altrettanto agevole, rispetto alla messa a punto di interventi strettamente economici. Si tratta infatti, perlopiù, di una questione culturale.

Una grossa fetta dei non utenti è però in condizioni di inattività: casalinghe e pensionati in primis. Si tratta dei quasi 11 milioni di membri delle 8 milioni di famiglie dell'ultimo gruppo. Raggiungere queste tipologie di persone, che non possono essere intercettate tipicamente in luoghi di aggregazione (se non in casi specifici, quali i centri anziani), rappresenta una sfida importante.

In conclusione, il lavoro esposto nel presente progetto di ricerca fornisce diversi elementi utili per una valutazione del digital divide in Italia.

Innanzitutto, il suggerimento fornito dai dati è che si sia di fronte a un digital divide composito, in cui convergono problematiche di diversa natura, cui non è quindi possibile fornire una risposta unica, in termini di policy. Ciò sia perché il digital divide è per sua natura molteplice, almeno allo stato attuale, sia perché si può suddividere la popolazione in specifici segmenti che mostrano dinamiche di adozione e uso molto diverse, con determinanti differenti.

La complessità del fenomeno è poi sensibilmente incrementata dalla presenza di interazioni forti tra decisioni individuali d'uso e decisioni familiari di sottoscrizione di servizi a banda larga da casa, che creano una circolarità tra le due, comportando anche alcuni problemi metodologici.

Inoltre, vi è un certo numero di input nel database che sono tra di loro fortemente associati, a coppie o gruppi più ampi. Si pensi ad esempio al caso del livello d'istruzione, la condizione professionale e la posizione nella professione. Il ruolo cruciale dell'età, che emerge dalle elaborazioni su uso e adozione, potrebbe essere anche motivato, soprattutto all'interno di alcune classi di età, dalla predominanza di alcune di queste variabili socio-economiche. Ad esempio, nella classe dei più anziani si riscontrano anche livelli di istruzione bassi e alti tassi di inattività e di posizioni professionali manuali.

Passando alle questioni più strettamente di policy, risulta dunque essere largamente minoritaria, sebbene ancora rilevante in termini numerici, la porzione di famiglie italiane che non adottano Internet per motivazioni economiche. E questo costituisce un importante problema, dato che la variabile economica è una delle più facilmente modificabili attraverso politiche di sostegno della domanda, che siano pubbliche attraverso incentivi all'adozione, o private attraverso offerte *ad-boc*. Un possibile ruolo dei fattori di costo è però riscontrabile nell'esistenza di una quota significativa di famiglie che dispongono solo della connessione mobile a banda larga (il 22% nel 2016, secondo i dati Eurostat). Si tratta probabilmente di utenti con ridotta disponibilità a pagare, che accettano un servizio fortemente penalizzato in termini di volume mensile di traffico consentito, a fronte di una spesa mensile sensibilmente inferiore, e di minori vincoli contrattuali.

La gran parte dei divari nell'adozione sono invece dovuti a problemi nella struttura stessa della popolazione italiana, in termini principalmente di età e livelli di istruzione e coinvolgimento attivo nella vita sociale del paese. Inoltre, si evidenzia come ampi segmenti della popolazione sono emarginati socialmente, perché non attivi sul mercato del lavoro o nello studio. Gli effetti negativi delle variabili appena citate sono poi acuite dall'impossibilità di un contagio (almeno nel contesto familiare), in quella parte della popolazione che vive in famiglie in cui non sono presenti utenti, o comunque membri con caratteristiche migliori in una o più delle dimensioni evidenziate.

Queste problematiche sono difficilmente superabili nel breve e medio periodo, in quanto intrinseche della struttura della popolazione, come si diceva. Il quadro è aggravato dalla palese difficoltà nell'intervenire in questo tipo di dinamiche attraverso politiche pubbliche o piani di marketing di imprese private.

Ancora, tra gli utenti stessi emergono differenze importanti, che non sono spiegabili in termini “classici”, attraverso l’uso delle sole variabili di segmentazione. Vengono riscontrati infatti profili di uso di Internet molto diversi, che vanno di pari passo con diversi livelli nei consumi culturali (cinema, teatro, musei ecc.) o nell’uso di servizi finanziari. Queste differenze dovrebbero essere indagati con l’aiuto di dati più fini, ma potrebbero comunque non trovare riscontro in variabili osservabili, dato il ruolo ipotizzato per la cultura personale.

In definitiva, il messaggio che emerge dalle differenti analisi dei dati sulla popolazione italiana sul periodo di 10 anni, tra il 2005 e il 2014, è che non si possa più parlare di digital divide al singolare, ma di una pluralità di situazioni di disuguaglianza su indicatori chiave dell’emarginazione sociale. Da una parte, le caratteristiche strutturali della popolazione italiana dividono nettamente, e, stando alle analisi multi-periodo, sempre di più, gli utenti dai non utenti, e le famiglie connesse da quelle non connesse. Dall’altra, anche all’interno di queste categorie, esistono differenze importanti, legate sostanzialmente alle classi sociali e alla cultura personale dei singoli individui, che non sono spiegabili solo attraverso le “classiche” variabili strutturali. Tutto questo dovrebbe essere ulteriormente indagato con l’ausilio di diverse fonti di dati, col fine ultimo di supportare politiche pubbliche e iniziative private di sostegno alla domanda Internet che siano davvero efficaci, perché basate sulla reale comprensione del fenomeno *dei* digital divide italiani.

1 Il ritardo italiano nella domanda di Internet nel confronto internazionale

La diffusione di Internet in un paese può essere misurata attraverso diversi indicatori, a seconda dello specifico segmento di domanda che si vuole analizzare. Infatti, per le ragioni esposte nell'introduzione, e come si evince anche dai report internazionali (ITU, 2016), l'uso di Internet in un paese può essere letto sul duplice livello familiare e individuale. I principali indicatori sono dunque i seguenti:

- l'adozione di una connessione Internet a banda larga da parte delle famiglie;
- l'uso effettivo di Internet e dei suoi servizi da parte degli individui della popolazione, che appartengono a tali famiglie.

Per ciascuno di essi, lo stato di diffusione può essere misurato come la quota di famiglie che possiede una connessione a banda larga, fissa o mobile, la quota di individui della popolazione che usa Internet con regolarità e la quota di imprese che possiede una connessione e utilizza Internet, rispettivamente.

Idealmente, si può ipotizzare che il processo di diffusione sia quindi effettivamente completato quando le due quote appena descritte raggiungano il 100%, o comunque un valore ad esso molto vicino.

In realtà, a causa di una serie di fattori, la soglia di saturazione di ciascuno di questi indicatori potrebbe essere sensibilmente inferiore al massimo teorico, anche se difficile da stabilire a priori. Ad esempio, le persone con disabilità tali da impedirne l'uso degli strumenti informatici per l'uso di Internet potrebbero essere escluse dalla soglia di saturazione dell'uso da parte degli individui. Anche le famiglie composte solamente da individui disabili potrebbero dunque essere escluse dagli obiettivi per la diffusione delle connessioni a banda larga.

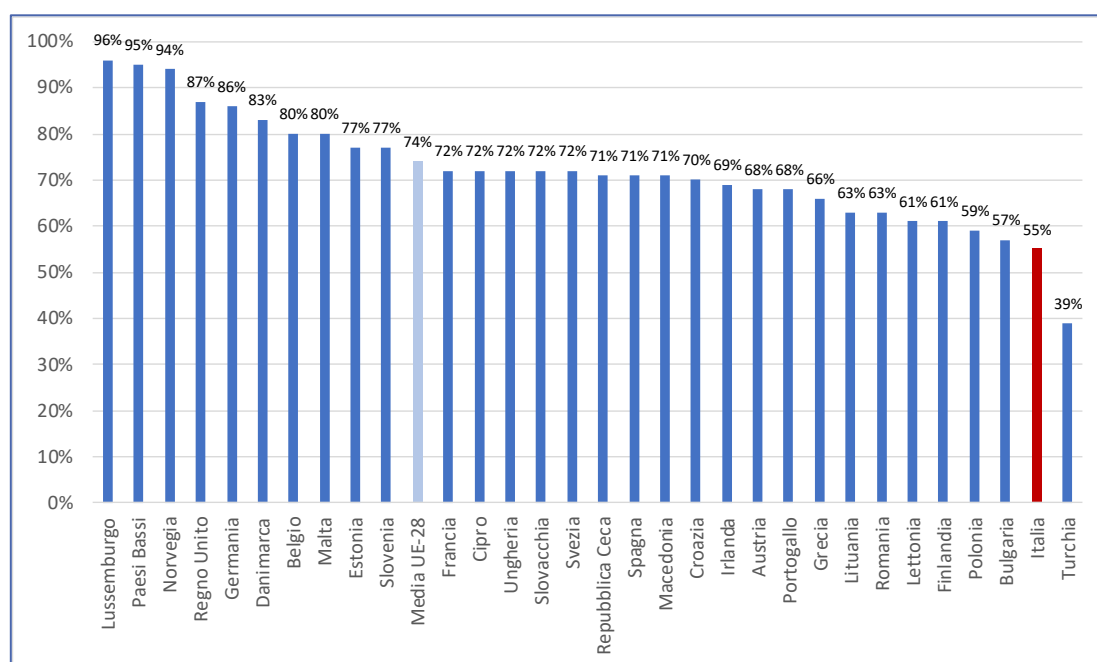
In generale, comunque, lo stato della diffusione di Internet in un dato paese può essere valutato mettendo in relazione i valori attuali, relativi agli indicatori presentati, con quelli di altri paesi comparabili, in termini di popolazione e avanzamento tecnologico ed economico.

Per l'Italia, in particolare, è possibile effettuare un raffronto con gli altri Stati Membri dell'Unione Europea. Tra questi, è particolarmente interessante paragonare i dati del nostro paese a quelli con dimensioni simili, in termini di popolazione, quali Germania, Francia e

Spagna. Al di fuori dell'Unione Europea, anche la Gran Bretagna presenta caratteristiche simili a quelle del nostro paese.

Nel grafico seguente sono mostrati gli ultimi dati disponibili rispetto alla diffusione di connessioni a banda larga tra le famiglie, per quanto riguarda le connessioni fisse (Eurostat, 2017b).

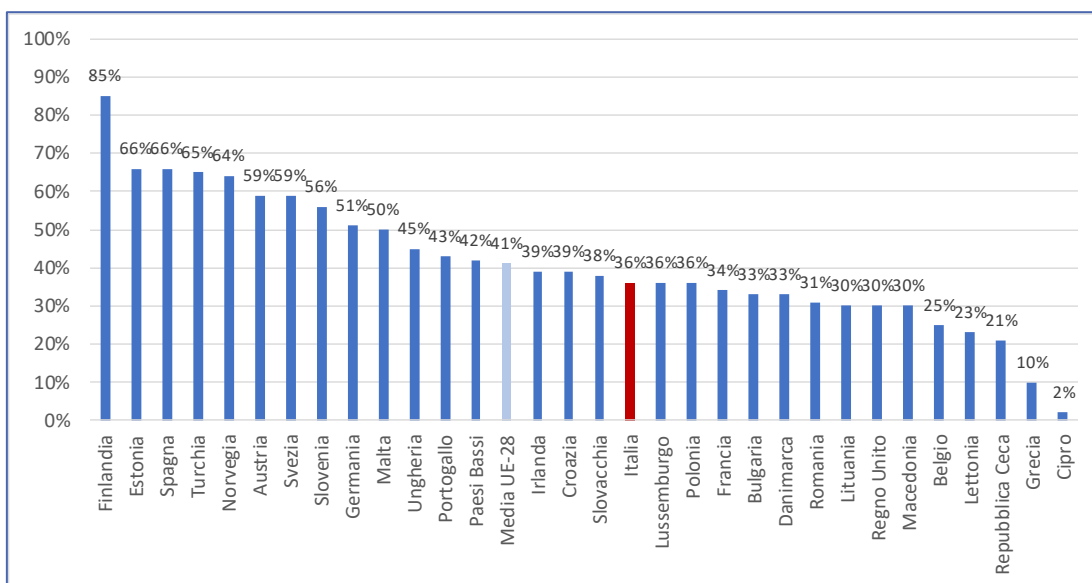
Figura 1 - Quota di famiglie (con almeno un componente tra 16 e 74 anni) con accesso ad Internet a banda larga fissa, per paese di residenza (anno 2016) - Fonte: Eurostat



Come si può notare, l'Italia con il suo 55% risulta essere, ancora al 2016, ultima per quota di famiglie con accesso ad Internet con una connessione fissa a banda larga in tutta l'Unione Europea, e penultima considerando anche altri paesi europei, davanti alla sola Turchia. Paesi con popolazione comparabile a quella italiana sono in posizione nettamente migliore: il Regno Unito presenta un tasso di adozione di banda larga fissa dell'86% delle famiglie, seguito dalla Germania con l'83%, la Francia con il 72% e la Spagna con il 71%. Si tratta di un dato particolarmente preoccupante, qualora si consideri che tipicamente la banda larga fissa, rispetto a quella mobile, garantisce connessioni maggiormente stabili, ed in generale viene fornita con contratti che prevedono una quantità illimitata di dati scambiati al mese.

Passando ai dati sulla diffusione delle connessioni a banda larga mobile tra le famiglie, la situazione per il nostro paese è invece sensibilmente migliore. Nel seguente grafico sono mostrati i valori relativi al 2016, per il nostro paese e per i principali paesi di benchmark.

Figura 2 - Quota di famiglie (con almeno un componente tra 16 e 74 anni) con accesso ad Internet a banda larga *mobile*, per paese di residenza (anno 2016) - Fonte: Eurostat

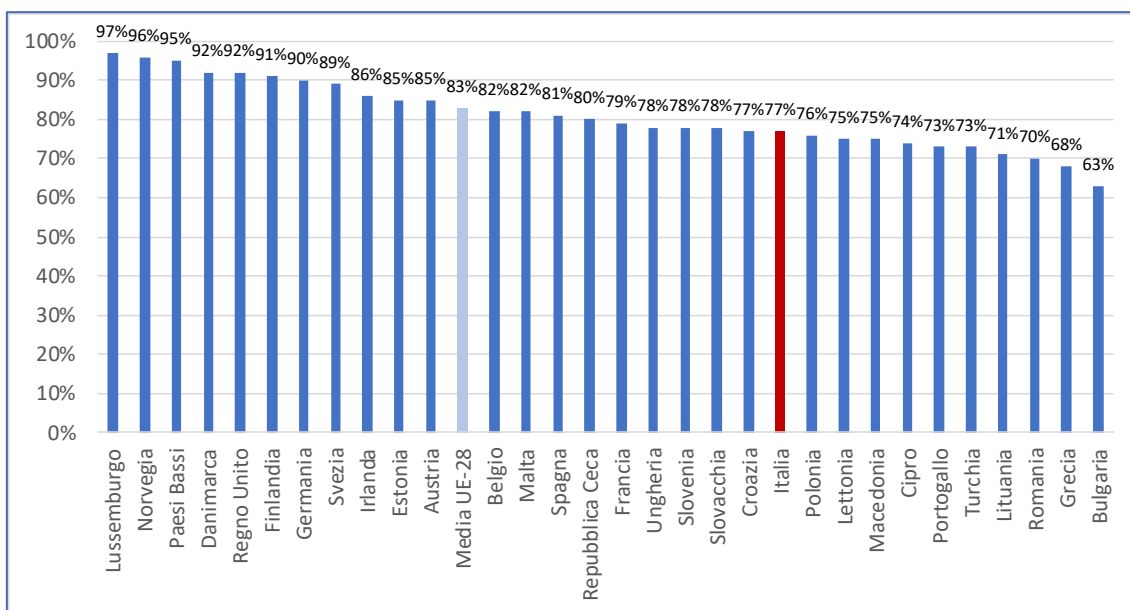


Il 36% delle famiglie italiane (con almeno un componente tra 16 e 74 anni) dispone dunque di una connessione mobile a banda larga da casa. Si tratta di un dato più vicino alla media europea (41%), di quello presentato in precedenza, che pone dunque il nostro paese in una situazione intermedia tra tutti i paesi considerati. In particolare, gli altri grandi paesi europei presentano situazioni molto differenziate. Mentre, infatti, la Spagna risulta essere tra coloro che presentano tassi di diffusione di larga banda mobile tra le famiglie più elevati, con il 66%, la Germania del 51%, mentre Francia (34%) e Regno Unito (30%) si ritrovano al di sotto dell'Italia in questa classifica.

Si noti che il possesso dei due tipi di connessione a banda larga, fissa e mobile, non è mutuamente escludente, motivo per il quale le due quote appena rappresentate non sono sommabili per ottenere la quota totale di famiglie connesse.

Per effetto della situazione riguardante la banda larga mobile, appena presentata, l'Italia recupera in parte il ritardo derivante dalla diffusione della sola banda larga fissa tra le famiglie, rispetto alla media UE e agli altri grandi paesi europei, come mostrato nel grafico seguente.

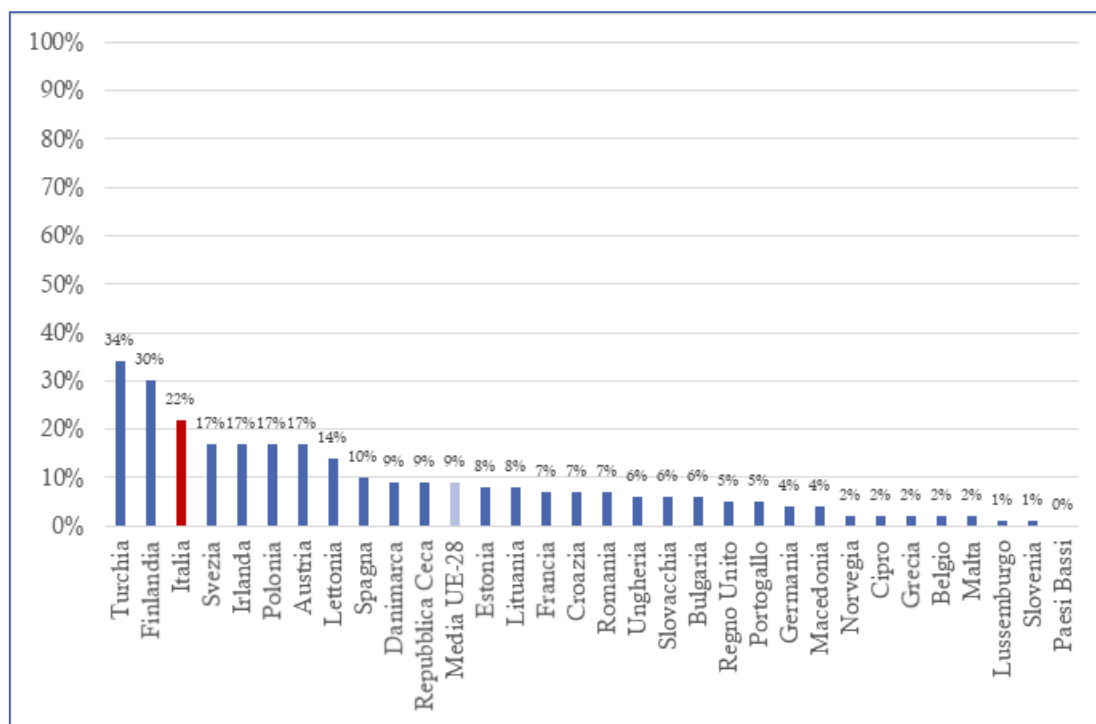
Figura 3 – Quota totale di famiglie (con almeno un componente tra 16 e 74 anni) con accesso ad Internet a banda larga, per paese di residenza (anno 2016) - Fonte: Eurostat



Ancora una volta, però, il nostro paese si ritrova ultimo tra i grandi paesi europei, laddove il Regno Unito sembra aver raggiunto un valore vicino alla saturazione (92% di famiglie connesse), così come la Germania (90%), mentre la Spagna e la Francia sembrano presentare una situazione più simile a quella italiana (81% e 79%, rispettivamente).

Si può però dedurre, per differenza tra la quota totale di famiglie connesse e la quota di famiglie con connessione a banda larga fissa, la quota di famiglie con sola connessione mobile a banda larga.

Figura 4 - Quota di famiglie (con almeno un componente tra 16 e 74 anni) con accesso ad Internet a banda larga solo mobile, per paese di residenza (anno 2016) - Fonte: Eurostat



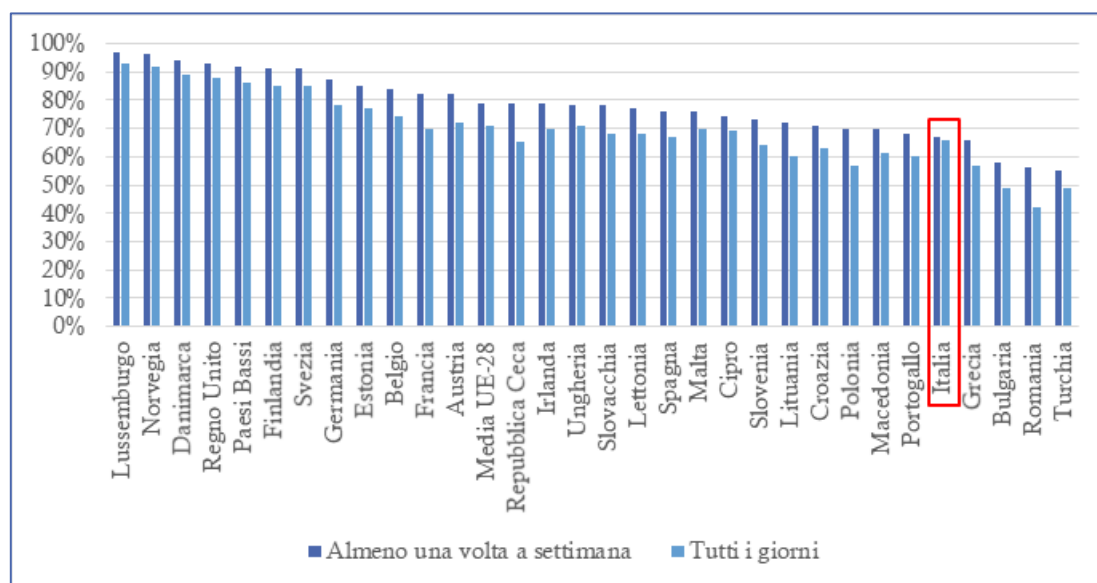
L'Italia risulta essere tra i primi stati europei per quota di famiglie con accesso a banda larga esclusivamente mobile, con il 22%, poco più di un quarto del totale delle famiglie connesse a banda larga. I paesi con maggiori tassi di diffusione di larga banda fissa (Lussemburgo, Paesi Bassi, Norvegia, Regno Unito, Germania) presentano invece, come era lecito attendersi, valori molto bassi di famiglie con sola connessione mobile, che vanno dal 5% del Regno Unito al valore sotto l'1% dei Paesi Bassi.

Come si diceva in apertura di paragrafo, un altro importante indicatore della diffusione di Internet in un paese è rappresentato dalla quota di popolazione residente che utilizza questo mezzo regolarmente. La maggior parte delle fonti statistiche disponibili, sia in Italia (Istat) che a livello di comparazione internazionale (OCSE e Eurostat principalmente) forniscono dati per diversi livelli di intensità di uso di Internet. Tuttavia, dato che l'obiettivo è valutare quanti siano effettivamente in grado di sfruttare appieno le opportunità offerte dalla rete, sembra inopportuno considerare come utenti anche coloro che dichiarano di utilizzare Internet solo sporadicamente, o che l'hanno utilizzato in passato ma non lo utilizzano più. La valutazione sulla frequenza di connessione degli utenti sarà sviluppata più diffusamente nel seguito (paragrafo 6.2), laddove sarà necessario pervenire a una suddivisione dell'utenza

di Internet articolata, al fine di descrivere l'uso in Italia da parte degli individui in maniera più precisa.

In questa sede di prima valutazione del fenomeno e di confronto internazionale, sembra utile utilizzare una delle classificazioni proposte dall'Eurostat, per la quale sono considerati utenti gli individui che dichiarano di utilizzare Internet almeno una volta a settimana. Sono esclusi da questa analisi sia i bambini e gli adolescenti sotto i 16 anni, perché ancora in fase di crescita, sia gli anziani oltre i 75 anni, che potrebbero essere progressivamente portati a non usare il mezzo per problemi di disabilità. Per raffronto, nel grafico seguente è inserita anche la quota di utenti che utilizzano Internet tutti i giorni, per ciascun paese considerato. Si noti come, a meno di alcuni casi specifici, l'andamento delle due variabili è fortemente correlato (coefficiente di correlazione¹=0,975).

Figura 5 – Quota di individui (tra 16 e 74 anni) che utilizzano Internet, per frequenza di utilizzo e paese di residenza (anno 2016). Fonte: Eurostat



Come si può notare, i valori relativi all'Italia, tra di loro molto simili, sono pari al 67% per gli utenti almeno una volta a settimana e 66% per gli utenti con frequenza quotidiana tutti i giorni. Questi dati collocano l'Italia agli ultimi posti della classifica, ben lontana da Lussemburgo (97% e 93%), Norvegia (96% e 92%) e Danimarca (94% e 89%). Se si considerano gli altri grandi paesi europei, la situazione non migliora. Il Regno Unito possiede

¹ Il coefficiente di correlazione tra due variabili X e Y, detto anche indice di correlazione di Pearson, è dato dal rapporto della covarianza delle variabili stesse sul prodotto delle rispettive deviazioni standard: $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$

il 93% di utenti regolari e l'88% di utenti quotidiani, la Germania rispettivamente l'87% e il 78% e la Francia l'82% e il 70%. Infine, la Spagna presenta un dato di utenti regolari più elevato di quello italiano, pari al 76%, ma un valore di utenti "tutti i giorni" pari a quello italiano (67%).

Dai dati analizzati emerge dunque un quadro di complessivo ritardo della diffusione del nostro paese, sia in termini di connessioni a banda larga fissa, che di connessioni mobili, che di uso regolare da parte dei cittadini. Tale ritardo si manifesta non solo rispetto agli altri paesi europei tra i primi posti per popolazione, ma a quasi tutti i paesi membri dell'Unione Europea e anche a quelli europei che non vi fanno parte.

L'ipotesi è dunque che esistano effettivamente delle specificità riguardanti il nostro paese, dall'analisi delle quali è necessario partire quando si voglia mettere a punto delle politiche, pubbliche ma anche private (in termini di marketing), atte a favorire l'uso di Internet da parte degli individui e la diffusione delle connessioni a banda larga tra le famiglie. Sull'analisi delle determinanti di questa duplice domanda si concentra il progetto di ricerca descritto in questo documento, a partire da quanto riportato in letteratura (capitolo 2).

2 Le determinanti della domanda Internet in letteratura

Così come esposto nell'introduzione, lo studio della domanda di Internet da parte della popolazione comprende diversi aspetti: l'adozione di una connessione ad Internet da parte delle famiglie e l'uso di Internet da parte dei cittadini, anche in termini di frequenza di utilizzo.

Inoltre, risulta particolarmente interessante studiare le attività svolte online da parte degli utenti, per indagare le possibili differenze all'interno di cluster di popolazione che sperimentano la stessa frequenza di connessione ad Internet.

Come si vedrà nel seguito del capitolo, i modelli di scelta discreta (McFadden, 1974; Train, McFadden e Ben-Akiva, 1987; Menard, 1995; Agresti, 2007) sono stati ampiamente utilizzati in letteratura per la modellizzazione della domanda di servizi di telecomunicazione.

2.1 Adozione della connessione e uso individuale

I lavori di analisi delle determinanti della domanda di Internet, che applichino modelli statistici a microdati, sia a livello familiare che individuale, hanno una buona diffusione in letteratura.

In uno dei primi lavori sulla banda larga, viene analizzata la domanda residenziale di questi servizi in Australia, attraverso l'uso di un modello logistico binario, che utilizza il prezzo, il reddito, la professione, l'età media e alcune caratteristiche della famiglia come variabili esplicative (Madden e Simpson, 1997). Gli autori hanno scoperto come il prezzo della connessione e il reddito della famiglia siano i fattori più importanti per l'adozione. Inoltre, l'adozione da casa risulta anche essere fortemente influenzata in positivo dall'età e dalla professione del capofamiglia e, in negativo, dalla dimensione della famiglia.

Chaudhuri et al. hanno invece analizzato l'impatto di una serie di variabili socioeconomiche sulla decisione delle famiglie di dotarsi di una connessione ad Internet *narrowband*, attraverso l'uso di un modello di scelta discreta applicato a un'indagine sulle famiglie statunitensi (Chaudhuri, Flamm e Horrigan, 2005). La maggior parte delle variabili sociodemografiche considerate hanno prodotto un effetto del segno atteso. L'influenza del reddito e del grado di istruzione sembrerebbe essere particolarmente rilevante. L'analisi suggerisce anche come la decisione di adozione sia moderatamente sensibile alle variazioni di prezzo.

Tale studio è stato poi esteso ai servizi a banda larga (Flamm e Chaudhuri, 2007). I risultati mostrano come la decisione di dotarsi di una connessione qualsiasi e quella più specifica di dotarsi di un accesso a banda larga, quando si è già in possesso di una connessione base, potrebbero essere influenzate in maniera differente dai diversi fattori socioeconomici della famiglia. Lo studio mostra come la domanda di banda larga sia sensibilmente influenzata dal prezzo della stessa. Sebbene essa sia relativamente inelastica, il coefficiente stimato è abbastanza elevato da indicare un impatto significativo. D'altro canto, gli autori hanno scoperto che altri fattori non collegati al prezzo, come lo stato civile e l'iscrizione ad un corso di studi, il genere e il risiedere in aree metropolitane, abbiano un impatto differente a seconda che si analizzi l'adozione di una connessione Internet base e il passaggio a connessioni a banda larga. Lo studio dimostra come i fattori più importanti che spingono al passaggio da dial-up a banda larga siano ancora il reddito e il livello di istruzione della famiglia.

Cerno e Pérez Amaral hanno invece costruito due diversi modelli econometrici per analizzare la connessione a banda larga da casa e la frequenza nell'uso di Internet in Spagna (Cerno e Pérez Amaral, 2006). Attraverso un modello probit in due fasi, gli autori mostrano come la domanda di banda larga delle famiglie spagnole sia correlata positivamente al reddito e al possesso di un PC e negativamente a fattori socio-demografici come la dimensione del comune di residenza e l'età. Essi hanno poi stimato un modello di domanda di uso di Internet, prendendo in considerazione il fatto che gli individui usano Internet in contesti differenti (casa, scuola, lavoro). Anche questo studio mostra come fondamentali il reddito e il livello di istruzione, questa volta in favore di un uso frequente della Rete.

Un altro interessante studio riguarda l'analisi della domanda di connessione ad Internet delle famiglie in Austria, con particolare riguardo alle connessioni a banda larga (Cardona *et al.*, 2009). Nella loro definizione di banda larga, gli autori hanno incluso anche la banda larga mobile via UMTS o HSDPA, anche al fine di derivare indicazioni sulla definizione regolamentare del relativo mercato di riferimento. I risultati della stima suggeriscono che la domanda per servizi DSL è più elastica nelle aree dove Internet via cavo e la banda larga mobile sono presenti, e le reti via cavo sono probabilmente nello stesso mercato delle connessioni DSL sia a livello *retail*, che *wholesale*.

Ancora, in un altro studio riguardante la Spagna (Lera-López, Billon e Gil, 2011) viene analizzato l'impatto di una serie di variabili socioeconomiche, demografiche e geografiche

nello spiegare l'adozione di Internet e la frequenza di utilizzo da parte degli individui in Spagna, attraverso l'uso di modelli binomiali e ordinali di tipo probit, con procedura di stima a due stadi di Heckman. Ciò consente di distinguere tra le diverse variabili e spiegare entrambe le uscite, l'utilizzo e la relativa intensità. L'uso di Internet risulta essere principalmente associato all'educazione, l'età, la condizione professionale, l'occupazione nel settore dei servizi, la nazionalità, la residenza in aree urbane e il PIL regionale pro capite. La frequenza di utilizzo di Internet è invece correlata positivamente alla presenza di una connessione a banda larga a casa, l'istruzione, le modalità attraverso le quali vengono acquisite le competenze Internet, il sesso, e la dimensione della popolazione nell'area di residenza.

Un modello simile è utilizzato in uno studio che analizza la domanda di banda larga fissa in Thailandia (Srinuan e Bohlin, 2013). Gli autori propongono infatti un modello probit bivariato, applicato ai dati ottenuti da un sondaggio nazionale del 2010 della Commissione Nazionale per la Radiodiffusione e le Telecomunicazioni (NBTC) della Thailandia. Nella prima fase, essi esaminano se la copertura effettiva di rete e alcune variabili socioeconomiche e geografiche hanno impatto sull'accesso fisso a banda larga. Nella seconda fase, viene stimato l'uso specifico di Internet, a condizione che l'accesso esista. I risultati mostrano che le variabili che hanno un potenziale impatto sono: presenza di accesso fisso, reddito, sesso, livello di istruzione, età dell'utenza e zona residenziale. Gli autori ritengono che l'impatto di questi fattori vari in relazione alle specifiche attività svolte online.

Uno studio originale (Whitacre e Rhinesmith, 2016) analizza invece il fenomeno abbastanza diffuso delle famiglie che hanno avuto una connessione ad Internet in passato, ma hanno deciso di non rinnovarla, che vengono denominate "de-adottanti" ("*un-adopters*"). Applicando un modello logit multinomiale ai dati di un'indagine sulla popolazione statunitense del 2013, lo studio mostra che, in confronto con i loro omologhi che non hanno mai avuto una connessione ad Internet, le famiglie "de-adottanti" sono significativamente più propense a citare i seguenti motivi principali per l'annullamento della sottoscrizione al servizio Internet: i costi elevati, l'uso di Internet in altro luogo da parte dei membri e l'inadeguatezza dei propri dispositivi per la connessione.

Diversi altri lavori in letteratura affrontano il tema delle determinanti domanda di Internet (Choudrie e Dwivedi, 2005; Savage e Waldman, 2005; Hitt e Tambe, 2007; Prieger e Hu,

2008; Nisbet, Stoycheff e Pearce, 2012; Weiner *et al.*, 2012; Roycroft, 2013; Lyons, 2014; Köksal e Anil, 2015; Sadowski, 2017).

Sono presenti anche alcuni lavori che provano ad indagare l'uso di Internet in Italia a partire dai dataset delle indagini Istat. In particolare, la stessa Istat ha condotto, insieme alla Fondazione Ugo Bordoni (FUB), una serie di analisi riguardanti l'uso di Internet da parte degli individui e delle imprese, in due studi successivi denominati rispettivamente *Internet@Italia 2013 e 2014* (Istat e FUB, 2014, 2015)². All'interno del primo soprattutto, è presente un'analisi tramite regressioni logistiche simile a parte di quella presentata nel seguito (cfr. 6.1.1). In una prima analisi relativa a tutta la popolazione di 6 anni e più, per le quali sono effettivamente rilevate la maggior parte delle variabili ICT, l'età risulta essere di gran lunga la variabile più rilevante nello spiegare l'uso regolare di Internet, seguita dal titolo di studio e dalla condizione professionale. Il lavoro presenta poi delle analisi di regressione separate per ciascuna delle classi di età individuate. Emerge così come all'interno di differenti classi, l'importanza relativa delle singole variabili socio-demografiche considerate sia sostanzialmente diversa. In particolar modo, il titolo di studio sembra di gran lunga predominante tra gli ultra-sessantacinquenni, mentre nelle altre classi lo stesso resta comunque la variabile più importante, ma man mano che si passa ad età più giovani altre determinanti assumono sempre maggiore rilevanza: la condizione professionale e la posizione professionale, per le classi intermedie; la situazione economica della famiglia per la classe più giovane di adulti (19-34 anni). I due lavori presentano anche un'analisi descrittiva delle attività svolte online, suddivisa per differenti variabili di segmentazione, e un'analisi delle attività culturali svolte *offline*, sulla base di alcune variabili di segmentazione e della tipologia di utenza di Internet.

Particolarmente interessante ai fini del presente studio, poiché ad esso in qualche modo “parallelo”, è anche il lavoro di (Quaglione *et al.*, 2017). Gli autori, utilizzando i dati provenienti dalla stessa indagine “Aspetti della vita quotidiana” di Istat, che come si vedrà costituiscono la fonte primaria anche di questa ricerca, relativi al solo 2014, costruiscono un modello probit in due stadi *à la* Heckman, per indagare le determinanti della frequenza d'uso di Internet e dell'adozione di connessioni a banda larga in Italia. Il modello utilizza in input, per quel che riguarda la frequenza d'uso, di una serie di variabili socio-demografiche (quali

² Del gruppo di lavoro Istat-FUB ha fatto parte anche Cosimo Dolente, in quanto ricercatore FUB.

l'età, la situazione economica, il titolo di studio ecc.) e di competenze digitali. La seconda parte del modello, relativo alla banda larga, utilizza anch'esso alcune variabili sociodemografiche, comuni alla prima parte, e una serie di variabili sullo specifico uso di Internet e in generale sulle abitudini ICT dell'individuo e sui possessi della famiglia di appartenenza. I risultati mostrano come l'età giovane, il genere maschile e la condizione di occupato giocano un ruolo importante nello spiegare un'elevata frequenza d'uso di Internet. In maniera controintuitiva, il modello mostra una probabilità più elevata di avere una connessione fissa a banda larga da casa per i più anziani. Gli autori spiegano questo risultato con l'argomentazione che, trattandosi di un modello di Heckman, il primo passo filtra già gli utenti più frequenti, e dunque è plausibile che anziani che siano già utenti frequenti, per motivi anche di carattere economico e di facilità di utilizzo, utilizzino più frequentemente connessioni fisse, laddove le fasce più giovani della popolazione, di per loro più innovatrici, sono più portate a sostituire le connessioni fisse con quelle mobili. Inoltre, un altro risultato rilevante è l'impatto del reddito sull'uso, ma non sull'adozione della connessione a banda larga. L'ipotesi è che il costo dei dispositivi di connessione costituisca l'ostacolo di gran lunga maggiore, da un punto di vista economico, nell'adozione di Internet, invece che quello della connessione. Dunque, le questioni di reddito hanno un impatto sull'uso, ma non sulla successiva decisione di adozione di una connessione a banda larga, dato che nel modello i fenomeni sono stimati in *step* successivi.

Dall'analisi della letteratura emerge dunque un ruolo molto importante dei fattori demografici, socioeconomici (principalmente l'istruzione, la condizione professionale e la posizione professionale) e culturali nell'uso regolare da parte degli individui e nella sottoscrizione della connessione da parte della famiglia, al fianco delle questioni più strettamente legate al costo di apparecchiature e connessione.

2.2 Attività online e differenze tra utenti

Le attività svolte online e le differenze tra utenti sono state studiate in particolare da Jan A.G.M. Van Dijk in diversi studi, da più di una decade (Van Dijk e Hacker, 2003). In uno in particolare (Van Dijk, 2012), l'autore propone un'interpretazione originale del divario digitale come prodotto di disuguaglianze in quattro successivi stadi di accesso: la motivazione, l'accesso fisico, le competenze digitali e l'uso differente. L'ipotesi è che il divario si sia man mano trasferito dal primo all'ultimo tipo di accesso nei dieci anni tra il 2002 e il 2012. Il

documento descrive l'attuale situazione europea di tutti e quattro i tipi di accesso, attraverso la valutazione delle risorse culturali degli individui e l'uso della teoria dell'appropriazione³.

In uno studio successivo, Van Dijk e Van Deursen verificano l'ipotesi dello spostamento del divario digitale verso lo specifico tipo di utilizzo che i diversi utenti fanno della Rete (Van Deursen e Van Dijk, 2013). Utilizzando un'indagine sulla popolazione dei Paesi Bassi, essi mostrano come le persone con bassi livelli di istruzione e le persone disabili utilizzano Internet per più ore al giorno, rispetto alle persone più istruite e occupate. Per spiegare questo risultato, gli autori effettuano un'analisi delle specifiche attività svolte online. Una prima analisi approfondita è svolta attraverso un *clustering* dei tipi di attività suddivisi in: informazione, notizie, sviluppo personale, interazione sociale, tempo libero, transazioni commerciali e gioco. Nella seconda parte del lavoro, sulla base di questa classificazione derivante dalla prima analisi, gli autori identificano un certo numero di differenze nell'utilizzo, influenzate dal genere, dall'età, dai livelli di istruzione e dall'esperienza su Internet. La conclusione generale del documento è che, man mano che l'uso di Internet tra la popolazione si diffonde e consolida nel tempo, in esso si riflettono sempre di più i rapporti sociali, economici e culturali già noti del mondo offline, ivi comprese le disuguaglianze tra differenti segmenti della popolazione. Una valutazione simile era stata già proposta nel 2001 da DiMaggio e Hargittai, che in un loro studio analizzavano come, oltre al digital divide in fase di superamento, si dovesse iniziare a valutare le *digital inequalities* (disuguaglianze digitali) emergenti all'interno dell'ampia categoria degli utenti Internet (DiMaggio e Hargittai, 2001). Il concetto è stato da allora ampiamente ripreso in letteratura (Robinson *et al.*, 2015).

Il modello teorico proposto in (Van Dijk, 2012) sui quattro tipi di accesso successivi viene poi ulteriormente testato in (Van Deursen, Van Dijk e Ten Klooster, 2015), ancora su di un campione rappresentativo della popolazione olandese. Lo studio dimostra che, sebbene il focus delle *policy* riguardanti il digital divide si sia spostato effettivamente sulle competenze digitali e sull'utilizzo, le questioni motivazionali e materiali assumono ancora rilevanza. Inoltre, nell'articolo si afferma che ogni stadio dell'accesso ha differenti determinanti, sebbene essi interagiscano tutti tra loro per produrre le disuguaglianze digitali effettivamente

³ Si tratta di una teoria sociologica che enfatizza le posizioni nella società e le relazioni tra gli individui, invece degli attributi dei singoli, sviluppata dallo stesso autore qualche anno prima (Van Dijk, 2005).

riscontrate nella società. Dunque, gli autori suggeriscono che le politiche sul digital divide dovrebbero contemporaneamente tener conto di tutti e quattro gli stadi di accesso.

Diversi altri studi seguono approcci simili a quello appena illustrato, evidenziando le differenze nella frequenza e nelle tipologie di uso all'interno degli utenti di Internet (Dutton e Blank, 2014; van Deursen e Helsper, 2015; Büchi, Just e Latzer, 2016). Sostanzialmente, ne emerge un quadro nel quale solo alcuni degli utenti della Rete sono effettivamente in grado di sfruttarne appieno le potenzialità, mentre molti altri ne fanno comunque un uso limitato, sostanzialmente edonistico e comunicativo.

Inoltre, come si diceva nel paragrafo precedente, anche i due report a firma Istat-FUB (Istat e FUB, 2014, 2015) presentano alcune analisi descrittive rispetto alle attività svolte online da diverse categorie di utenti.

In generale, dunque, nell'analisi delle determinanti della domanda Internet e del digital divide, sembra importante valutare, al fianco delle questioni relative all'adozione della connessione da parte della famiglia e dell'uso regolare da parte degli individui, anche la frequenza di accesso degli individui e i differenti tipi di uso, in termini di attività svolte online.

3 Stato della diffusione di Internet in Italia

Dal confronto internazionale presentato nel capitolo 0 emerge dunque un quadro di complessivo ritardo della diffusione di Internet in Italia, sia in termini di connessioni che di utilizzo. Il ritardo è ancora più accentuato considerando i soli grandi paesi europei, soprattutto Regno Unito e Germania. Quali sono le cause di questo ritardo?

L'analisi della letteratura (capitolo 2) aiuta a individuare le variabili socio-economiche e demografiche rilevanti, sulle quali valutare la suddivisione tra famiglie connesse e non connesse, e tra utenti e non utenti. Le determinanti principali possono dunque essere individuate in: età, livello di istruzione, condizione professionale e posizione professionale.

Nel seguito del capitolo si presenta dunque una prima analisi descrittiva della suddivisione dell'uso (paragrafo 3.3) e delle connessioni (paragrafo 3.4) su queste variabili, svolta a partire dalla fonte di dati principali della ricerca, l'indagine annuale "Aspetti della Vita Quotidiana" (AVQ) di Istat, presentata nel paragrafo 3.1. Al fine di poter comprendere al meglio tale analisi, viene dapprima proposta una veloce disamina della struttura della popolazione italiana sulle stesse variabili selezionate (paragrafo 3.2).

Il capitolo si chiude poi con alcune statistiche descrittive sui motivi della non adozione dichiarati dalle famiglie, e del non uso dichiarati dagli individui, contenuti anch'essi all'interno dell'indagine AVQ di Istat (paragrafo 3.5).

3.1 Fonti di dati della ricerca

Per analizzare più a fondo la questione italiana, all'interno di questa ricerca ci si è avvalsi in particolar modo dei microdati (l'insieme delle singole risposte) provenienti dalla rilevazione "Aspetti della Vita Quotidiana" (AVQ), che viene effettuata da Istat ogni anno su un campione variabile tra i 45.000 e i 50.000 individui, raccolti in circa 20.000 famiglie. Il campione è rappresentativo di tutta la popolazione italiana, su entrambi i livelli (familiare e individuale), con significatività fino a un dettaglio geografico NUTS-2⁴, che in Italia corrisponde alle Regioni. Utilizzando i coefficienti di peso forniti dall'Istat è possibile proiettare all'universo di riferimento le osservazioni che rispettano determinati criteri di

⁴ classe 2 della Nomenclatura delle Unità Territoriali Statistiche (NUTS), classificazione utilizzata dall'Unione Europea che suddivide i territori economici dell'UE col fine di raccogliere e armonizzare le statistiche europee. La classe 2 in Italia corrisponde alla suddivisione amministrativa delle Regioni.

selezione, ottenendo il corrispondente numero di individui nella popolazione residente in Italia.

All'interno della ricerca sono stati utilizzati, per elaborazioni diverse, i microdati relativi al periodo di 10 anni che va dal 2005 al 2014, anno più recente disponibile nell'ultima data utile per l'inizio delle elaborazioni finali.

I microdati provenienti da questa indagine possono essere utilizzati al fine di stimare un gran numero di tavole statistiche. Come si vedrà più avanti, però, l'uso dei microdati permette un livello di dettaglio molto elevato, nell'esplorare i comportamenti individuali e familiari. Gli aspetti coperti dall'indagine riguardano diversi campi della vita quotidiana dei cittadini italiani, tra cui:

- Caratteristiche socio-demografiche e professionali degli individui;
- Istruzione;
- Struttura e composizione familiare;
- Caratteristiche e problemi dell'abitazione di residenza e della zona;
- Spostamenti quotidiani;
- Abitudini alimentari e stile di vita;
- Stato di salute e consumo di medicinali;
- Abitudini e capacità relative all'ICT;
- Consumi culturali, tempo libero e partecipazione sociale;
- Possesso di beni da parte della famiglia;
- Problematiche ambientali e riciclo;
- Sicurezza;
- Soddisfazione per diversi aspetti della vita.

In particolare, le principali variabili di segmentazione presenti nel database sono mostrate in Tabella 1, con l'elenco delle relative modalità, estratte dal database relativo al 2012. Tali modalità sono sostanzialmente invariate negli anni di rilevazione, salvo alcune aggregazioni operate dall'Istat a partire dal 2013, soprattutto per quel che riguarda l'età e il titolo di studio.

Tabella 1 - Principali variabili di segmentazione della rilevazione "Aspetti della vita quotidiana" di Istat, e relative modalità

| Variabile | Modalità (anno 2012) |
|---|---|
| Genere | Maschio |
| | Femmina |
| Età | Variabile continua (anni) |
| Titolo di studio | Dottorato di ricerca o specializzazione post-laurea |
| | Laurea di 4 anni o più (vecchio ordinamento o nuova laurea specialistica a ciclo unico) |
| | Laurea specialistica di 2 anni di secondo livello (nuovo ordinamento) |
| | Laurea di 3 anni di primo livello (nuovo ordinamento) |
| | Diploma universitario, laurea breve vecchio ordinamento, scuola diretta a fini speciali, scuola parauniversitaria |
| | Accademia belle arti, istituto superiore industrie artistiche, accademia di arte drammatica, scuola interpreti e traduttori, perfezionamento accademia di danza, perfezionamento conservatorio, scuola di archivistica, perfezionamento istituto di musica pareggiato, paleografia, diplomatica |
| | Diploma di scuola superiore 4-5 anni che permette l'iscrizione all'Università |
| | Diploma di scuola superiore 2-3 anni che non permette l'iscrizione all'Università |
| | Licenza media (o avviamento professionale) |
| | Licenza elementare |
| | Nessun titolo (sa leggere e scrivere) |
| | Nessun titolo (non sa leggere e/o scrivere) |
| | Condizione professionale (solo per rispondenti di 15 anni e più) |
| In cerca di nuova occupazione | |
| In cerca di prima occupazione | |
| Studente | |
| Casalinga | |
| Inabile al lavoro | |
| Ritirato dal lavoro | |
| Altra condizione | |
| Posizione nella professione (solo per rispondenti di 15 anni e più) | <i>alle dipendenze come:</i> |
| | dirigente |
| | direttivo, quadro |
| | impiegato, intermedio |
| | capo operaio, operaio subalterno e assimilati |
| | apprendista |
| | lavorante a domicilio per conto d'impresa |
| | <i>autonomo come:</i> |
| | imprenditore |
| | libero professionista |

| Variabile | Modalità (anno 2012) |
|---------------------------|---|
| | lavoratore in proprio |
| | socio cooperativa Produzione Beni e/o prestazioni di servizio |
| | coadiuvante |
| | <i>* Se non occupato il codice, quando presente, si riferisce all'occupazione passata</i> |
| Ripartizione territoriale | Nord Ovest |
| | Nord Est |
| | Centro |
| | Sud |
| | Isole |
| Dimensione del comune | Comuni centro di area metropolitana |
| | Comuni periferia dell'area metropolitana |
| | Comuni fino a 2.000 abitanti |
| | Comuni tra 2.001 e 10.000 abitanti |
| | Comuni tra 10.001 e 50.000 abitanti |
| | Comuni con oltre 50.000 abitanti |

Passando invece all'uso di Internet, di seguito sono riportate alcune delle principali variabili che descrivono la frequenza di uso da parte degli individui e il tipo di connessione da casa. Riguardo alla seconda, le modalità qui riportate fanno riferimento alle indagini svolte a partire dal 2010, mentre nelle precedenti la suddivisione era più semplificata, non tenendo conto delle connessioni mobili a banda larga, non ancora diffuse sul mercato. La gran parte delle variabili relative all'uso dell'ICT, ivi comprese quelle elencate in Tabella 2, sono rilevate a partire dai 6 anni di età.

Tabella 2 - Principali variabili relative all'uso di Internet e alla connessione, contenute nella rilevazione "Aspetti della vita quotidiana" di Istat, e relative modalità

| Variabile | Modalità (anno 2012) |
|--|---|
| Uso di Internet | sì, negli ultimi 3 mesi |
| | sì, da 3 mesi a 1 anno fa |
| | sì, più di 1 anno fa |
| | mai |
| Frequenza d'uso di Internet negli ultimi 12 mesi | tutti i giorni |
| | qualche volta alla settimana |
| | una volta alla settimana |
| | qualche volta al mese (meno di quattro volte) |
| | meno di una volta al mese |
| Tipo di connessione con cui la famiglia accede ad Internet da casa | Linea telefonica tradizionale o ISDN |
| | DSL (ADSL, SHDSL, VDSL ecc.) |
| | Connessione a banda larga via cavo (fibra ottica, rete locale, PLC, ecc.) |
| | Connessione senza fili fissa (satellite, rete pubblica WIFI, Wimax) |

| Variabile | Modalità (anno 2012) |
|-----------|--|
| | Connessione senza fili mobile con palmare o telefonino 3G (UMTS, mobile, ecc.) |
| | Connessione senza fili mobile con modem 3G (chiavetta USB, scheda) |
| | Connessione senza fili mobile a banda stretta (GPRS, 2G+ ecc.) |

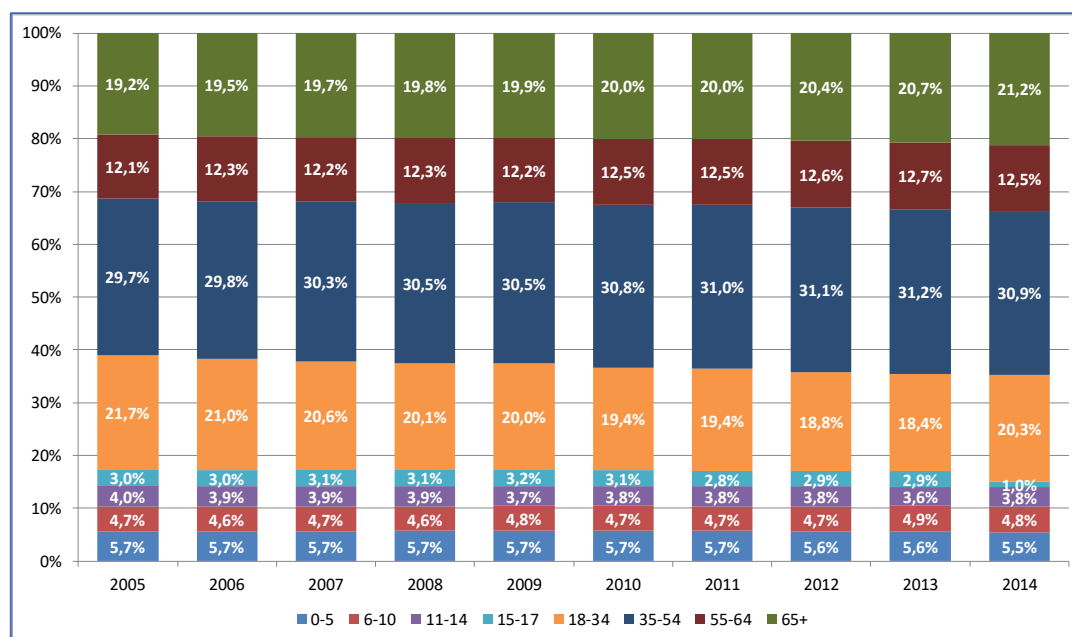
La stessa Istat produce ogni anno delle statistiche sintetiche derivanti dall'indagine AVQ, che negli ultimi anni sono presentate, congiuntamente alle statistiche sulle imprese provenienti da altre indagini, all'interno dei comunicati "Cittadini, imprese e ICT".

Inoltre, come si diceva nel corso dell'analisi della letteratura (cfr. paragrafo 2.1), Istat ha pubblicato due approfondimenti sul tema dell'uso di Internet in Italia, dal titolo "Internet@Italia" (Istat e FUB, 2014, 2015), in collaborazione con la Fondazione Ugo Bordoni.

3.2 Struttura demografica della popolazione italiana

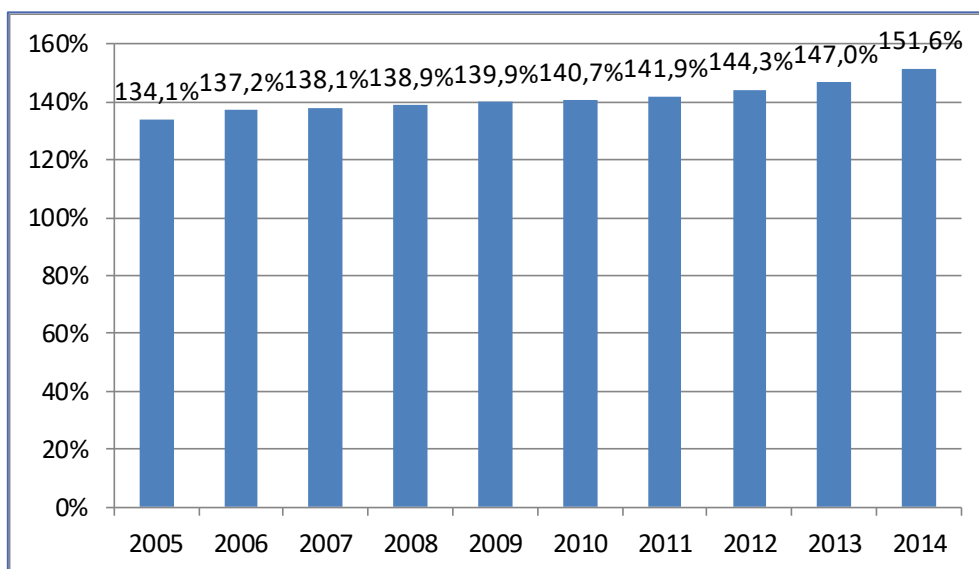
Prima di presentare una prima analisi descrittiva dell'uso di Internet, è opportuno analizzare la composizione socio-demografica della popolazione italiana, partendo dalla composizione della stessa per classi di età, variabile che in letteratura ricorre come determinante della domanda Internet.

Figura 6 - Struttura per età della popolazione italiana (anni 2005-2014). Fonte: elaborazione su dati microdati Istat AVQ



L'evoluzione nel tempo della struttura per età è alquanto stabile nel periodo relativamente breve considerato, come era lecito attendersi. Tuttavia, anche su tale breve lasso di 10 anni, si può notare la tendenza al progressivo invecchiamento della popolazione del nostro paese. Infatti, le persone di 35 anni e più passano dal 61,0% nel 2005 al 64,7% nel 2014, un aumento nel periodo di 3,7 punti percentuali del peso relativo sulla popolazione. Quella italiana si configura come una popolazione matura, che sta continuando ad invecchiare. Valutando l'indice di vecchiaia⁵ della popolazione nello stesso periodo, si ottengono i valori mostrati in Figura 7.

Figura 7 - Indice di vecchiaia della popolazione italiana (anni 2005-2014). Fonte: elaborazione su dati microdati Istat AVQ



La crescita dell'indice è alquanto importante, pari a 17,5 punti percentuali nel periodo considerato. Il trend di invecchiamento della popolazione è quindi consolidato, e si iscrive nel più generale invecchiamento della popolazione europea, mitigato in parte dall'immigrazione massiva degli ultimi anni dai paesi africani e asiatici (Eurostat, 2017a).

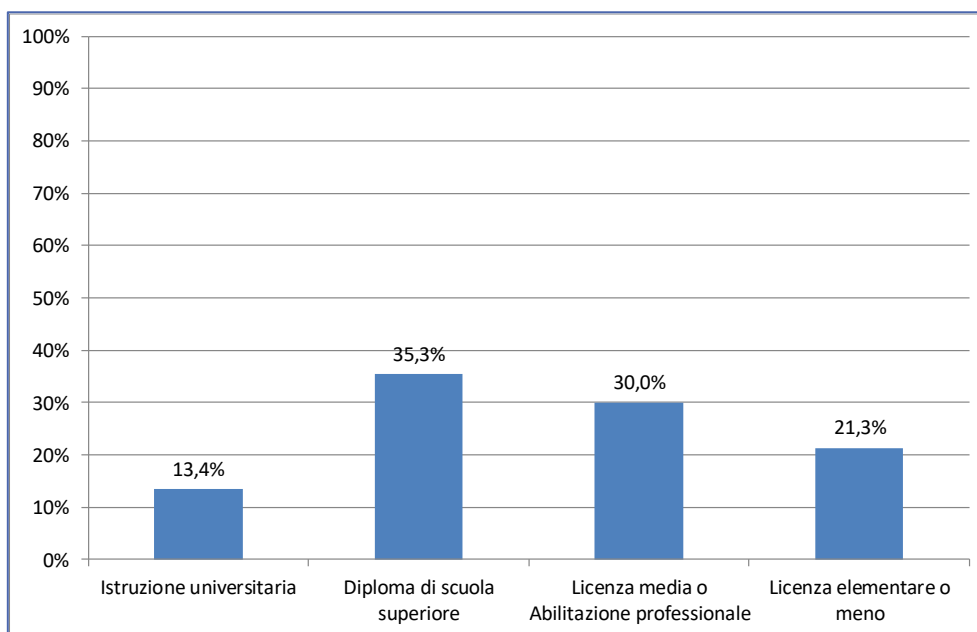
Passiamo ora ad una breve valutazione della popolazione italiana per livello di studio e condizione professionale.

Per quanto riguarda il primo, è possibile riassumere le modalità del titolo di studio, illustrate in Tabella 1, in quattro categorie, corrispondenti appunto ai diversi livelli di istruzione:

⁵ Esso è definito come rapporto tra la quota di ultrasessantacinquenni nella popolazione e la quota di minori fino a 14 anni, espresso in percentuale.

istruzione universitaria, diploma di scuola media superiore, diploma di scuola media inferiore, licenza elementare o nessun titolo. Nella Figura 8 è riportata la suddivisione della popolazione italiana su queste 4 categorie. Sono stati esclusi gli individui ancora in formazione, poiché il loro livello di studio non può essere considerato definitivo.

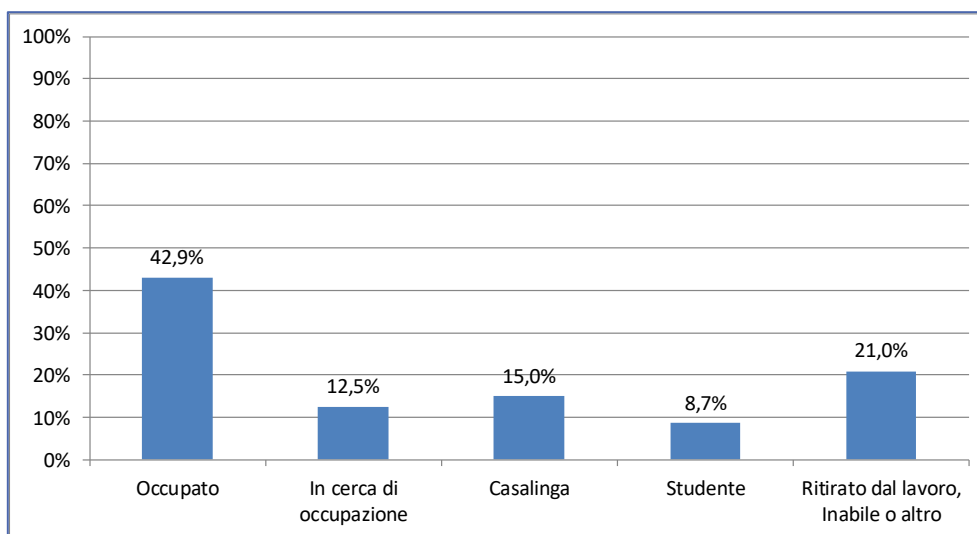
Figura 8- Popolazione italiana, fuori dal circuito scolastico, per titolo di studio (anno 2014). Fonte: elaborazione su dati microdati Istat AVQ



Dunque, la popolazione italiana risulta essere ancora composta in maggioranza di persone con livelli di istruzione bassi, quali licenza media o elementare, mentre l'istruzione universitaria non raggiunge il 15% del totale. È una condizione che influenza negativamente diversi aspetti della vita del nostro paese, e sicuramente anche la diffusione e l'uso pervasivo di Internet, come si vedrà nel seguito.

Per quanto riguarda la condizione professionale (Figura 9), solo il 43% delle persone di almeno 15 anni è occupato, valore cui va aggiunto un 13% circa di persone in cerca occupazione. Ad essi si può aggiungere il 9% di studenti, considerandoli comunque parte attiva della società.

Figura 9 - Popolazione italiana di 15 anni o più per condizione professionale (anno 2014). Fonte: elaborazione su dati microdati Istat AVQ



Vi è dunque una quota molto significativa della popolazione, circa il 37%, che risulta essere inattiva, in quanto non partecipa a processi di formazione, né al mercato del lavoro. Questa si compone di principalmente da casalinghe, inabili o ritirati dal lavoro. Si tratta di segmenti ad alta marginalità sociale, difficili da raggiungere per qualsiasi politica pubblica che voglia incentivare l'uso della Rete e l'adozione di una connessione da parte delle famiglie (cfr. paragrafo 7.2, ad esempio).

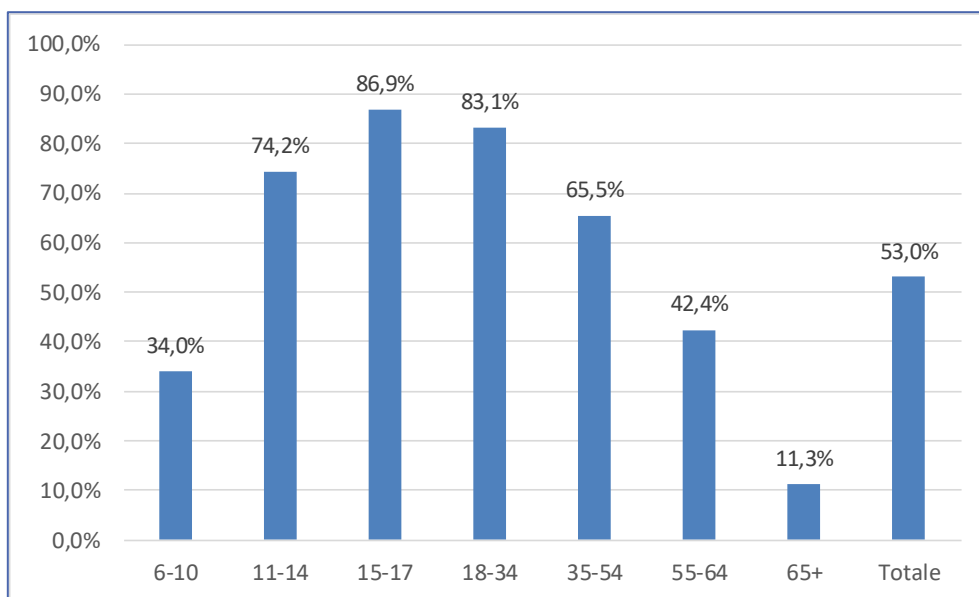
3.3 L'uso di Internet da parte degli individui

La situazione appena descritta, rispetto alla composizione socio-demografica della popolazione italiana sulle variabili principali (età, livello di istruzione e condizione professionale), non può che incidere sulla propensione all'utilizzo di nuove tecnologie, in particolare sulla diffusione di Internet tra la popolazione italiana.

Nella Figura 10 è riportata la diffusione dell'uso regolare di Internet (almeno una volta a settimana) per le diverse classi di età della popolazione italiana. Le classi risultano essere più aggregate di quelle utilizzate per l'analisi di regressione presentata più avanti (paragrafo 6.1.1), al fine di renderne più chiara la visualizzazione. La classe 0-5 non è presente, in quanto per essa non vengono rilevate le abitudini relative all'uso delle ICT.

Figura 10 - Quota di utenti Internet (almeno una volta a settimana) per classe di età, in Italia (anno 2014).

Fonte: elaborazione su dati microdati Istat AVQ

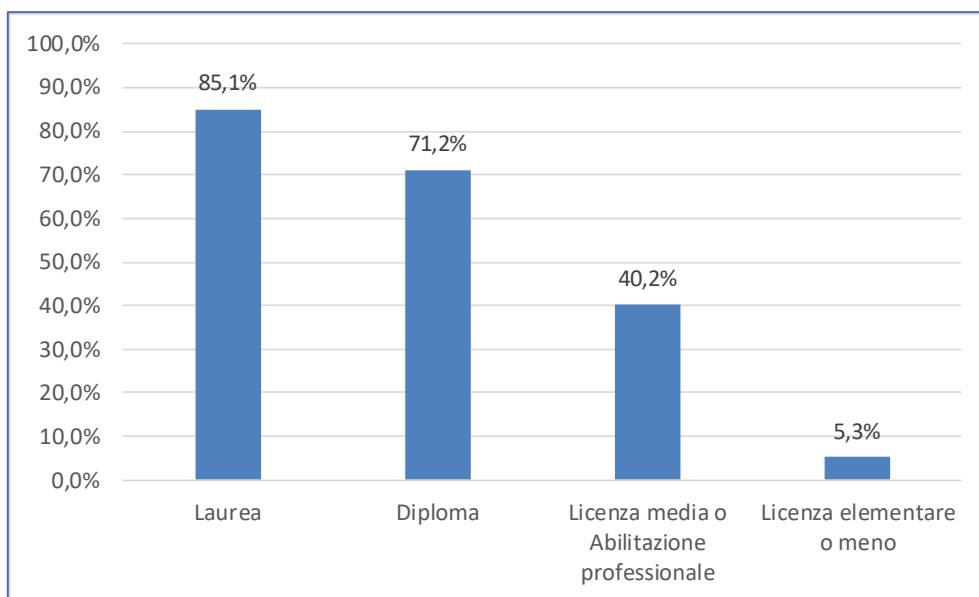


Nelle classi di età più giovani, a partire dalla pre-adolescenza (quindi tralasciando quella tra 6 e 10 anni), si riscontra dunque la maggiore inclinazione all'uso di questo medium. Se si legge questo dato congiuntamente con la struttura della popolazione italiana per età (Figura 6), anche considerando il trend di invecchiamento evidenziato dall'evoluzione dell'indice di vecchiaia (Figura 7), si può ipotizzare che la naturale diffusione di Internet tra la popolazione potrebbe non essere abbastanza rapida, dato che ancora la 2014 il 47% della popolazione di 6 anni e più non usava Internet. Tale valore è preoccupante, anche alla luce obiettivi posti dall'Agenda Digitale Europea (Commissione Europea, 2010), che già per il 2015 prevedeva un uso regolare di Internet da parte del 75% della popolazione, superiore di ben 22 punti a quello effettivamente stimato per il nostro paese.

Per quanto l'età, anche dall'analisi della letteratura, risulti tra i fattori principali che spingono all'uso di Internet, si può ipotizzare che essa agisca congiuntamente con le altre variabili già citate.

Nella Figura 11 viene mostrato come si distribuisce l'utenza di Internet rispetto al livello di istruzione.

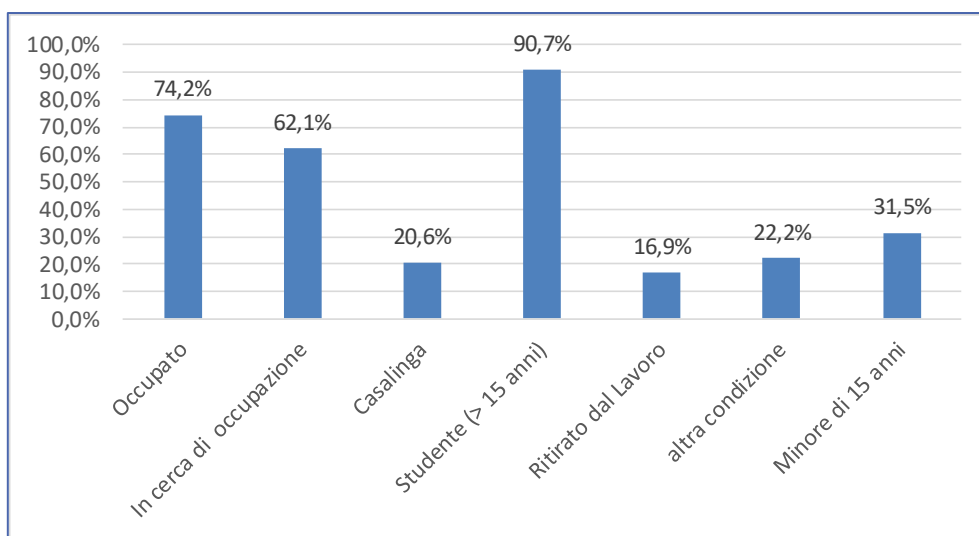
Figura 11 – Quota di utenti Internet per livello di istruzione (persone fuori dal percorso scolastico, anno 2014). Fonte: elaborazione su microdati Istat AVQ



Emerge dunque un'importante relazione tra l'uso di Internet e il livello di istruzione. In particolare, si noti la quota molto minoritaria (5% circa) di utenti tra le persone con licenza elementare o meno, che ricordiamo essere ancora il 21% circa (10 milioni di persone) della popolazione che ha terminato il proprio percorso di istruzione (cfr. Figura 8).

Per quel che riguarda la condizione professionale, di seguito è riportata la quota di utenti regolari per ciascuna categoria di questa variabile, rilevata per le persone di 15 anni o più.

Figura 12 – Quota di utenti Internet in Italia per condizione professionale (persone di 15 anni o più, anno 2014). Fonte: elaborazione su microdati Istat AVQ

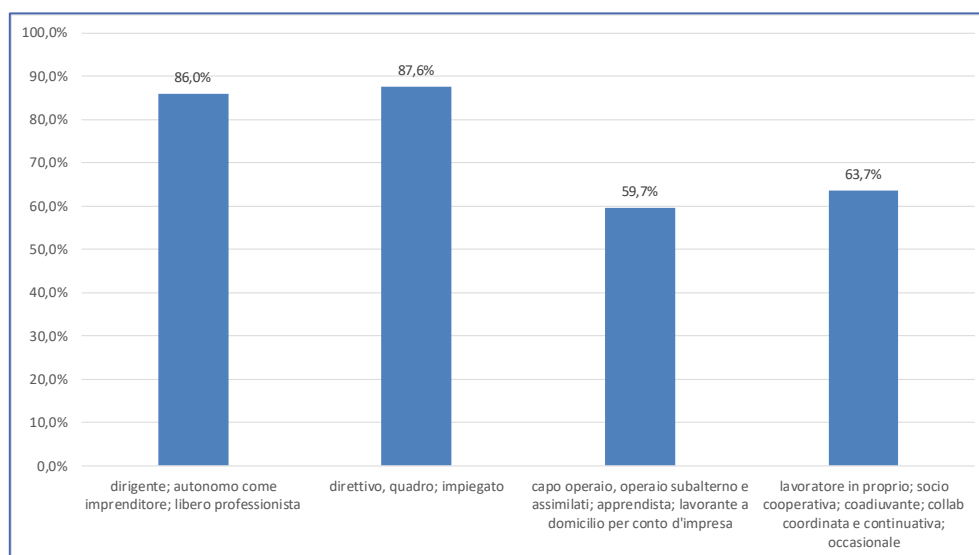


Dunque, sembrerebbe che anche l'essere parte attiva della società, dal punto di vista della condizione professionale, costituisce condizione fortemente abilitante all'uso di Internet.

Gli studenti, complice anche un'età in genere molto favorevole (cfr. Figura 10), presentano il tasso di utenza più elevato, pari al 90,7%, gli occupati del 72,0% e i lavoratori in cerca di occupazione del 62,1%. Viceversa, le casalinghe sono utenti Internet solamente nel 20,6% dei casi, i pensionati nel 16,9%, gli inabili e le altre condizioni residuali nel 22,2%. Migliore la situazione che riguarda i minori di 15 anni, che sono utenti nel 31,5% dei casi.

Tra coloro che lavorano poi, si può ipotizzare che vi siano delle differenze a seconda della posizione occupata. Nella Figura 13 è rappresentata la quota di utenti per categoria professionale, come codificata da Istat, con riferimento ai soli occupati.

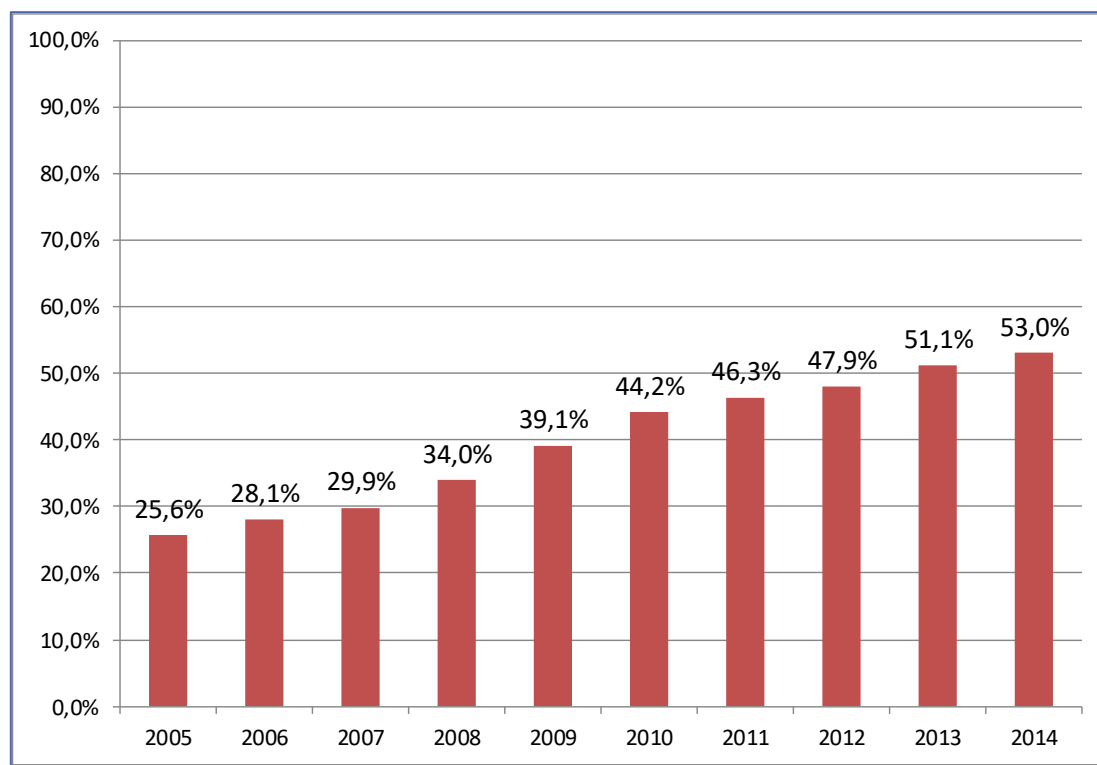
Figura 13 – Quota di utenti Internet in Italia per posizione professionale (occupati, anno 2014). Fonte: elaborazione su microdati Istat AVQ



Si può notare come le posizioni professionali che presentano tassi di utenza più elevati siano quelle che probabilmente utilizzano l'ICT per le attività lavorative, nonché quelle per ricoprire le quali normalmente è richiesto un titolo di studio non inferiore al diploma di scuola superiore.

Per valutare quale sia stata la dinamica di diffusione dell'uso regolare di Internet nella popolazione italiana, nella Figura 14 è riportato l'andamento della quota di utenti complessiva nel periodo 2005-2014.

Figura 14 – Quota di utenti Internet (almeno una volta a settimana) in Italia (popolazione di 6 anni o più, anni 2005-2014). Fonte: elaborazione su microdati Istat AVQ



Si è dunque sperimentata una crescita totale di poco più di 27 punti in 10 anni, con un aumento annuo medio pari a 2,7 punti percentuali. Considerati anche gli obiettivi dell'Agenda Digitale Europea, che ricordiamo prevedeva il 75% della popolazione utente di Internet al 2015, si può valutare tale crescita come troppo lenta.

A seguito della differente distribuzione degli utenti sulle variabili strutturali finora mostrate, si può ipotizzare che anche l'evoluzione nel tempo dell'uso di Internet in segmenti diversi della popolazione, individuata da una o più di tali variabili, sia stata differente, con un aumento della quota di utenti più veloce nei cluster che al momento mostrano tassi di adozione più elevati. In tal senso, alcuni segmenti particolarmente significativi possono essere, a scopo di esempio:

- 1) gli studenti (di 15 anni e più);
- 2) gli occupati;
- 3) le persone in cerca di occupazione;
- 4) le casalinghe che hanno lavorato in passato;
- 5) le casalinghe che non hanno mai lavorato;

6) i pensionati che vivono soli e quelli che vivono in famiglia.

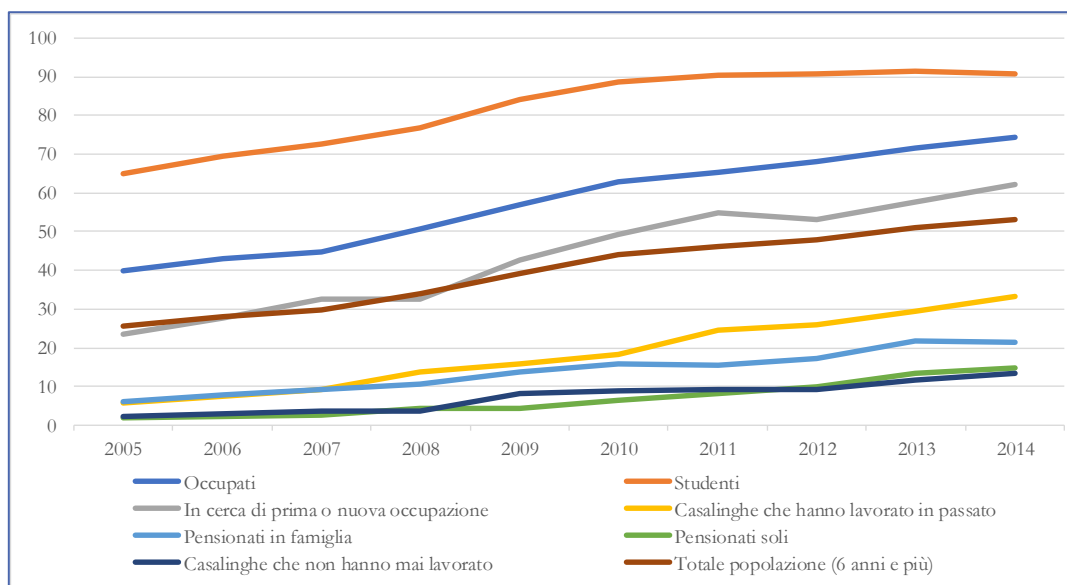
In Tabella 3 sono riportate le quote di utenza, divise per anno, per ciascun segmento individuato, nel periodo 2005-2014.

Tabella 3 - Utenza Internet per diversi segmenti significativi della popolazione (anni 2005-2014). Fonte: elaborazione su microdati Istat AVQ

| Segmento | Quota di utenti (%) | | | | | | | | | |
|---|---------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 |
| <i>Occupati</i> | 39,9 | 43,0 | 44,6 | 50,7 | 56,8 | 63,0 | 65,3 | 68,0 | 71,7 | 74,2 |
| <i>Studenti</i> | 64,9 | 69,6 | 72,5 | 76,7 | 84,1 | 88,7 | 90,3 | 90,7 | 91,4 | 90,7 |
| <i>In cerca di prima o nuova occupazione</i> | 23,6 | 27,8 | 32,4 | 32,6 | 42,7 | 49,4 | 54,9 | 53,1 | 57,6 | 62,1 |
| <i>Casalinghe che hanno lavorato in passato</i> | 5,9 | 7,5 | 9,4 | 13,7 | 15,8 | 18,2 | 24,6 | 26,0 | 29,5 | 33,4 |
| <i>Pensionati in famiglia</i> | 6,0 | 8,0 | 9,2 | 10,7 | 13,6 | 16 | 15,6 | 17,3 | 21,8 | 21,4 |
| <i>Pensionati soli</i> | 2,0 | 2,3 | 2,6 | 4,3 | 4,3 | 6,6 | 8,1 | 9,8 | 13,3 | 14,7 |
| <i>Casalinghe che non hanno mai lavorato</i> | 2,2 | 2,8 | 3,6 | 3,5 | 8,2 | 8,9 | 9,1 | 9,2 | 11,7 | 13,3 |
| Totale popolazione (6 anni e più) | <i>25,6</i> | <i>28,1</i> | <i>29,9</i> | <i>34,0</i> | <i>39,1</i> | <i>44,2</i> | <i>46,3</i> | <i>47,9</i> | <i>51,1</i> | <i>53,0</i> |

L'ipotesi per cui i diversi segmenti della popolazione presi in esame hanno avuto differenti tassi di crescita nel tempo sembra dunque essere confermata dai dati empirici. Di seguito si riporta il grafico relativo ai dati appena esposti, che visualizza in maniera migliore le differenze tra i diversi gruppi sociali.

Figura 15 - Utenza Internet per diversi segmenti significativi della popolazione (anni 2005-2014). Fonte: elaborazione su microdati Istat AVQ



Si può notare non solo come al 2005 i vari segmenti individuati avessero già tassi di adozione molto differenti (si va dal 2,0% dei pensionati soli al 64,9% degli studenti con 15 anni o più), ma anche come la crescita nell'utilizzo sia stata molto più rilevante per la gli studenti stessi, ed in generale anche per i segmenti che sono attivi nel mercato del lavoro, rispetto agli inattivi, con i pensionati soli e le casalinghe a contendersi l'ultimo posto. In particolare, la crescita media annua nel periodo considerato è rappresentata in Tabella 4, per i diversi segmenti individuati.

Tabella 4 - Crescita media annua dell'utenza regolare di Internet in Italia, in punti percentuali, per segmenti significativi della popolazione (anni 2005-2014). Fonte: elaborazione su microdati Istat AVQ

| Segmento | Tasso medio annuo |
|--|-------------------|
| Occupati | 3,4 |
| Studenti | 2,6 |
| In cerca di prima o nuova occupazione | 3,9 |
| Casalinghe che hanno lavorato in passato | 2,8 |
| Pensionati in famiglia | 1,5 |
| Pensionati soli | 1,3 |
| Casalinghe che non hanno mai lavorato | 1,1 |
| Totale popolazione (6 anni e più) | 2,7 |

Dunque, sono le persone in cerca di occupazione ad aver sperimentato la crescita più rapida, seguiti dagli occupati. Gli studenti e le casalinghe che hanno lavorato in passato presentano tassi simili, e sostanzialmente allineati alla media. Per i primi, si può ipotizzare che tale crescita

più rallentata rispetto ai segmenti della forza lavoro sia dovuta al raggiungimento di tassi molto vicini alla saturazione (sopra al 90% già a partire dal 2011). Invece, la crescita di questo particolare segmento di casalinghe è interessante, poiché pare testimoniare la capacità delle attività lavorative di incidere sull'abitudine all'uso di Internet anche una volta che si esce dalla forza lavoro. Si noti come i pensionati che vivono con componenti di altro tipo in famiglia evidenzino un tasso di crescita nella quota di utenti Internet solo di 0,2 punti percentuali annui superiore ai pensionati che vivono con altri pensionati o con casalinghe, pur mantenendo una quota totale di utenti di circa un 50% superiore a quella dei secondi. Ciò sembra rappresentare, in qualche modo, una prova a sfavore dell'ipotesi di un effetto "contagio" da parte di altri membri della famiglia, più favoriti sulla base delle variabili considerate.

Bisogna comunque considerare che le diverse variabili socio-economiche e demografiche fin qui analizzate sono tra di loro correlate, e quindi è ipotizzabile che esse influenzino congiuntamente l'uso di Internet. Utilizzando sempre i microdati dell'indagine "Aspetti della vita quotidiana", si possono incrociare le quattro variabili principali: età, titolo di studio, condizione professionale e posizione professionale. All'interno di ciascun segmento di popolazione individuato da ciascun livello di queste variabili, opportunamente dicotomizzate per favorire la leggibilità del fenomeno, è allora possibile valutare il tasso di uso di Internet, riportato in tabella seguente.

Tabella 5 - Quota di utenti di Internet (almeno una volta a settimana) all'interno di ciascun segmento individuato dalle quattro variabili principali (anno 2012). Fonte: elaborazione su microdati Istat AVQ

| | | | | Classe di età | | | | | |
|--------------------------|----------|-------------------------|---------------------|-------------------------|----------------------------|-------------------------|----------------------------|-------------------------|----------------------------|
| | | | | 15-24 | | 25-44 | | 45+ | |
| | | | | Titolo di studio | | Titolo di studio | | Titolo di studio | |
| | | | | <i>Laurea o Diploma</i> | <i>Scuola media o meno</i> | <i>Laurea o Diploma</i> | <i>Scuola media o meno</i> | <i>Laurea o Diploma</i> | <i>Scuola media o meno</i> |
| Condizione occupazionale | Attivo | Posizione professionale | <i>White collar</i> | 93% | 73% | 92% | 72% | 82% | 59% |
| | | | <i>Mai lavorato</i> | 91% | 83% | 81% | 33% | 58% | 21% |
| | | | <i>Blue collar</i> | 88% | 73% | 74% | 49% | 57% | 29% |
| | Inattivo | Posizione professionale | <i>White collar</i> | 100% | - | 74% | 55% | 42% | 20% |
| | | | <i>Mai lavorato</i> | 79% | 26% | 50% | 19% | 33% | 6% |
| | | | <i>Blue collar</i> | 72% | 37% | 58% | 34% | 34% | 7% |

Risulta evidente come i segmenti più in ritardo, dal punto di vista della diffusione di Internet, siano anche quelli in cui si trovano più facilmente le condizioni di marginalità nella società: età medio-alta (oltre i 45 anni), titolo di studio basso, inattivi che non hanno mai lavorato o hanno svolto lavori manuali.

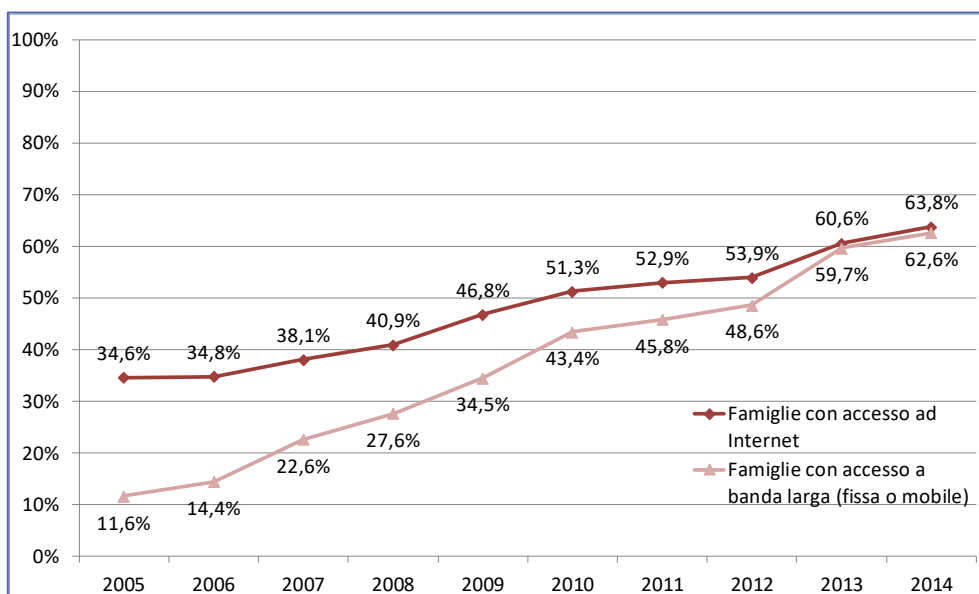
L'interazione tra queste variabili, e tra di esse e altre variabili di input, nello spiegare i fenomeni oggetto di studio, sarà indagata in maniera sensibilmente più diffusa nel corso di tutto il capitolo 6.

3.4 La diffusione delle connessioni Internet tra le famiglie

Come si diceva in apertura di capitolo 0, l'adozione di Internet in un dato paese può essere misurata su due livelli. Fin qui è stato analizzato il livello dell'uso individuale, regolare su base settimanale.

Passiamo ora alla diffusione delle connessioni presso le famiglie. Il grafico seguente mostra l'evoluzione nei tassi di penetrazione di due variabili: l'accesso a Internet da casa e l'adozione di una connessione a Internet a banda larga, fissa o mobile, componente della prima. Il periodo considerato è di nuovo l'intervallo 2005-2014.

Figura 16 - Diffusione ICT tra le famiglie italiane (anni 2005-2014). Fonte: elaborazione su dati Istat AVQ

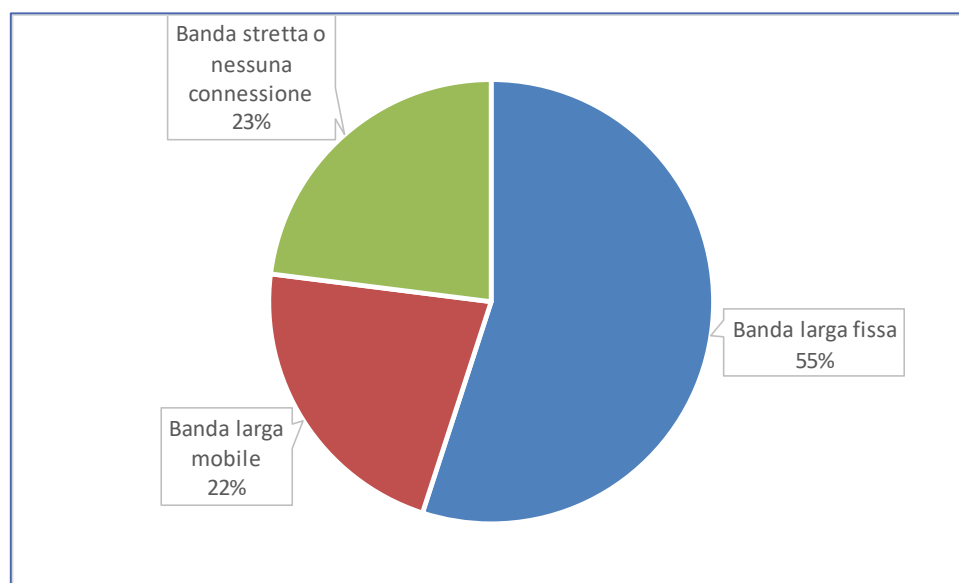


La crescita di famiglie con accesso ad Internet è quindi di circa 29 punti percentuali nel periodo di 10 anni considerato, con una crescita media annua di 2,9 punti. C'è da notare,

però, la più rapida evoluzione delle connessioni a banda larga, anche per effetto dell'introduzione sul mercato di quelle mobili, avvenuta tra la fine del 2009 e l'inizio del 2010. L'aumento totale nei 10 anni della quota di famiglie con accesso a banda larga è pari dunque a 51 punti percentuali, con aumento annuo di 5,1 punti, quasi doppio rispetto al tasso di crescita delle connessioni totali. Ciò implica una netta riduzione della quota di famiglie con connessione a banda stretta, che nel 2014 si attesta a solo l'1,2% del totale delle famiglie italiane, contro il dato di inizio periodo (2005), pari al 23,0%.

La distribuzione delle diverse tipologie di connessione tra le famiglie italiane, al 2016, risulta essere quella emersa durante l'analisi comparativa internazionale, e riassunta nel Figura 17.

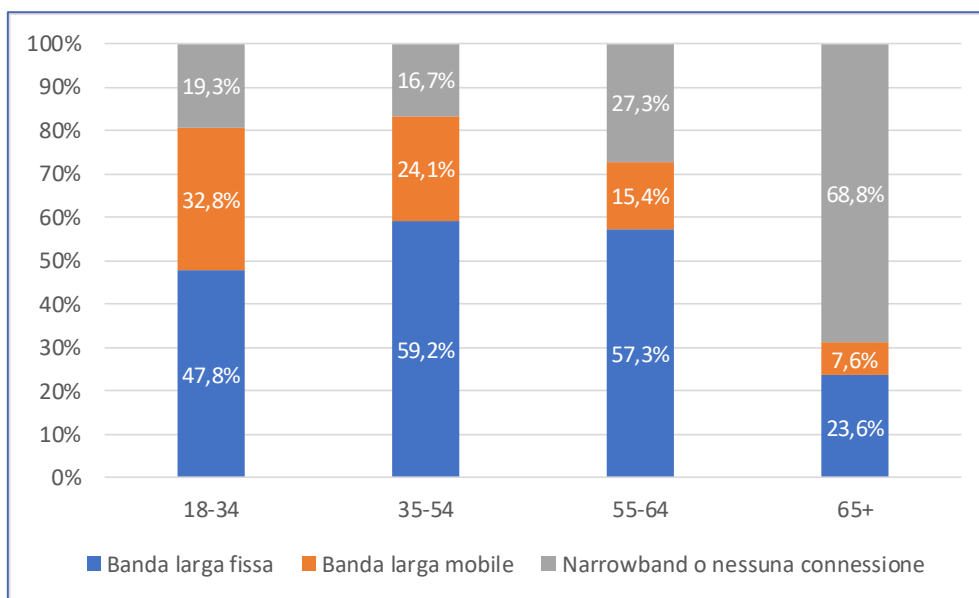
*Figura 17 - Diffusione delle diverse tecnologie di connessione Internet tra le famiglie italiane (anno 2016).
Fonte: Eurostat*



È possibile, così come fatto per gli individui, incrociare i dati di connessione delle famiglie con alcune variabili strutturali delle stesse. In particolare, poiché le variabili di struttura sono tutte individuali, può risultare interessante verificare la diffusione delle connessioni Internet, soprattutto quelle a banda larga, sulla base di alcune caratteristiche aggregate della famiglia o della Persona di Riferimento (figura simile al più noto “capofamiglia”). Tale analisi è possibile solo utilizzando i microdati del database Istat AVQ, il cui ultimo anno utile all’inizio delle elaborazioni, lo ricordiamo, era il 2014, mentre i dati Eurostat sono riferiti al 2016. Tuttavia, dati i relativamente bassi tassi di crescita appena esposti, si ritiene che un’analisi di questo tipo resti comunque significativa al momento della scrittura di questo testo.

Nel grafico seguente è mostrata la suddivisione delle connessioni per età della Persona di Riferimento (PR).

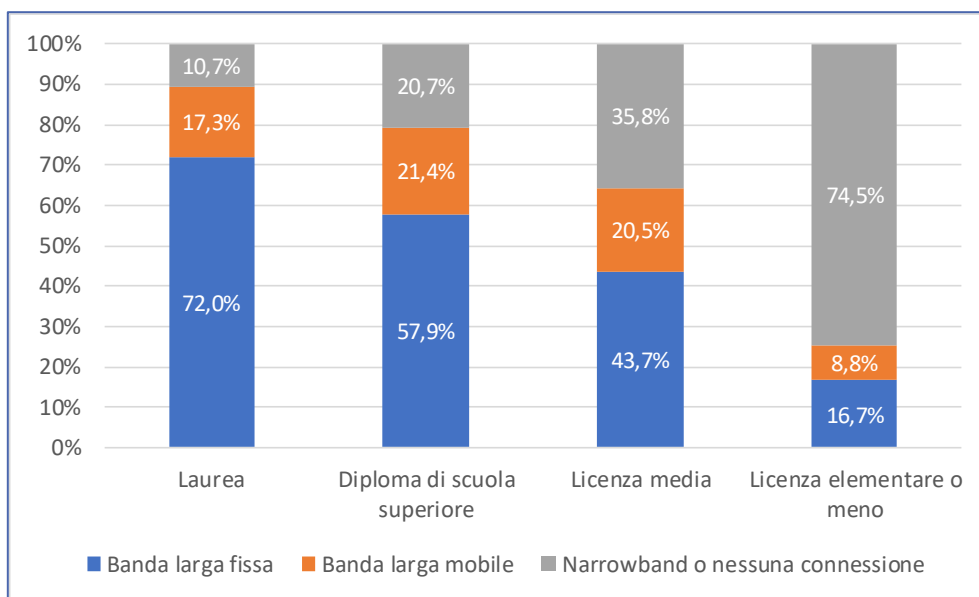
Figura 18 - Famiglie italiane per tipologia di connessione ad Internet e classe di età della Persona di Riferimento (anno 2014). Fonte: elaborazione su microdati Istat AVQ



Si evidenzia un tasso di connessioni a banda larga fissa più basso nella fascia di età della PR tra 18 e 34 anni, rispetto alle successive due classi di età intermedie. Ciò è probabilmente dovuto a questioni di reddito, che potrebbero essere più rilevanti per le famiglie con componenti relativamente più giovani. Tuttavia, tale dato è compensato dal 32,8% relativo alla diffusione della banda larga mobile nella stessa classe di età, valore più elevato su tutte le classi. Sembra evidenziarsi in questo caso un meccanismo di sostituzione fisso-mobile, in cui il secondo consente a famiglie giovani e quindi, presumibilmente, con reddito mediamente più basso, di utilizzare i servizi Internet contenendo la spesa annuale (Agcom, 2017). All'altro estremo, si noti come tra le famiglie con Persona di Riferimento ultrasessantacinquenne vi sia una scarsa diffusione delle connessioni a banda larga, con il 68,8% senza alcuna connessione, o con connessione a banda stretta.

Nel Figura 19 è riportata la suddivisione delle connessioni per titolo di studio della Persona di Riferimento.

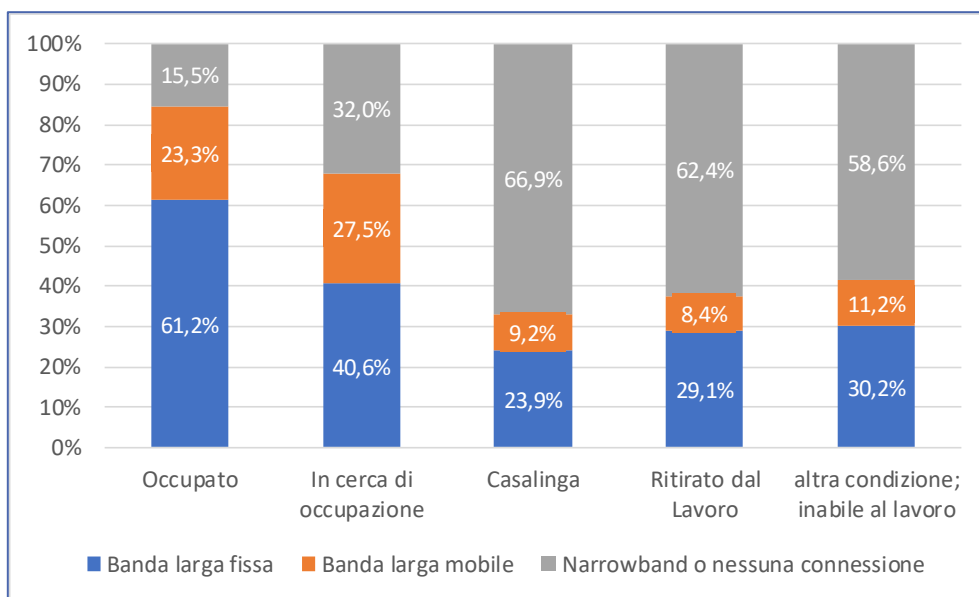
Figura 19 - Famiglie italiane per tipologia di connessione ad Internet e titolo di studio della Persona di Riferimento (anno 2014). Fonte: elaborazione su microdati Istat AVQ



L'andamento delle connessioni a banda larga ricalca sostanzialmente quello dell'uso di Internet per gli individui: le famiglie con Persona di Riferimento in possesso di titoli di studio più elevati sono anche quelle che presentano tassi di adozione maggiori, soprattutto per quello che riguarda le connessioni fisse. Per quanto riguarda invece le connessioni mobili, vi è una sostanziale parità tra i diversi titoli di studio, con eccezione dell'ultima categoria, licenza elementare o meno, in cui si registra il tasso più basso (9% circa), come accade peraltro con le connessioni fisse, con i tre quarti delle famiglie in questa categoria che non dispone di una connessione a banda larga.

Passando poi alla suddivisione delle connessioni per condizione professionale della Persona di Riferimento, la situazione delle famiglie italiane al 2014 è evidenziata nella Figura 20, in cui i pochi casi in cui la Persona di Riferimento risulta essere studente sono stati esclusi, poiché non rappresentativi.

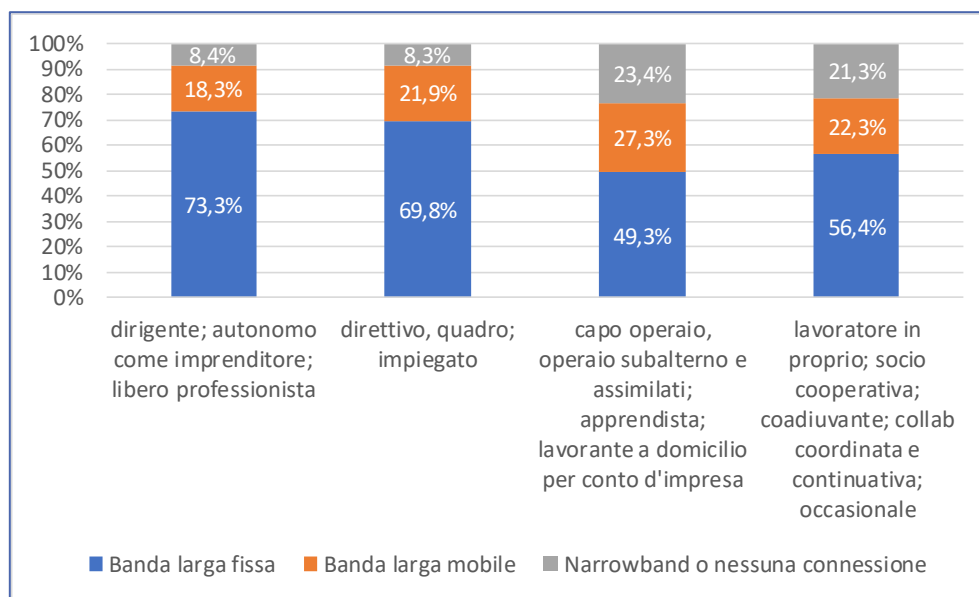
Figura 20 - Famiglie italiane per tipologia di connessione ad Internet e condizione professionale della Persona di Riferimento (anno 2014). Fonte: elaborazione su microdati Istat AVQ



In questo caso, così come avvenuto nell'uso di Internet da parte degli individui (cfr. Figura 12), si assiste ad una sostanziale bipartizione delle famiglie italiane, rispetto alla dotazione di una qualche connessione a banda larga, sia essa fissa o mobile: da un lato, le famiglie con Persona di Riferimento appartenente alla forza lavoro, occupata o in cerca di occupazione, mostrano tassi di adozione della banda larga più elevati, 84,5% e 68,0%, rispettivamente; dall'altro, le famiglie con Persona di Riferimento inattiva, con tassi di adozione sensibilmente inferiori: 41,4% per gli inabili o in altra condizione, 37,6% per i pensionati e addirittura 33,1% delle casalinghe.

Infine, come per il caso dell'uso di Internet da parte degli individui, nella Figura 21 è riportata la suddivisione delle connessioni ad Internet per la posizione nella professione della Persona di Riferimento, per le sole famiglie in cui tale persona risulta occupata.

Figura 21 - Famiglie italiane per tipologia di connessione ad Internet e condizione professionale della Persona di Riferimento (solo famiglie con PR occupata, anno 2014). Fonte: elaborazione su microdati Istat AVQ



Questi dati possono essere letti raffrontandoli alla media delle famiglie con Persona di Riferimento occupata, che ricordiamo è di 61,2% di connessioni a banda larga fissa, 23,3% di connessioni mobili e 15,5% di connessioni a banda stretta o nessuna connessione. Anche in questo caso, dunque, si ha una sostanziale bipartizione: da un lato quelli che in inglese vengono comunemente definiti come *white collar*, cioè i ruoli dirigenziali e impiegatizi, con quote di connessione a banda larga fissa ben al di sopra della media (+12,1 e +8,6 punti, rispettivamente) e tassi di banda stretta o non adozione pari a quasi la metà della media, in entrambe le categorie; dall'altro, le famiglie con Persona di Riferimento *blue collar*, che svolge cioè ruoli manuali, che presentano tassi di non adozione (o di banda stretta) molto superiori alla media (+7,9 punti percentuali per gli operai e simili e +5,8 per i lavoratori autonomi e assimilati). Le connessioni mobili sono sostanzialmente in linea con la media per gli impiegati e i lavoratori autonomi, mentre sono meno diffuse nelle famiglie con Persona di Riferimento dirigente, imprenditore o libero professionista (-5 punti percentuali rispetto alla media) e più diffuse per gli operai e simili (+4 punti). Questo risultato sembra confermare l'ipotesi, avanzata durante l'analisi della suddivisione delle connessioni per classe di età della Persona di Riferimento, che le connessioni mobili siano preferite a quelle fisse da utenti con redditi più bassi.

3.5 Le motivazioni del non uso⁶

Dato il quadro fin qui dipinto dalle statistiche descrittive presentate, è importante indagare anche le motivazioni della non adozione di una connessione Internet a casa, fornite dalle famiglie. Di seguito è riportata la distribuzione di tali motivazioni per i diversi paesi europei.

Tabella 6 - Motivi della non adozione di una connessione Internet da parte delle famiglie (percentuale di tutte le famiglie che non adottano) (anno 2015). Fonte: Eurostat

| Paese | Mancanza di capacità e costi troppo elevati (esclusi i non interessati) | Tra l'altro, per costi troppo elevati | Solo per mancanza di interesse | Solo per altre ragioni | Solo per mancanza di capacità | Solo per preoccupazioni riguardanti la privacy | Solo per costi troppo elevati |
|---------------------------|---|---------------------------------------|--------------------------------|------------------------|-------------------------------|--|-------------------------------|
| Austria | 2,4% | 12,0% | 57,4% | 4,0% | 5,2% | 1,1% | 2,1% |
| Belgio | 6,7% | 33,9% | 24,1% | 11,3% | 11,9% | 1,2% | 15,1% |
| Bulgaria | 12,7% | 38,5% | 21,6% | 4,5% | 20,6% | 0,1% | 14,1% |
| Croazia | 3,3% | 39,6% | 22,7% | 0,9% | 1,4% | 0,6% | 7,8% |
| Cipro | 9,0% | 51,9% | 15,4% | 3,1% | 6,9% | 0,2% | 17,1% |
| Repubblica Ceca | 6,0% | 34,1% | 38,1% | 3,1% | 5,4% | 0,1% | 12,8% |
| Danimarca | 2,9% | 13,0% | 24,5% | 12,1% | 14,1% | 1,5% | 7,4% |
| Estonia | 5,0% | 26,3% | 46,4% | 1,1% | 6,1% | 0,3% | 7,1% |
| Unione Europea (28 paesi) | 6,6% | 32,7% | 23,0% | 6,2% | 15,7% | 1,5% | 13,8% |
| Finlandia | 2,5% | 27,3% | 16,6% | 1,4% | 3,6% | 2,9% | 0,8% |
| Macedonia | 7,4% | 33,5% | 34,3% | 8,8% | 11,2% | 0,2% | 13,5% |
| Francia | 8,0% | 41,8% | 12,6% | 15,5% | 11,3% | 5,4% | 21,7% |
| Germania | 5,8% | 35,6% | 26,0% | 1,4% | 4,7% | 1,9% | 11,1% |
| Grecia | 9,3% | 24,5% | 14,3% | 5,1% | 40,5% | 0,8% | 12,4% |
| Ungheria | 10,0% | 60,0% | 13,5% | 3,9% | 2,9% | 0,0% | 10,8% |
| Irlanda | 4,8% | 20,8% | 26,9% | 8,2% | 23,8% | 0,1% | 9,7% |
| Italia | 4,0% | 21,3% | 16,6% | 5,0% | 35,0% | 1,7% | 14,5% |
| Lettonia | 7,1% | 26,7% | 29,3% | 1,8% | 7,5% | 0,0% | 9,1% |
| Lituania | 8,9% | 28,1% | 46,3% | 0,4% | 7,7% | 0,2% | 9,0% |
| Lussemburgo | 5,5% | 16,5% | 44,6% | 5,9% | 5,5% | 5,5% | 5,5% |
| Malta | 3,1% | 16,1% | 11,0% | 3,2% | 31,0% | 0,0% | 3,6% |
| Paesi Bassi | 2,5% | 26,1% | 34,8% | 13,0% | 9,8% | 2,8% | 1,9% |
| Norvegia | 0,0% | 21,3% | 24,2% | 27,0% | 2,0% | 6,5% | 13,0% |

⁶ Questa analisi è stata pubblicata da Claudio Leporelli (DIAG -Sapienza) e Cosimo Dolente sulla rivista online AgendaDigitale.eu (<http://www.agendadigitale.eu>).

| Paese | Mancanza di capacità e costi troppo elevati (esclusi i non interessati) | Tra l'altro, per costi troppo elevati | Solo per mancanza di interesse | Solo per altre ragioni | Solo per mancanza di capacità | Solo per preoccupazioni riguardanti la privacy | Solo per costi troppo elevati |
|-------------|---|---------------------------------------|--------------------------------|------------------------|-------------------------------|--|-------------------------------|
| Polonia | 7,8% | 31,4% | 28,6% | 4,1% | 11,7% | 0,3% | 9,4% |
| Portogallo | 7,8% | 43,1% | 8,0% | 0,9% | 30,1% | 0,3% | 12,4% |
| Romania | 13,6% | 45,5% | 19,9% | 4,9% | 14,7% | 0,0% | 20,7% |
| Serbia | 2,0% | 32,6% | 46,3% | 2,7% | 3,4% | 0,2% | 12,3% |
| Slovacchia | 4,2% | 18,4% | 23,7% | 4,6% | 16,8% | 0,9% | 5,3% |
| Slovenia | 4,5% | 32,5% | 34,7% | 12,7% | 5,1% | 1,0% | 10,2% |
| Spagna | 4,4% | 33,5% | 31,1% | 3,2% | 6,9% | 0,2% | 14,2% |
| Svezia | 2,5% | 18,9% | 24,2% | 9,7% | 2,2% | 0,7% | 2,1% |
| Turchia | 5,5% | 41,3% | 19,4% | 0,6% | 10,0% | 0,4% | 16,1% |
| Regno Unito | 3,3% | 24,2% | 31,3% | 12,1% | 13,4% | 0,9% | 14,1% |

Come si vede, il motivo principale della mancata adozione da parte delle famiglie italiane è la mancanza di capacità (35% delle famiglie senza accesso ad Internet a casa). Tuttavia, molto elevata risulta essere anche la quota di famiglie che segnalano il fattore di costo. Esso compare tra le motivazioni nel 21% dei casi, ed è una motivazione esclusiva nel 14,5%. Anche la mancanza di interesse gioca un ruolo importante, essendo segnalata come unico motivo di non adozione da parte del 16,6% delle famiglie senza accesso.

Questi dati però non dicono tutto, per diverse ragioni. In primo luogo, si riferiscono a risposte fornite da famiglie prive di un accesso a casa, ma bisogna tener presente che alcuni componenti della famiglia potrebbero usare Internet altrove (questo succede in Italia per il 3% delle famiglie, una su otto di quelle che non hanno accesso a casa); viceversa non rispondono a questa domanda i tantissimi cittadini che dispongono di un accesso a casa, ma non usano Internet personalmente. In secondo luogo, esistono famiglie che non specificano alcun motivo, nemmeno tra i più importanti (vedremo che ciò è collegato al fatto che molti individui dichiarano, all'interno della stessa indagine, di non sapere cosa sia Internet). Si tratta del 14,2% delle famiglie italiane, contro il 7,0% della media UE-28.

Nella Tabella 7 mostriamo come, in questa triste classifica, l'Italia sia in posizione migliore di soli 4 paesi, nell'ordine: Grecia, Bulgaria, Romania, e Portogallo; ai primi posti troviamo invece i Paesi Bassi, il Lussemburgo, la Svezia, la Finlandia e la Norvegia. La nuova tabella

presenta tutti i dati elaborati da Eurostat rispetto al 2016, circa le ragioni per cui una famiglia non si è dotata di un accesso Internet a casa, incluse quelle meno rilevanti. I dati vengono presentati per l'Italia, e per l'UE a 15 e a 28. Per dare un'idea dello spettro dei valori in gioco, senza perdere in leggibilità, viene inoltre riportata, per ogni domanda, la percentuale massima riscontrata rispettivamente per i quattro Paesi più "deboli", in questo, dell'Italia, e per i cinque Paesi sopra citati che appaiono, in questo, i più forti dell'UE. Una differenza importante rispetto alla tabella precedente è che qui le percentuali sono calcolate sul totale delle famiglie di ciascun paese e non sul numero di quelle che non dispongono di un accesso a Internet. Ciò consente di misurare la gravità del problema, ai fini del perseguimento degli obiettivi dell'Agenda Digitale Europea, e di confrontare paesi con diverso grado di penetrazione dell'accesso.

Tabella 7 - Analisi dei motivi della non adozione di una connessione Internet a casa da parte delle famiglie (anno 2015). Fonte: Eurostat

| Ragioni per la non adozione di connessione Internet | Quota massima tra gli ultimi 4 paesi | Italia | EU-28 | EU-15 | Quota massima tra i primi 5 paesi |
|--|---|---------------|--------------|--------------|--|
| <i>Una qualsiasi ragione (escluso "non serve")</i> | 20,0% | 14,2% | 7,0% | 6,5% | 1,5% |
| <i>Mancanza di capacità</i> | 20,9% | 11,3% | 7,0% | 5,7% | 5,7% |
| <i>Solo mancanza di capacità</i> | 12,5% | 8,5% | 2,7% | 2,6% | 0,4% |
| <i>Mancanza di capacità e costi troppo elevate (escluso "non serve")</i> | 5,2% | 1,0% | 1,1% | 0,8% | 0,2% |
| <i>Non serve (non è utile, non è interessante, ecc.)</i> | 17,7% | 6,0% | 7,9% | 6,3% | 7,6% |
| <i>Solo non serve (non è utile, non è interessante, ecc.)</i> | 8,7% | 4,0% | 3,9% | 3,3% | 2,0% |
| <i>Costi troppo elevati</i> | 15,6% | 5,2% | 5,6% | 4,6% | 2,7% |
| <i>Solamente costi troppo elevati</i> | 6,5% | 3,5% | 2,4% | 2,1% | 0,4% |
| <i>Costo di accesso troppo elevato</i> | 11,9% | 3,4% | 3,9% | 3,3% | 2,3% |
| <i>Costo degli strumenti troppo elevato</i> | 12,1% | 3,0% | 4,5% | 3,6% | 2,4% |
| <i>Accede altrove</i> | 3,2% | 3,0% | 2,0% | 2,0% | 2,4% |
| <i>Altre ragioni</i> | 2,5% | 1,4% | 1,9% | 1,8% | 1,1% |
| <i>Solo altre ragioni</i> | 1,8% | 1,2% | 1,1% | 1,1% | 0,9% |
| <i>Preoccupazione per la privacy o la sicurezza</i> | 4,0% | 0,8% | 1,6% | 1,7% | 2,6% |
| <i>Solo preoccupazioni per la privacy e la sicurezza</i> | 0,3% | 0,4% | 0,3% | 0,3% | 0,3% |
| <i>Banda larga non disponibile</i> | 1,0% | 0,6% | 0,5% | 0,5% | 0,6% |

| Ragioni per la non adozione di connessione Internet | Quota massima tra gli <u>ultimi</u> 4 paesi | Italia | EU-28 | EU-15 | Quota massima tra i <u>primi</u> 5 paesi |
|---|---|--------|-------|-------|--|
| Somma delle righe non sovrapposte | 46,6% | 35,1% | 19,6% | 18,2% | 6,7% |

La mancanza di capacità, citata come unico motivo dall'8,5% delle famiglie italiane, e anche insieme ad altri motivi dall'11,3% di esse, è il problema più grave, seguito dalla mancanza di interesse, citato da solo dal 4,0% delle famiglie, e insieme ad altri motivi dal 6,0%, e dal costo (3,5% da soli, 5,2% con altri motivi).

È interessante notare come risulti essere molto bassa la percentuale delle famiglie che cita come motivo la mancanza di copertura a banda larga per la propria abitazione. Tuttavia, occorre tener presente che questo impedimento è conosciuto soprattutto dalle famiglie che hanno valutato approfonditamente la scelta di dotarsi di un accesso e hanno dunque riscontrato questo problema.

Rispetto poi all'uso di Internet, è utile capire più in dettaglio le ragioni del non uso da parte degli individui, e analizzarle in base alle loro caratteristiche socio-economiche e demografiche. Questa analisi è svolta, in riferimento alla sola Italia, utilizzando i microdati della edizione 2014 dell'indagine Istat "Aspetti della Vita Quotidiana" (cfr. paragrafo 3.1). In questo caso, utilizziamo le risposte di ogni individuo che si è dichiarato non utilizzatore di Internet, in nessun luogo, a prescindere dalla disponibilità di un accesso a casa.

Nella Tabella 8 sono presentate le risposte fornite alle quattro domande che sono risultate più rilevanti. Le risposte a quesiti simili riguardanti il costo e l'interesse/utilità dell'uso sono state aggregate, perché affini. Inoltre, sono stati esclusi: da un lato, i ragazzi fino a 14 anni, per i quali la risposta largamente prevalente era il divieto da parte dei genitori (circa un milione di persone stimate); dall'altro, gli anziani con almeno 75 anni, stimati in circa 6.270.000, non utenti al 95%, che in circa il 47% dei casi rispondono di non sapere cosa sia Internet. Sulla diagonale della tabella è stimato il numero di individui che fornisce una sola risposta; nelle altre caselle è stimato il numero di persone nell'universo corrispondente a chi fornisce le risposte indicate in riga e colonna. Per ragioni di leggibilità, non sono visualizzate le consistenze dei *cluster* di coloro che segnalano 3 di questi motivi principali, o tutti e 4.

Tabella 8 – Persone che segnalano le motivazioni principali del non uso di Internet da parte degli individui (anno 2014). Fonte: elaborazione su microdati Istat AVQ

| | Non so cosa sia | Costa troppo | Non interessa | Non so usarlo |
|------------------------|------------------------|---------------------|----------------------|----------------------|
| Non so cosa sia | 1.848.683 | | | |
| Costa troppo | 157.777 | 1.417.266 | | |
| Non interessa | 347.269 | 453.544 | 4.271.946 | |
| Non so usarlo | 149.676 | 358.053 | 948.408 | 1.811.787 |

Osserviamo intanto che, a differenza di quanto accadeva per le famiglie senza accesso Internet da casa, il numero degli individui che dichiarano di essere disinteressati a Internet è assai più alto di quello di coloro che dichiarano di non avere le capacità per utilizzare Internet. Una spiegazione può risiedere nel fatto che molti di coloro che esprimono disinteresse dispongono in realtà di un accesso domestico utilizzato da altri componenti della famiglia, e quindi non contribuiscono alla misura precedente. Il numero stimato di individui tra 15 e 74 anni che rispondono di non sapere cosa sia Internet, o di non essere interessati all'uso, o di non saperla usare, pur disponendo di una connessione domestica, è dunque stimato in 5,2 milioni di individui.

Di seguito è riportata un'analisi delle risposte dagli individui che forniscono una delle 4 risposte principali, da sola o in combinazione con qualcun'altra delle stesse 4. Il numero delle osservazioni così selezionate è rappresentativo di una popolazione di 12,2 milioni di persone tra 15 e 74 anni, il 94% del totale dei non utenti rispondenti. I motivi del mancato utilizzo sono messi in relazione, separatamente, con l'età, il titolo di studio, la condizione e la posizione professionale. La Tabella 9 riporta i valori relativi alle classi di età.

Tabella 9 - Suddivisione per classi di età delle motivazioni principali al non uso di Internet da parte degli individui (anno 2014). Fonte: elaborazione su microdati Istat AVQ

| Classe di età | Non so cosa sia | | Costa troppo | | Non interessa | | Non so usarlo | |
|----------------------|----------------------------------|------------------------------------|----------------------------------|------------------------------------|----------------------------------|------------------------------------|----------------------------------|------------------------------------|
| | <i>Unica risposta principale</i> | <i>Insieme ad altre principali</i> | <i>Unica risposta principale</i> | <i>Insieme ad altre principali</i> | <i>Unica risposta principale</i> | <i>Insieme ad altre principali</i> | <i>Unica risposta principale</i> | <i>Insieme ad altre principali</i> |
| 15-24 | 1,4% | 1,1% | 2,2% | 6,2% | 1,4% | 1,0% | 1,1% | 1,1% |
| 25-34 | 2,2% | 2,4% | 8,5% | 13,9% | 4,3% | 5,0% | 4,4% | 2,9% |
| 35-44 | 7,1% | 6,3% | 14,0% | 21,1% | 9,9% | 15,2% | 9,5% | 10,0% |

| Classe di età | Non so cosa sia | | Costa troppo | | Non interessa | | Non so usarlo | |
|---------------|---------------------------|-----------------------------|---------------------------|-----------------------------|---------------------------|-----------------------------|---------------------------|-----------------------------|
| | Unica risposta principale | Insieme ad altre principali | Unica risposta principale | Insieme ad altre principali | Unica risposta principale | Insieme ad altre principali | Unica risposta principale | Insieme ad altre principali |
| 45-54 | 15,7% | 13,8% | 15,0% | 18,5% | 20,0% | 21,6% | 19,4% | 20,3% |
| 55-64 | 26,3% | 26,3% | 25,4% | 17,4% | 27,9% | 27,3% | 27,5% | 27,3% |
| 65-74 | 47,4% | 50,2% | 34,9% | 22,8% | 36,4% | 29,8% | 38,0% | 38,4% |
| Totale | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% |

Come era logico attendersi, sono soprattutto gli anziani, anche sotto i 75 anni, a non sapere cosa sia Internet. Le percentuali riscontrate per questa motivazione nelle fasce di età più bassa derivano invece da un numero modesto di casi osservati e meriterebbero ulteriori indagini *ad hoc*. In queste indagini, infatti, sono possibili errori di risposta ma, anche presupponendo che le risposte siano state fornite in maniera corretta, il soggetto presente nel campione potrebbe avere caratteristiche individuali molto specifiche che non lo rendono rappresentativo. Anche considerando affidabile tale valore, andrebbe sicuramente indagata l'estensione di fenomeni di emarginazione e degrado che potrebbero coinvolgere una porzione consistente delle classi più giovani.

Ancora, si può notare come il costo eccessivo di utilizzo assume rilevanza maggiore nelle fasce di età più basse, in special modo tra i 15 e i 44 anni. Invece, il disinteresse, inteso anche come mancata percezione di utilità, è la ragione di mancato utilizzo percentualmente più rilevante tra 25 e 64 anni. La mancanza di *skill* è infine molto diffusa soprattutto a partire dalla fascia di età 45-55 anni.

Passando al titolo di studio, nella Tabella 10 sono considerati separatamente i soggetti in formazione, a tutti i livelli. Per essi vale il caveat sopra espresso circa il ridotto numero di casi osservati, tuttavia resta confermato che per i giovani la principale motivazione per il mancato utilizzo risulta essere di natura economica.

Tabella 10 - Suddivisione per titolo di studio delle motivazioni principali al non uso di Internet da parte degli individui (anno 2014). Fonte: elaborazione su microdati Istat AVQ

| Titolo di studio | Non so cosa sia | | Costa troppo | | Non interessa | | Non so usarlo | | |
|-------------------|---------------------------|-----------------------------|---------------------------|-----------------------------|---------------------------|-----------------------------|---------------------------|-----------------------------|-------|
| | Unica risposta principale | Insieme ad altre principali | Unica risposta principale | Insieme ad altre principali | Unica risposta principale | Insieme ad altre principali | Unica risposta principale | Insieme ad altre principali | |
| Non in formazione | Laurea | 0,9% | 1,1% | 2,4% | 3,3% | 1,9% | 3,2% | 1,6% | 1,7% |
| | Diploma | 11,8% | 8,8% | 16,7% | 24,2% | 16,3% | 22,7% | 13,9% | 16,8% |

| Titolo di studio | Non so cosa sia | | Costa troppo | | Non interessa | | Non so usarlo | |
|----------------------|---------------------------|-----------------------------|---------------------------|-----------------------------|---------------------------|-----------------------------|---------------------------|-----------------------------|
| | Unica risposta principale | Insieme ad altre principali | Unica risposta principale | Insieme ad altre principali | Unica risposta principale | Insieme ad altre principali | Unica risposta principale | Insieme ad altre principali |
| Licenza media | 37,5% | 32,9% | 42,8% | 47,4% | 42,6% | 47,4% | 43,0% | 43,4% |
| Licenza elementare | 42,3% | 48,0% | 32,4% | 17,2% | 34,1% | 24,1% | 37,2% | 34,2% |
| Nessun titolo | 6,5% | 8,9% | 4,5% | 3,7% | 4,3% | 2,0% | 3,7% | 3,7% |
| In formazione | 0,9% | 0,3% | 1,3% | 4,2% | 0,7% | 0,6% | 0,6% | 0,2% |
| Totale | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% |

Anche nel caso dei laureati, il numero assoluto stimato è ridotto, ma il loro comportamento appare non dissimile da quello dei diplomati. Per queste due classi, nel loro complesso, il disinteresse appare la motivazione residuale più probabile, dando per scontato che, a questo livello di istruzione, non si possa ignorare il fenomeno Internet, si disponga delle risorse economiche necessarie all'utilizzo (almeno mobile) e si abbia attitudine ad acquisire le skill necessarie. I dati indicano che in effetti nel disinteresse il peso di queste categorie è tra i maggiori, ma esse appaiono rilevanti anche rispetto alle motivazioni di costo e alla mancanza di skill. Resta confermato che un basso titolo di studio rende ancora più probabile che si ignori l'esistenza di Internet (circa il 90% dei rispondenti, sia come solo motivo che insieme ad altri). Più in generale, i dati sembrano confermare le intuizioni emerse nei paragrafi precedenti, che indicano nel livello di istruzione, insieme all'età, una delle cause più importanti di *divide* del paese.

Nella Tabella 11, relativa all'occupazione, vengono definiti *white collar* gli imprenditori, liberi professionisti, dirigenti, quadri e impiegati (nell'attuale posto di lavoro o nell'ultima posizione ricoperta prima di andare in pensione); vengono definiti *blue collar* gli operai, i lavoratori autonomi e coadiuvanti, i soci di cooperativa, i lavoratori occasionali o a progetto. In questa ampia categoria ricadono quindi molte posizioni "precarie" non necessariamente assimilabili a quelle operaie o del lavoro autonomo. Questa suddivisione, già introdotta brevemente nei paragrafi precedenti, sarà utilizzata sovente anche nel seguito del testo.

Tabella 11 - Suddivisione per condizione e posizione professionale delle motivazioni principali al non uso di Internet da parte degli individui (anno 2014). Fonte: elaborazione su microdati Istat AVQ

| Condizione e posizione professionale | Non so cosa sia | | Costa troppo | | Non interessa | | Non so usarlo | |
|--------------------------------------|---------------------------|-----------------------------|---------------------------|-----------------------------|---------------------------|-----------------------------|---------------------------|-----------------------------|
| | Unica risposta principale | Insieme ad altre principali | Unica risposta principale | Insieme ad altre principali | Unica risposta principale | Insieme ad altre principali | Unica risposta principale | Insieme ad altre principali |
| Occupato White Collar | 2,0% | 2,0% | 3,4% | 6,4% | 2,4% | 6,5% | 2,1% | 4,9% |
| Occupato Blue Collar | 14,9% | 14,0% | 21,5% | 26,5% | 21,0% | 26,5% | 21,2% | 21,1% |
| Pensionato White Collar | 6,9% | 7,2% | 8,7% | 9,0% | 8,8% | 8,8% | 8,4% | 8,2% |
| Pensionato Blue Collar | 33,1% | 33,8% | 22,8% | 15,6% | 24,7% | 19,7% | 25,5% | 25,8% |
| Casalinga | 30,1% | 33,9% | 24,8% | 19,6% | 29,8% | 29,1% | 30,4% | 31,9% |
| In cerca di nuova occupazione | 10,8% | 7,7% | 15,5% | 17,7% | 11,0% | 8,2% | 10,7% | 7,2% |
| In cerca di prima occupazione | 2,2% | 1,3% | 3,2% | 5,2% | 2,4% | 1,2% | 1,7% | 0,8% |
| Totale | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% |

I dati mostrano come la distinzione tra *white* e *blue collar*, cui possiamo assimilare le casalinghe, risulta essere un altro *divide* fondamentale della società italiana, confermando anche qui le intuizioni precedenti. Analizzando le motivazioni, appare evidente come la maggioranza assoluta di coloro che segnalano cosa non sia Internet, da sola o con altre motivazioni, è composta da pensionati *blue collar* o casalinghe. Percentuali importanti, per questi due gruppi, si notano anche per tutte le altre motivazioni. Come era lecito attendersi, i disoccupati, sia in cerca di nuova che di prima occupazione, hanno un peso più rilevante per quel che riguarda le motivazioni di costo.

L'analisi descrittiva qui presentata consente dunque di realizzare una prima valutazione della questione della diffusione delle connessioni Internet tra le famiglie e dell'uso di Internet tra gli individui in Italia, nonché dei motivi del non uso. Tuttavia, al fine di comprenderne a fondo le dinamiche, è opportuno utilizzare metodologie più complesse, in grado di stabilire l'impatto relativo di ciascuna variabile al netto delle altre, e di prendere in considerazione un numero maggiore di variabili. Una descrizione delle metodologie individuate è fornita nel capitolo 4. Attraverso di queste sono svolte poi le analisi del capitolo 6, a seguito di una nuova definizione di digital divide in quattro stadi, proposta nel capitolo 5.

4 Metodologie per l'analisi dei microdati

La struttura dei database legati all'indagine annuale "Aspetti della Vita Quotidiana" di Istat, è di tipo microdati (cfr. paragrafo 3.1). Ciò significa che ciascun *dataset* contiene l'intero insieme di risposte di ciascun intervistato, per ogni anno, a meno di eliminazioni effettuate dall'Istat per motivi di privacy o per inconsistenze nei dati. La disponibilità di microdati apre alla possibilità di svolgere una pluralità di analisi, utilizzando metodi statistici e matematici molto differenti. La scelta del metodo specifico da utilizzare per l'analisi della domanda è legata a differenti aspetti, tra cui la tipologia delle variabili a disposizione (categoriale, ordinale, continua) e lo specifico problema che si vuole analizzare (per una trattazione diffusa delle diverse tecniche disponibili si veda, ad esempio, (Agresti, 2007)).

In particolare, all'interno di questo studio si ha l'obiettivo di indagare diversi aspetti della domanda di Internet, e in particolare della banda larga, in Italia. Le variabili obiettivo sono tutte di natura categoriale (cfr. Tabella 2), sia per quanto riguarda l'uso da parte degli individui che le connessioni delle famiglie. Lo stesso dicasi per la maggior parte delle variabili indipendenti che si può ipotizzare di utilizzare, a iniziare da quelle di segmentazione (cfr. Tabella 1).

Le seguenti tecniche sono dunque state identificate come utili e utilizzate nelle analisi delle determinanti della domanda del successivo capitolo 6, date le caratteristiche del database di analisi e del fenomeno oggetto di esame:

- 1) Regressione logistica (paragrafo 4.1);
- 2) Algoritmi di analisi logica dei dati (LAD) (paragrafo 4.2);
- 3) Alberi di classificazione (paragrafo 4.3);
- 4) Analisi delle classi latenti (paragrafo 4.4).

Ognuna di tali tecniche presenta dei vantaggi e degli svantaggi nell'utilizzo, e fornisce risultati di natura differente.

In particolare, la regressione logistica è la prima tecnica individuata, molto utilizzata in letteratura, laddove si deve indagare un'uscita dicotomica o categoriale (ma ne esiste anche una versione ordinale) sulla base di un insieme di variabili di input anch'esse categoriali o continue. Essa verrà utilizzata nella prima modellizzazione del problema dell'adozione della tecnologia (paragrafo 6.1.1).

Come si vedrà nel seguito, è poi interessante considerare un numero ampio di variabili indipendenti, per cercare di valutarne l'impatto sulle variabili di uscita, e cercare associazioni tra variabili che non siano prevedibili a priori. A tal fine è stata individuata la tecnica di classificazione SLAD, opportunamente adattata al problema in esame attraverso la proposta di due nuove procedure, originali del progetto di ricerca (paragrafo 4.2.2). Tale metodologia sarà poi applicata ai microdati provenienti dall'indagine AVQ (paragrafo 6.1.2).

Ancora, i risultati della SLAD, sebbene non forniscano per costruzione una descrizione del fenomeno in esame, saranno poi utilizzati per costruire una partizione della popolazione italiana sulla base di un set più ampio di variabili (paragrafo 6.1.3), attraverso un albero di classificazione (paragrafo 4.3).

Infine, al fine di analizzare una forte associazione tra diverse variabili di uscita, è stata individuata la tecnica di analisi per classi latenti (paragrafo 4.4), che verrà poi utilizzata per realizzare una classificazione più puntuale della frequenza d'uso e del profilo di utilizzo di Internet, anche rispetto a variabili culturali, per gli utenti, e per valutare le diverse tipologie di non utenti (paragrafo 6.2.1).

4.1 Regressione logistica

Il primo strumento metodologico individuato, anche a seguito dell'analisi della letteratura, è dunque la regressione logistica (per una trattazione estesa, si veda per esempio (H. Stock e W. Watson, 2010)). Con questo termine si indica una particolare regressione che permette di individuare e isolare l'impatto di una serie di variabili di input, siano esse continue, categoriali o ordinali, su un'uscita dicotomica o a più categorie. Essa ipotizza che le probabilità delle diverse categorie di uscita si possa esprimere attraverso una funzione logit, del tipo:

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

Equazione 1

I modelli basati sulla funzione logit, a differenza di quelli basati sulla funzione probit, anch'essi utili alla valutazione dell'impatto dei predittori su un'uscita a due o più categorie, hanno il principale vantaggio di consentire l'espressione delle probabilità di uscita in forma chiusa, invertendo facilmente l'Equazione 1 come segue:

$$p = \frac{e^{\text{logit}}}{1 + e^{\text{logit}}}$$

Equazione 2

L'obiettivo di un'analisi che utilizzi tale metodo è identificare un modello matematico di relazione tra input e output, che calzi al meglio (*fit*, in inglese, termine spesso utilizzato anche in italiano) sui dati disponibili, che sia al contempo abbastanza parsimonioso in termini computazionali. Tra i più comuni esempi di modellizzazione di questo tipo vi è il modello di regressione lineare, in cui la variabile dipendente si assume essere continua, così come quelle di input.

Ciò che distingue la regressione logistica da quella lineare è principalmente l'assunzione sulla natura dell'uscita, che nel primo non è lineare, come nel secondo, ma per l'appunto dicotomica o categoriale. Questa differenza fra i modelli si riflette sia nella scelta del modello parametrico che nelle assunzioni. Una volta che si sia tenuta presente questa differenza, i metodi impiegati nell'analisi che utilizza una regressione logistica seguono in gran parte gli stessi principi generali della regressione lineare, con alcune differenze. Per ora si farà riferimento, per semplicità di esposizione, a un modello univariato, estendendolo poi al caso multivariato.

La prima differenza sostanziale fra le due metodologie è la natura della relazione tra uscita e variabili indipendenti. In ogni problema di regressione, uno dei valori fondamentali è rappresentato dal valor medio dell'uscita, dati i valori delle covariate. Questa quantità viene definita come media condizionale, ed è espressa come $E(Y|x)$ dove Y denota la variabile dipendente (o uscita, appunto) e x un valore della variabile indipendente (o covariata). Nella più nota regressione lineare si assume che questa media possa essere espressa con una equazione lineare in x , quale

$$E(Y | x) = \beta_0 + \beta_1 x$$

Equazione 3

Dove β_0 è una costante e β_1 il coefficiente della variabile x , che misura l'impatto della stessa sul livello dell'uscita.

Dunque, questa media può assumere un qualsiasi valore continuo, al variare di x tra $-\infty$ e $+\infty$, ma questo non sia adatta a un'uscita binaria o comunque categoriale.

Tra le altre possibili forme matematiche disponibili, atte a prevedere un'uscita di questo tipo, vi è appunto la logit, mostrata pocanzi.

Ci sono due buone ragioni per scegliere questo tipo di formula e la relativa distribuzione logistica. In primo luogo, da un punto di vista più squisitamente matematico, si tratta di una funzione estremamente flessibile, che si può invertire al fine di ottenere le probabilità in forma chiusa, e linearizzare attraverso il logaritmo, ottenendo una funzione lineare nelle variabili di input, esprimendo il logit come segue.

Nel seguito, al fine di rendere più semplice la notazione, la media condizionale di Y dato x , nel caso di distribuzione logistica, sarà indicata con $\pi(x) = E(Y | x)$. La forma specifica del modello di regressione logistica che viene utilizzato, è:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Equazione 4

Si può ottenere una trasformazione di $\pi(x)$, particolarmente importante all'interno della regressione logistica, che è definita, per l'appunto, come trasformazione logit:

$$g(x) = \ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] = \beta_0 + \beta_1 x$$

Equazione 5

L'importanza di tale trasformazione risiede nelle proprietà di $g(x)$, che di fatto sono le stesse di un modello di regressione lineare. Il logit $g(x)$ è infatti lineare nei suoi parametri, anche continui, e può variare da $-\infty$ a $+\infty$, a seconda delle variazioni di x .

Inoltre, la distribuzione logistica si presta ad interpretazioni cliniche od economiche molto significative, dato che rappresenta la distribuzione tipica della diffusione per contagio delle malattie, e della diffusione dei nuovi prodotti sul mercato (Bass, 1969).

Tornando alle differenze tra regressione lineare e logistica, una seconda riguarda la distribuzione condizionale della variabile di uscita.

Nei modelli di regressione lineare, possiamo assumere che una osservazione dell'uscita possa essere espressa come $y = E(Y|x) + \varepsilon$. La quantità ε esprime la deviazione dell'osservazione dalla media condizionale, ed è perciò definita "errore". L'assunzione più comune per ε è che essa segua una distribuzione normale con media zero e una varianza costante, indipendente

dai livelli della variabile indipendente. Ne consegue che la distribuzione condizionale della variabile di uscita dato x sarà normale con media $E(Y|x)$ e varianza costante.

Non può avvenire lo stesso nel caso di un'uscita dicotomica. Infatti, in questo contesto si può esprimere il valore dell'uscita, dato x , come $y = \pi(x) + \varepsilon$. Perciò, la quantità ε non è più libera, ma può assumere solo uno di due possibili valori, a seconda dei rispettivi valori di y . Se $y = 1$, allora $\varepsilon = 1 - \pi(x)$ con probabilità $\pi(x)$, e se $y = 0$ allora $\varepsilon = -\pi(x)$ con probabilità $1 - \pi(x)$. Ne consegue che ε ha una distribuzione con media nulla e varianza pari a $\pi(x) * [1 - \pi(x)]$ e dunque la distribuzione condizionale dell'uscita segue una distribuzione binomiale con probabilità data dalla media condizionale, $\pi(x)$.

Si è fin qui introdotta la regressione logistica in un contesto univariato. Si tralascia qui la trattazione diffusa della calibrazione, della verifica della significatività e della stima degli intervalli di confidenza nel caso di modelli univariati, che non sono usati nel corso del presente lavoro, data la loro relativa semplicità e limitatezza. Molto più utile risulta infatti l'abilità di modellare molte variabili di input, su scale di misura differenti (categoriali e continue).

Di seguito, il modello logistico fin qui esposto viene generalizzato al caso di più variabili indipendenti, definito caso multivariato.

Consideriamo un insieme di p variabili indipendenti contenute nel vettore $x' = (x_1, x_2, \dots, x_p)$, che vengono considerate rilevanti per il calcolo di un'uscita y . Poniamo per il momento che ognuna di queste variabili sia almeno definita su un intervallo continuo. Denotiamo la probabilità condizionata che la risposta sia presente come $P(Y=1 | x) = \pi(x)$. Estendendo banalmente quanto rappresentato nell'Equazione 5, il logit del modello di regressione logistica multipla è dato dall'equazione seguente:

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Equazione 6

In questo caso, il modello di regressione logistica è dunque

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}}$$

Equazione 7

All'interno dell'Equazione 6, non è possibile introdurre delle variabili categoriali così come sono, poiché essa è lineare. Le variabili categoriali, infatti, sono codificate tipicamente

attraverso numeri che ne rappresentano i vari livelli, ma che non hanno alcun significato numerico stretto. In questo caso, dunque, è necessario procedere alla sostituzione di ogni variabile categorica con una serie di variabili binarie (o *dummy*), ciascuna per una categoria della variabile originaria (a meno di una, se si vuole evitare la ridondanza).

In generale, se una variabile discreta ha k possibili valori, è necessario dunque creare almeno $k-1$ variabili binarie. L'eliminazione della k -esima variabile binaria è resa possibile dalla presenza di un termine costante, al cui interno vengono contemplati gli effetti di tale variabile. In seguito, per tali variabili binarie, sarà utilizzata la seguente notazione: supponiamo che la j -esima variabile indipendente x_j abbia k_j livelli, le k_j-1 variabili binarie saranno denotate come D_{jl} e i coefficienti relativi saranno denotati come $\beta_{jl}, l=1,2,\dots, k_j-1$. La formula del logit, per un modello con p variabili, di cui la j -esima categoriale, sarà

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \dots + \sum_{l=1}^{k_j-1} \beta_{jl} D_{jl} + \dots + \beta_p x_p$$

Equazione 8

Si supponga di avere un campione di n osservazioni indipendenti (x_i, y_i) , $i=1,2,\dots,n$. La calibrazione del modello richiede la stima del vettore dei coefficienti $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$.

Nella regressione lineare, il metodo più comunemente utilizzato per stimare i parametri ignoti è quello dei minimi quadrati. In questo metodo, vengono scelti quei valori di β' che minimizzano la somma dei quadrati delle differenze tra i valori osservati di Y e quelli stimati attraverso il modello. Sotto le usuali assunzioni della regressione lineare, il metodo dei minimi quadrati consente di ottenere stimatori che presentano alcune proprietà statistiche desiderabili. Tuttavia, quando il metodo dei minimi quadrati è applicato ad un modello con risposta dicotomica o categoriale, gli stimatori perdono tali proprietà.

Il metodo generale di stima, che nel caso in cui l'errore sia assunto distribuito normalmente produce la funzione dei minimi quadrati nella regressione lineare, è chiamato di massima verosimiglianza. Le equazioni di verosimiglianza da ottimizzare possono essere espresse come segue:

$$\sum_{i=1}^n [y_i - \pi(\mathbf{x}_i)] = 0$$

Equazione 9

e

$$\sum_{i=1}^n x_{ij} [y_i - \pi(\mathbf{x}_i)] = 0 \quad \text{con } j=1,2,\dots,p.$$

Equazione 10

La soluzione delle equazioni appena esposte richiede software specifici che sono implementati in quasi tutti i programmi statistici. All'interno del presente lavoro è stato principalmente utilizzato IBM SPSS.

4.1.1 Scelta delle variabili indipendenti

Finora si è discusso di quale forma funzionale scegliere per effettuare una regressione di un modello per la stima di un'uscita categoriale sulla base di una serie di variabili indipendenti che si suppone possano influenzare questi livelli, siano esse continue o categoriali. Resta aperta la questione di come stabilire se una variabile indipendente deve essere effettivamente inclusa in un dato modello.

I criteri per effettuare questa valutazione possono essere variabili da un problema al successivo e da una disciplina scientifica ad un'altra. L'approccio tradizionale alla costruzione di un modello statistico implica l'utilizzo del modello più parsimonioso possibile, in termini di variabili utilizzate, che al contempo spieghi comunque adeguatamente i dati, in termini di misure di *fitting*. La ratio è che più variabili sono incluse nel modello, più grandi diventano gli errori standard stimati, e più il modello diventa dipendente dal particolare campione di dati osservati. Spesso, un approccio osservato in letteratura, ad esempio all'interno degli studi epidemiologici, è quello di inserire tutte le variabili rilevanti da un punto di vista clinico e intuitivo, senza considerare la loro significatività statistica, al fine di tenere presente l'azione congiunta di variabili che prese singolarmente non mostrino un effetto, ma che prese tutte insieme possano effettivamente esercitare una forte influenza sull'uscita.

Il punto dell'includere molte variabili nel modello di regressione è che così si rischia una sovrastima (*overfit*) dei coefficienti, che risultano poi essere irrealisticamente grandi. Questo può essere particolarmente problematico quando il numero delle variabili incluse nel modello è ampio in rapporto all'effettiva dimensione del campione, oppure quando la proporzione tra le classi dell'uscita è molto sbilanciata, in termini di frequenza, verso una di esse.

Di seguito sono elencati i passi da seguire (Hosmer e Lemeshow, 2004) nella valutazione delle variabili:

- 1) il processo di selezione dovrebbe iniziare con un'attenta analisi univariata di ogni variabile. Per variabili con pochi valori interi, o categoriali, è possibile costruire una tabella di contingenza dell'uscita rispetto ai k livelli della variabile indipendente. È necessario inoltre prestare particolare attenzione alle celle nulle delle tabelle di contingenza, poiché ciò porta a una stima puntuale della probabilità pari a zero o infinito. Includere una tale variabile in un qualsiasi programma di regressione logistica porta alla comparsa di risultati numerici non desiderabili. Dunque, per eliminare le celle nulle si possono accorpate le categorie della variabile indipendente, eliminare completamente la categoria (se ritenuta non rilevante) o, se la variabile è ordinale, modellare la variabile come continua. Per le variabili continue, invece, l'analisi univariata può essere svolta attraverso la calibrazione di un modello di regressione logistica per poter valutare i coefficienti stimati e l'impatto della variabile.
 - 2) Dopo il completamento dell'analisi univariata, si procede alla selezione delle variabili per l'analisi multivariata. Ogni variabile il cui test univariato abbia valore di significatività $p < 0,25$ è una candidata per il modello multivariato, insieme a tutte quelle variabili riconosciute come importanti con approccio *top-down*, ad esempio perché utilizzate tipicamente in letteratura. Una volta che le variabili sono state identificate, si procede a testare il modello che contiene tutte le variabili selezionate. Il valore soglia di significatività viene posto pari 0,25 in vari lavori sperimentali in letteratura. Questo perché l'utilizzo di una soglia tipica (che può essere 0,05) può rivelarsi spesso troppo stringente, eliminando variabili che, invece, sono note a priori come rilevanti. D'altro canto, l'utilizzo di un valore più alto porta lo svantaggio di includere variabili di importanza discutibile, dunque è importante rivedere criticamente tutte le variabili aggiunte al modello prima di stabilire definitivamente il modello finale.
- Inoltre, l'approccio univariato ha il problema di ignorare la possibilità che un insieme di variabili, debolmente correlata con l'uscita quando presa singolarmente, possano diventare importanti predittori dell'uscita, quando presi congiuntamente. Se si crede che ciò sia possibile, si può conseguentemente scegliere un valore soglia di ammissione sufficientemente alto da permettere che più variabili diventino candidate all'inclusione nel modello multivariata.

In generale, la decisione di quali variabili includere in un modello multivariato includendo tutte le possibili variabili, dipende dalla dimensione totale del campione e dalla distribuzione di frequenze dell'uscita sulle diverse variabili esplicative.

Un altro approccio alla selezione delle variabili è di utilizzare un metodo *stepwise*, in cui le variabili vengono selezionate per inclusione od esclusione dal modello in modo sequenziale sulla base di soli criteri statistici. Ci sono due versioni principali della procedura *stepwise*: (a) selezione in avanti con test per eliminazione all'indietro o (b) eliminazione all'indietro seguita da un test per re-immissione in avanti. L'approccio a passi successivi è utile e intuitivamente attraente poiché costruisce il modello in maniera sequenziale e permette ad ogni passo l'esame del modello fino a quel momento costruito, cosa che non potrebbe essere fatta altrimenti.

Le grandi disponibilità e facilità, con cui i metodi *stepwise* possono essere utilizzati, rischiano di ridurre gli analisti al ruolo di assistenti del computer, piuttosto che di decisori ultimi sul modello più adeguato da utilizzare. Tuttavia, è solo tramite la comprensione da parte dell'analista delle potenzialità e soprattutto dei limiti di questi metodi, che questi ultimi possono essere sfruttati al meglio. È infatti l'analista, e non il computer, l'unico responsabile per la valutazione e la revisione del modello, poiché, al di là dei test statistici, può comprendere l'effettiva interpretabilità delle variabili nella spiegazione del fenomeno in esame.

- 3) L'importanza di ogni singola variabile inclusa nel modello deve essere verificata poi a seguito della calibrazione dello stesso. Ciò avviene attraverso un esame della statistica di Wald per ogni variabile e un paragone di ogni coefficiente stimato con il coefficiente del modello che contenga quella sola variabile. Le variabili che non contribuiscono significativamente al modello, sulla base di questi criteri, dovrebbero essere eliminate, calibrando poi nuovamente un nuovo modello. Il nuovo modello dovrebbe essere comparato con il precedente e più ampio, utilizzando il likelihood ratio test. Inoltre, i coefficienti stimati per le variabili rimaste dovrebbero essere comparati con quelli del modello completo. In particolare, bisogna verificare che i coefficienti di alcune variabili non abbiano subito notevoli variazioni di dimensione. Ciò infatti indicherebbe che una o più delle variabili escluse aveva in realtà importanza, apportando un correttivo all'effetto della variabile rimasta nel modello,

il cui coefficiente si è modificato (sia cioè un *confounder*, si veda il successivo par. 4.1.3.3).

Il processo di eliminazione, ri-calibrazione e verifica viene eseguito fino a che tutte le variabili rilevanti sono incluse nel modello, e quelle escluse sono irrilevanti da un punto di vista statistico o interpretativo.

- 4) Una volta ottenuto il modello che si crede contenga le variabili essenziali, dovremmo esaminare più attentamente le variabili incluse. La questione delle categorie appropriate per le variabili discrete dovrebbe essere affrontato a livello univariato. Per le variabili continue, bisogna verificare l'assunzione di linearità del logit.
- 5) Dopo aver raffinato il modello e accertato che ogni variabile continua abbia la scala corretta, il suggerimento è di ipotizzare possibili interazioni fra le variabili incluse. In qualsiasi modello, l'interazione fra due variabili implica che l'effetto di una variabile non è costante al variare dell'altra. La decisione riguardo all'adozione di un nuovo termine di interazione dovrebbe essere basata tanto su valutazioni statistiche quanto pratiche e interpretative del fenomeno in esame. Le variabili di interazione si costruiscono attraverso il prodotto delle coppie di variabili interagenti.

Si aggiunge una variabile di interazione per volta al modello e si verifica la sua significatività con il likelihood ratio test, così come esposto al passo 3.

4.1.2 Misure della bontà della calibrazione

Una volta che si è sostanzialmente certi di aver incluso, per quanto noto, quelle variabili (effetti principali e interazioni) che dovrebbero essere nel modello, e che tali variabili sono state inserite nella corretta forma funzionale, si procede a stabilire quanto il modello costruito descriva realmente la variabile di risposta. Questa misura è chiamata *goodness-of-fit*, o bontà della calibrazione.

Supponiamo di denotare i valori dell'uscita, osservati dal campione di dati, in forma vettoriale come $y' = (y_1, y_2, y_3, \dots, y_n)$. Indichiamo invece i valori stimati dal modello, o *valori calibrati*, con y^{\wedge} , dove $y^{\wedge} = (y^{\wedge}_1, y^{\wedge}_2, y^{\wedge}_3, \dots, y^{\wedge}_n)$. Il modello può essere considerato ben calibrato se: (1) le misure di sintesi della distanza tra y e y^{\wedge} assumono valori sufficientemente limitati e (2) il contributo di ogni coppia (y_i, y^{\wedge}_i) , $i=1,2,3,\dots,n$ a queste misure di sintesi è non sistematico e piccolo relativamente all'errore totale del modello.

Quindi, una valutazione completa del modello calibrato prevede sia il calcolo delle misure di sintesi della distanza fra y e y^{\wedge} che una profonda analisi dei componenti individuali di queste misure.

Le misure di sintesi della *goodness-of-fit* sono generalmente fornite di routine come output dei programmi di statistica, insieme ad ogni calibrazione del modello, e forniscono una indicazione generale della bontà del modello. In genere tali misure sono espresse attraverso il metodo statistico noto come test di Pearson del Chi-quadro (Agresti, 2007).

4.1.3 Interpretazione del modello calibrato

Dopo aver adeguatamente calibrato il modello, si può procedere all'interpretazione dei valori, che costituiscono il risultato principale dell'analisi di regressione, che consente di comprendere le relazioni tra l'output e i diversi input.

L'interpretazione di ogni modello calibrato implica una certa capacità di trarre delle indicazioni esplicative sul fenomeno oggetto di esame, a partire dai coefficienti stimati del modello. Soprattutto quando nella binarizzazione delle variabili categoriali si sia perseguita la strada della non ridondanza (cfr. par. 4.1.1), anche l'intercetta del modello costituisce oggetto d'interesse per l'indagine. I coefficienti delle variabili indipendenti possono essere anche letti come l'inclinazione di una funzione della variabile dipendente, per unità di cambiamento della variabile indipendente. Quindi, l'interpretazione riguarda due questioni: determinare la relazione funzionale, comunemente detta *link function*, fra la variabile dipendente e quella indipendente, e definire in maniera appropriata l'unità di cambiamento della variabile indipendente.

Il primo passo è determinare quale funzione della variabile dipendente porta a una funzione lineare delle covariate. Nel caso del modello di regressione lineare, essa è la funzione identità, poiché la variabile dipendente è lineare nelle variabili indipendenti, per definizione. Nel modello di regressione logistica, la *link function* è rappresentata dalla trasformazione logit, già mostrata in apertura di paragrafo 4.1.

Per la regressione lineare, il coefficiente di pendenza, β_1 , è pari alla differenza tra il valore della variabile dipendente a $x+1$ e quello a x , per ogni valore di x . In questo caso, l'interpretazione del coefficiente è relativamente diretta, poiché esprime il cambiamento risultante nella misurazione del livello della variabile dipendente per una variazione unitaria

della variabile indipendente. Per esempio, se in una regressione della quota di famiglie connesse a banda larga, rispetto al PIL (espresso in milioni) di una provincia, la pendenza è 1, significa che un incremento di 1 milione nel PIL della provincia è associato all'aumento di 1 punto percentuale della quota di famiglie connesse.

Nel modello di regressione logistica, invece, il coefficiente di pendenza rappresenta la variazione lineare nel logit corrispondente alla variazione di un'unità della variabile indipendente (cioè $\beta_1 = g(x+1) - g(x)$), oppure alla presenza (1) o assenza (0) di una data caratteristica, o di una categoria della stessa, dopo la binarizzazione. Un'interpretazione adeguata del coefficiente nella regressione logistica dipende dunque dalla capacità di dare un significato alla differenza tra questi due logit. L'interpretazione di questa differenza è discussa in dettaglio a seconda della tipologia di variabile indipendente nei seguenti paragrafi, prima per le variabili dicotomiche e categoriali, poi per quelle continue. Per semplicità di esposizione ci si riferisce qui al caso univariato, senza perdita di generalità.

4.1.3.1 Variabili dicotomiche e categoriali

Iniziamo col considerare la situazione in cui la variabile indipendente è misurata solamente su due livelli, ad esempio presenza o assenza di una data caratteristica. Questo caso fornisce anche la base concettuale per le altre possibili tipologie di variabili.

Assumiamo che la variabile indipendente, x , sia dunque codificata nella forma (0,1). La differenza nel logit per $x=1$ e $x=0$ è

$$g(1) - g(0) = [\beta_0 + \beta_1] - [\beta_0] = \beta_1$$

Equazione 11

L'algebra di quest'ultima espressione è banale, ma è riportata dettagliatamente al fine di sottolineare come il primo passo nell'interpretazione dell'effetto della covariata nel modello è esprimere la differenza del logit. In questo caso la differenza è pari a β_1 . Per poter interpretare questo risultato, è necessario introdurre e discutere una misura di associazione denominata *odds ratio*.

I possibili valori delle probabilità logistica sono rappresentati in Tabella 12.

Tabella 12 – Valori di probabilità dell'uscita, all'interno della regressione logistica, secondo le variazioni di una sola variabile indipendente dicotomica

| Variabile di uscita (Y) | Variabile indipendente | |
|-------------------------|--|--|
| | x = 1 | x = 0 |
| y = 1 | $\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$ | $\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$ |
| y = 0 | $1 - \pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$ | $1 - \pi(0) = \frac{1}{1+e^{\beta_0}}$ |

L'odd che l'uscita si manifesti effettivamente con livello 1 tra gli individui con $x=1$ è definita come $\pi(1) / [1 - \pi(1)]$. Similarmente, l'odd che l'uscita sia presente fra gli individui con $x=0$ è $\pi(0) / [1 - \pi(0)]$. L'odds ratio, indicato con OR, è definito dunque come il rapporto tra questi due odd, come nell'Equazione 12.

$$OR = \frac{\pi(1) / [1 - \pi(1)]}{\pi(0) / [1 - \pi(0)]}$$

Equazione 12

Sostituendo in quest'ultima equazione le espressioni di probabilità mostrate in Tabella 12, si ottiene la seguente formula.

$$OR = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right) / \left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) / \left(\frac{1}{1+e^{\beta_0}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1}$$

Equazione 13

Dunque, nel caso della regressione logistica con una variabile indipendente dicotomica codificata su (0,1), la relazione fra l'odds ratio e il coefficiente di regressione è

$$OR = e^{\beta_1}$$

Equazione 14

La semplicità di tale relazione è alla base del successo della regressione logistica come potente strumento di ricerca analitica.

L'odds ratio è una misura dell'associazione input/output che ha trovato ampio utilizzo, soprattutto in epidemiologia, poiché approssima ad esempio quanto sia più probabile che la

variabile di uscita sia presente tra i soggetti con presenza di una data caratteristica ($x=1$) rispetto a coloro che non la presentano ($x=0$), o viceversa.

Prima di concludere la discussione sul caso dicotomico, è importante considerare l'effetto che produce la codifica della variabile di uscita sul calcolo dell'odds ratio stimato. La stima di OR fornita nell'Equazione 14 è corretta quando la variabile indipendente è codificata sui due valori (0,1). Altre codifiche possono richiedere il calcolo del valore della differenza del logit per la specifica codifica utilizzata. In generale, infatti, la stima dell'odds ratio è ottenuta tramite l'esponenziale della differenza del logit:

$$OR = e^{\beta_1 * (a-b)}$$

Equazione 15

Si noti che questa espressione risulta essere pari a e^{β_1} solo quando $(a-b) = 1$.

Supponiamo ora che, invece di due sole categorie, la variabile indipendente si presenti con $k > 2$ valori distinti. Come visto in fase di calibrazione, verrà costruita, a partire dalla variabile originaria, una serie di variabili binarie per rappresentare tutte le categorie della variabile. Ad ogni variabile di questo tipo si potrà poi applicare quanto visto finora per le variabili dicotomiche e fornire un odds ratio che aiuti a spiegarne l'impatto sulla variabile di uscita.

4.1.3.2 Variabili continue

Nel caso di una variabile indipendente continua, ipotizziamo che il logit sia lineare in questa variabile. Ne consegue che il coefficiente β_1 rappresenta la variazione prodotta nel logit da una variazione per un incremento di 1 unità in x , cioè $\beta_1 = g(x+1) - g(x)$ per ogni valore di x . Molto spesso una variazione di una unità per la variabile indipendente non è significativa per il fenomeno. Ad esempio, se la variabile indipendente rappresenta l'età di una persona, una variazione di un solo anno potrebbe essere troppo piccola per incidere su un dato livello di uscita. D'altro canto, se invece la variabile continua x è definita solo su di un intervallo ristretto, per esempio da 0 a 1, chiaramente una variazione unitaria risulta essere molto elevata, e dunque bisognerebbe considerare frazioni dell'unità, quando si voglia valutare l'effetto di tale modifica l'effetto sull'uscita. In definitiva, quando si voglia fornire un'interpretazione utile sulle covariabili continue, è necessario sviluppare un metodo per stima puntuale e di intervallo per una variazione arbitraria di n unità della covariata.

Il logaritmo dell'odds ratio, per una variazione di c unità in x , è ottenuto dalla differenza del logit $g(x+c) - g(x) = n\beta_1$ e l'odds ratio associato viene calcolato facendo l'esponenziale di quest'ultima espressione.

$$OR(c) = OR(x+c, x) = n^*e^{\beta_1}$$

Equazione 16

Tale stima è dunque sensibile al particolare valore scelto per n . Dunque, questa scelta va fatta dunque in maniera ragionevole, di modo da fornire una indicazione chiara di quanto la variabile incida sull'uscita.

Inoltre, dev'essere anche valutato che un valore costante potrebbe essere inadeguato, poiché una variazione di una stessa quantità potrebbe avere effetti diversi a seconda del livello di partenza della variabile. Nell'esempio sull'età, si può ipotizzare che una variazione di 10 anni possa avere effetti diversi, a seconda che esso avvenga, per esempio, tra i 15 ed i 25 anni e tra i 50 ed i 60 anni. Si può allora pensare di suddividere la variabile in variabili dicotomiche. Ciò implica che l'assunzione che il logit sia lineare nella covariata iniziale non è più valida. Alternativamente, si potrebbe anche considerare l'utilizzo di termini di grado più alto o di altri tipi di forme non lineari, per tener conto dell'impatto della covariata sull'uscita.

4.1.3.3 Interazione e *confounding*

Il termine *confounder* è utilizzato in epidemiologia per descrivere una covariata associata sia all'uscita che ad un'altra variabile indipendente.

Il modello più semplice per includere l'interazione è quello in cui il logit è lineare su tutti i livelli del *confounder*, ma con una pendenza differente. Possono essere formulati modelli alternativi, in cui viene prevista una relazione non lineare fra il logit e le variabili nel modello, all'interno di ciascun gruppo. Un altro modo per incorporare tale interazione è inserire nel logit il prodotto delle due variabili associate, creando dunque una variabile derivata. Ovviamente, l'odds ratio risultante sarà relativo all'interazione delle due variabili, e non all'effetto singolo delle stesse.

4.2 Tecniche di classificazione LAD (Logical Analysis of Data) ⁷

La classificazione, intesa come l'attività di assegnazione a nuovi record in un dataset una delle categorie di una variabile di uscita, chiamata in genere etichetta, è tra le più diffuse nell'ambito del *data mining*. Diversi sono gli approcci proposti per risolvere questo problema, basati su diversi paradigmi e adattabili a diverse strutture di dati. Quelle più affermate includono: le reti neurali, le macchine vettoriali di supporto, il *k-nearest neighbor* (k-NN), gli approcci bayesiani, gli alberi decisionali (cfr. paragrafo 4.3), la regressione logistica (cfr. paragrafo 4.1), gli approcci booleani (Hastie *et al.*, 2005; Tsoumakas, Katakis e Vlahavas, 2010). In particolare, tra gli approcci booleani, uno particolarmente efficace è rappresentato dall'analisi logica dei dati (LAD) (Crama, Hammer e Ibaraki, 1988; Boros *et al.*, 1997, 2000, 2011), ispirato ai processi mentali che l'essere umano mette in atto nell'apprendere la classificazione da esempi del mondo reale.

In questo approccio, si suppone che i dati siano organizzati in record. Ogni record rappresenta un'osservazione diversa del fenomeno, ed è composto da campi contenenti i diversi valori osservati. Ogni campo ha il proprio dominio, cioè l'insieme dei possibili valori che esso può assumere. Un record può anche avere un'etichetta di classe. In questo caso, la classe viene chiamata anche output, mentre gli altri campi sono chiamati anche variabili esplicative. Per applicare l'approccio LAD, tutti i valori devono essere convertiti in forma binaria mediante un processo di discretizzazione chiamato binarizzazione. Il dominio di ogni campo viene partizionato in un numero finito di sottodomini codificati utilizzando attributi binari. Poiché il numero di attributi binari ottenuti è spesso molto grande, è necessario eseguire una fase di selezione degli stessi, al fine di individuare i più importanti. Tramite tale processo di selezione, vengono ottenuti gli attributi binari, utilizzati per creare dei pattern, che sono una congiunzione di attributi binari, chiamati anche condizioni, che caratterizzano una data classe. Ogni pattern riceve poi un peso, che misura l'importanza per la classificazione. Infine, ogni record non etichettato è classificato sulla base della somma ponderata dei pattern che coprono quel record. Nel successivo paragrafo 4.2.1 tale metodologia è descritta con maggiore dettaglio.

⁷ La descrizione dell'algoritmo SLAD e le relative procedure proposte in questo paragrafo provengono dal lavoro del gruppo composto da Renato Bruni (DIAG – Sapienza), Giampiero Bianchi (Istat), Claudio Leporelli (DIAG – Sapienza) e Cosimo Dolente, contenuto in (Bruni *et al.*, 2017).

La metodologia LAD è strettamente correlata agli alberi decisionali (Quinlan, 1992) e ai metodi di tipo k-NN (Cover e Hart, 1967), e costituisce un'estensione di questi due approcci, come mostrato in (Boros *et al.*, 2011), per citare solo alcune delle similitudini con altri algoritmi di classificazione.

Un ulteriore elemento chiave delle metodologie LAD risiede nel fatto che i pattern da esse generati possono fornire utili chiavi interpretative degli stessi fenomeni presi in esame (Crama e Hammer, 2011).

Dunque, in molti casi pratici, l'estrazione di regole, che identifichino l'appartenenza ad una data classe sulla base di un insieme di valori di alcuni variabili fissate, può risultare ancora più utile della classificazione stessa. Infatti, negli studi socioeconomici, quale questo, l'obiettivo principale è spesso la comprensione del comportamento delle persone e delle relative determinanti.

In particolare, quando si tratta del comportamento probabilistico di agenti economici, in generale è disponibile un gran numero di variabili esplicative, poiché moltissime possono essere le variabili che hanno un impatto su tali comportamenti complessi.

Come conseguenza di tale numero elevato di variabili, queste tecniche possono generare un numero estremamente elevato di pattern, la maggior parte dei quali potrebbe però avere scarso significato pratico, poiché individuano un numero troppo ristretto di casi. Inoltre, in questi casi, molte variabili esplicative sono spesso altamente associate; di conseguenza, sottoinsiemi di record ricoperti da pattern diversi risultano essere in gran parte sovrapposti.

Qui (par. 4.2.2) sono presentati alcuni criteri per individuare un insieme ridotto di regole che possano risultare significative, da un punto di vista pratico, all'interno del più ampio insieme dei pattern generati, insieme alle tecniche di ordinamento e filtraggio per la loro effettiva implementazione. Queste tecniche sono scalabili, a seconda dello specifico problema cui vengono applicate, e sono in grado, a differenza delle generali regole LAD, di produrre anche una serie di regole che siano internamente ortogonali, cioè le coperture di ogni coppia di regole hanno l'intersezione vuota. I modelli vengono generati utilizzando una versione della metodologia LAD sviluppata per gestire set di dati molto grandi (Bruni e Bianchi, 2015), progettata per ridurre il carico computazionale delle elaborazioni.

4.2.1 Binarizzazione delle variabili e generazione dei pattern

La struttura dei record, chiamata schema dei record R , consiste di un insieme di campi f_i , con $i = 1, \dots, m$. Un'istanza di record r , chiamata semplicemente record, consiste di un insieme di valori v_i , ciascuno dei quali riferito a un singolo campo. Come si diceva in apertura di paragrafo, tali campi, similmente a quanto accade per la regressione logistica sono essenzialmente di due tipi: quantitativo, o ordinato (quindi definito su di un intervallo di numeri interi) e qualitativo, anche denominato categoriale. Un record r è definito come etichettato, o classificato, se esso risulta assegnato a un elemento dell'insieme di possibili classi C di una variabile di uscita, o obiettivo. Viene definito come classificazione binaria il caso in cui C consiste solo di due elementi, normalmente denotati come $+$ (positivo) e $-$ (negativo). Per semplicità, qui di seguito verrà considerato questo caso binario. Come conseguenza, un record etichettato come positivo sarà denotato come r^+ , un negativo come r^- .

Viene costituito dunque per l'algoritmo un *training set* S di casi etichettati, identificando con S^+ il sottoinsieme di record positive e con S^- il sottoinsieme di quelli negativi. Tale insieme costituisce la fonte informativa dell'algoritmo di classificazione, da cui inferire le regole per la classificazione di casi la cui etichetta non sia nota a priori.

La metodologia LAD inizia con la binarizzazione, attraverso cui ogni campo non binario f viene convertito in un insieme di attributi binari a_i^j , con $j = 1 \dots n_i$. A seguito di tale operazione, si considera dunque un Nuovo schema di record R_b , che rappresenta un insieme di attributi binari a_i^j , e una nuova tipologia di record binarizzato r_b come insieme di valori binari $b_i^j \in \{0,1\}$ per tali attributi.

Per ogni campo qualitativo f_i , tutti i valori possono essere codificati in maniera semplice, attraverso un numero n_i di attributi binari a_i^j , in grado di discretizzare un campo con 2^{n_i} valori.

Per i campi quantitativi è invece necessario introdurre n_i soglie denominate *cut-point* e indicate ancora con a_i^j , e la binarizzazione di un dato valore v_i è ottenuto considerando se esso si trova al di sotto di ciascuna soglia:

$$b_i^j = \begin{cases} 1 & \text{se } v_i \geq a_i^j \\ 0 & \text{se } v_i < a_i^j \end{cases}$$

Equazione 17

I *cut-point* a_i^j dovrebbero essere posti su valori che rappresentano una qualche soglia significativa per il fenomeno oggetto di analisi. Essi sono calcolati come la semi-somma di ciascuna coppia di valori v_i' e v_i'' , appartenenti ai record adiacenti di classi opposte nel training set.

Questo identifica confini tra regioni corrispondenti a classi opposte. Quando i record positivi e negativi si sovrappongono e regioni puramente positive o negative diventano rare, questa tecnica può ancora fornire i confini tra regioni con predominanza di classe opposta (Bruni e Bianchi, 2015).

Un insieme di attributi binari $\{a_i^j\}$ utilizzati per binarizzare un insieme di dati S è definito come insieme di supporto U. Poiché il numero di attributi binari ottenibili nei problemi pratici è tipicamente molto elevato, e molti di essi potrebbero risultare irrilevanti per spiegare il fenomeno in esame, è importante poter selezionare un set di supporto ristretto (e significativo). Tale riduzione è necessaria anche per mitigare la complessità computazionale della rimanente parte di qualsiasi procedura di tipo LAD, che potrebbe altrimenti divenire impraticabile. Questo problema di ottimizzazione combinatoria è modellato utilizzando una variabile di decisione binaria x_i^j per ogni d_i , in modo tale che

$$x_i^j = \begin{cases} 1 & \text{se } a_i^j \text{ è tenuta nell'insieme di supporto} \\ 0 & \text{se } a_i^j \text{ è rimossa dall'insieme di supporto} \end{cases}$$

Equazione 18

Nella metodologia LAD classica, il problema viene risolto selezionando il set di supporto di dimensione minima che rimane esattamente separato, cioè in cui nessuna coppia di record positivi e negativi hanno la stessa codifica binaria. Ciò si traduce in un problema di copertura del set non ponderato (vedere, ad esempio, (Boros *et al.*, 1997)). D'altra parte (Bruni e Bianchi, 2015) propongono una tecnica per valutare il potere di separazione, definito come qualità e indicato con q_i^j , per ogni a_i^j . I q_i^j sono calcolati in modo che la qualità totale di un insieme di attributi binari corrisponda alla somma dei singoli valori di qualità. Pertanto, il problema può essere modellato come un *set covering* di pesi, con il vantaggio computazionale

rispetto al caso non ponderato, e il beneficio di mantenere nel set di supporto quegli attributi che presentano i migliori valori di qualità.

Da un punto di vista formale, con pattern si intende una congiunzione di condizioni che caratterizzano una data classe. Denotiamo un pattern positivo con P^+ e un negativo con P^- ; quando la classe non è specificata, si usa semplicemente P . Le condizioni sono rappresentate da attributi binari $a_i^j \in U$ o dai loro negati $\neg a_i^j$.

Abbiamo dunque che $P = 1$ se tutte le condizioni di P sono pari a 1, e $P = 0$ altrimenti. Diciamo che un modello P copre un record r , e viceversa che r attiva P , se l'insieme dei valori del record per le differenti condizioni porta a $P = 1$. Scriviamo $P(r)$ per indicare il valore del modello P applicato al record r :

$$P(r) = \begin{cases} 1 & \text{se } P \text{ copre } r \\ 0 & \text{se } P \text{ non copre } r \end{cases}$$

Equazione 19

Un pattern positivo P^+ (P^-) è definito come un modello che copra almeno c^+ (c^-) record positivi (negativi), ma non più di e^+ negativi (positivi). Definiamo la coppia di valori (c^+, e^+) come requisiti per un pattern positivo; allo stesso modo, (c^-, e^-) sono i requisiti per un pattern negativo. Tali valori sono stabiliti in modo che la copertura minima corretta (c^+ o c^-) è sempre maggiore della corrispondente copertura errata massima (e^+ o e^-). I pattern con $e^+ = 0$ o $e^- = 0$, vale a dire che non coprono alcun record della classe opposta, sono chiamati puri, mentre i pattern con $e^+ > 0$ o $e^- > 0$ sono chiamati fuzzy.

Infine, per eseguire la classificazione, i pesi w_b sono assegnati a tutti i pattern, con $w_b \geq 0$ per quelli positivi e $w_b \leq 0$ per i negativi. Per determinare tali pesi, sono state proposte diversi criteri (ad esempio, (Boros *et al.*, 2000; Bruni e Bianchi, 2015)). Un record r non classificato viene dunque etichettato sulla base dei pattern da esso attivati, calcolando la seguente somma ponderata, chiamata discriminante: $\Delta(r) = \sum_h w_h P_h(r)$ e selezionando una soglia δ : r . Il record r sarà classificato come positivo se $\Delta(r) > \delta$ e negativo altrimenti.

I pattern possono essere generati utilizzando tecniche di enumerazione combinatoria, basate su due possibili tipi di procedure: *bottom-up* o *top-down*. La generazione *bottom-up* di un pattern consiste nell'aggiungere una condizione per volta fino a ottenere una formula che rispetti i requisiti definiti, in termini di copertura ed errore. La generazione *top-down*, al contrario,

procede rimuovendo le condizioni una per una dall'insieme delle condizioni che coprono un solo record, finché i requisiti non vengono violati. Inoltre, sono stati studiati diversi tipi di pattern (ad esempio, *strong*, *prime*, *spanned*, *maximum*, vedere anche il paragrafo successivo), e sono stati sviluppati diversi algoritmi di generazione specifici per la loro enumerazione (Hammer *et al.*, 2004; Alexe *et al.*, 2007; Bonates, Hammer e Kogan, 2008; Ryoo e Jang, 2009; Guo e Ryoo, 2012). L'approccio scelto all'interno di questo lavoro è di generazione *bottom-up* dei pattern, utilizzando le condizioni con ordinamento *greedy*, cioè riducendo i valori q^i ed evitando le specializzazioni di congiunzioni che costituiscono già dei pattern.

4.2.2 Identificazione dei pattern più interessanti per la caratterizzazione di comportamenti socio-economici probabilistici

Tutti i pattern generati dalla procedura forniscono il loro contributo per la classificazione e, in generale, un ampio insieme di pattern garantisce una migliore precisione. Tuttavia, se l'obiettivo dell'analisi è rappresentato dalla ricerca di spiegazioni di fenomeni socio-economici, l'interpretazione di un grande insieme di pattern può essere problematica. In questo caso, l'insieme di tutti i record disponibili etichettati viene utilizzato come *training set* S e non è necessario un passo di classificazione dopo la generazione dei pattern. Invece, l'obiettivo è mettere a punto una procedura algoritmica che dovrebbe identificare i pattern più significativi, da un punto di vista interpretativo, all'interno del grande insieme P di tutti i pattern.

Dato un generico pattern P_b , i seguenti insiemi e valori sono definiti al fine di descriverne le caratteristiche. $Lit(P_b)$ rappresenta l'insieme delle condizioni che compongono il pattern, mentre $l(P_b) = |Lit(P_b)|$ è la cardinalità dell'insieme, pari dunque al numero di condizioni.

Nello spazio definito dagli d attributi binari di U , l'ipercubo booleano B^d è l'insieme dei punti 2^d che hanno come coordinate tutte le possibili stringhe binarie di lunghezza d , cioè tutti i possibili record binarizzati. Un sottocubo di dimensione $(d - l)$ è costituito dai punti 2^{d-l} di B^d per cui le coordinate $l < d$ sono fissate a 0 o 1. Un modello positivo P_h^+ è un particolare sottocubo di dimensione $(d - l(P_h^+))$ tale che le cardinalità delle sue intersezioni con $S^+ \subset B^d$ e $S^- \subset B^d$ soddisfano i requisiti (c^+, e^+) per i pattern positivi.

Per i pattern negativi si ha una situazione del tutto speculare. La dimensione $s(P_b^+)$ rappresenta il numero di punti $2^{d-l(P_h^+)}$ del corrispondente sottocubo.

La copertura corretta $Cov(P_b^+)$ è l'insieme di osservazioni di S^+ coperte da P_b^+ , la cui cardinalità è indicata con $c(P_b^+) = |Cov(P_b^+)|$. Allo stesso modo, la sua copertura errata $Err(P_b^+)$ rappresenta l'insieme di osservazioni di S^- coperte da P_b^+ , con cardinalità $e(P_b^+) = |Err(P_b^+)|$. Inoltre, si tenga presente che nel caso di dati provenienti da un campione estratto da una popolazione più ampia, come quello in esame, ogni record r non rappresenta solo un individuo, ma un numero μ_r di individui, di cui è rappresentativo nell'universo di indagine. Riportando dunque all'universo ogni record, i valori appena esposti vengono modificati come segue: il numero di casi corretti è $c^u(P_h^+) = \sum_{r \in Cov(P_h^+)} \mu_r$, il numero di errori è $e^u(P_h^+) = \sum_{r \in Err(P_h^+)} \mu_r$. Definizioni simili possono essere formulate per pattern negativi.

Le caratteristiche dei pattern sono state oggetto di diversi studi all'interno della ricerca dedicata alla LAD (Hammer *et al.*, 2004; Alexe *et al.*, 2007; Bonates, Hammer e Kogan, 2008). Un pattern P_b è definito come forte se non esiste alcun altro pattern P_k tale che $Cov(P_b) \subset Cov(P_k)$. Un pattern P_b è definito primo se l'eliminazione di una qualsiasi condizione produce una formula che non rispetta più i requisiti per i pattern.

Alcuni criteri di preferenza sono descritti in (Hammer *et al.*, 2004):

- *semplicità*: un pattern P_b è preferito ad un pattern P_k se e solo se $Lit(P_b) \subseteq Lit(P_k)$, dunque il secondo rappresenta una specificazione del primo;
- *selettività*: un pattern P_b è preferito ad un pattern P_k se e solo se $s(P_b) \subseteq s(P_k)$, cioè ha una dimensione contenuta nel secondo;
- *copertura*: un pattern P_b è preferito ad un pattern P_k se e solo se $Cov(P_k) \subseteq Cov(P_b)$, cioè la copertura del secondo è interamente contenuta in quella del primo.

Tuttavia, questi criteri potrebbero non essere idonei a selezionare pattern che aiutino a interpretare i fenomeni socioeconomici. In questi casi, supponiamo che ogni record etichettato del campione S riporti, per un dato individuo, l'esito osservato di una scelta discreta e l'insieme di fattori che hanno influenzato tale scelta. Il primo è chiamato l'output, mentre i secondi sono chiamati variabili esplicative. I metodi di scelta discreta (Manski e McFadden, 1981; Train, 2009), quale la regressione logistica presentata in 4.1, costruiscono modelli del processo decisionale che determinano la stima della probabilità di ciascuna scelta (l'output) dato i valori di m variabili esplicative. La natura probabilistica di questi modelli riflette l'eterogeneità dei decisori e i limiti delle variabili esplicative nel caratterizzare la scelta.

Teoricamente, potendo disporre di un campione sufficientemente ampio, si potrebbero stimare le probabilità $Pr(+|r)$ e $Pr(-|r)$ dell'output calcolando le frequenze delle sue classi per ogni combinazione r di tutte le variabili esplicative. Nella pratica, la dimensione del campione disponibile normalmente non consente tale granularità, e il numero di variabili esplicative deve essere ridotto al fine di evitare problematiche di *overfitting* sui dati e multicollinearità.

Per definire un sottoinsieme (cioè una categoria di individui), si fissa normalmente una tupla di valori per $k < m$ delle m variabili esplicative, denotandole con (v_{i1}, \dots, v_{ik}) . In generale, ogni variabile esplicativa ha una diversa importanza nella determinazione del comportamento. Se le k variabili scelte risultano essere abbastanza importanti, le frequenze delle diverse classi di output all'interno del sottogruppo definito da (v_{i1}, \dots, v_{ik}) approssimerà adeguatamente le rispettive probabilità. Poiché ogni pattern corrisponde alla binarizzazione di una tupla (v_{i1}, \dots, v_{ik}) , esso definisce una categoria specifica di individui. Dato, senza perdita di generalità, un pattern positivo P_b^+ , gli individui in $Cov(P_b^+)$ sono quelli che, in quella categoria di individui, si comportano in modo da avere output positivo, mentre quelli in $Err(P_b^+)$ sono quelli che, nella stessa categoria di individui, si comportano in modo opposto.

I pattern significativi, da un punto di vista pratico, dovrebbero dunque avere le seguenti proprietà:

- *generalità*, cioè presentare un'ampia copertura corretta (grande valore di c);
- *precisione*, cioè presentare una piccola copertura errata (piccolo valore di e);
- *semplicità*, contenere un numero ristretto di condizioni (piccolo valore di l).

Dunque, nello scegliere i pattern, è necessario di più criteri, che sono spesso, nella pratica, contrastanti tra loro. Tendenzialmente, infatti, nelle analisi del presente progetto di ricerca, si è evidenziata una certa correlazione positiva tra generalità e semplicità, laddove pattern semplici, cioè con poche condizioni (al limite solamente una), individuano normalmente un ampio numero di casi. Al contempo, si è sperimentata una correlazione di segno opposto tra questi due obiettivi e la precisione, spiegabile dal fatto che l'alto numero di casi coperti da pattern semplici portano con sé, necessariamente data la natura complessa del fenomeno, anche un certo numero di errori, cioè di casi identificati erroneamente in una delle due classi.

Per tenere conto di questo *trade-off* emerso tra copertura ed errore, è stato dunque introdotta una misura della precisione, denominata errore percentuale e definita come segue:

$$\varepsilon(P_h^+) = \frac{100 e(P_h^+)}{c(P_h^+) + e(P_h^+)} \%$$

Equazione 20

Si noti che tale valore è in realtà la probabilità che un individuo appartenente a una data categoria della popolazione, definita da P_b^+ , presenti un output negativo. Ora, è possibile combinare $c(P_b)$ e $\varepsilon(P_b)$ in diversi modi. Nel seguito, saranno presentate le procedure originali del lavoro (Bruni *et al.*, 2017), utilizzando il valore $c(P_b)/\varepsilon(P_b)$. Tuttavia, possono essere prese in considerazione altre combinazioni di copertura e accuratezza, a seconda di quale aspetto si voglia privilegiare, utilizzando anche i valori $c''(P)$ and $e''(P)$ relative all'universo di indagine.

Un primo criterio di preferenza, di probabilità empirica, è definito come segue: un pattern P_b è preferito a un pattern P_k se e solo se $c(P_b)/\varepsilon(P_b) > c(P_k)/\varepsilon(P_k)$. Ciò significa che P_b definisce una categoria abbastanza ampia di individui che si comportano in modo sufficientemente uniforme, rispetto al fenomeno in esame e pertanto risulta essere significativo. Per applicare questo criterio, possiamo ordinare tutti i pattern, separatamente per ogni classe, per valori decrescenti di $c(P)/\varepsilon(P)$.

Tuttavia, per ottenere pattern significativi, è necessario considerare anche un altro aspetto. In fenomeni complessi, vi sono variabili fortemente associate (nel nostro esempio, livello di istruzione e posizione professionale). Quindi, lo stesso aspetto del fenomeno può essere spiegato con pattern alternativi, ciascuno dei quali utilizza una o più delle di queste variabili fortemente associate, in quanto tali modelli copriranno prevedibilmente insiemi altamente sovrapposti di record. In generale, è desiderabile evitare questo tipo di ridondanza, anche se la presenza di tali spiegazioni alternative potrebbe risultare interessante, da un punto di vista esplicativo, quando esse derivano dall'effetto congiunto di diversi gruppi di variabili in due pattern diversi, anziché dall'associazione diretta di due singoli variabili. Infatti, si evidenzerebbero così correlazione tra diversi aspetti del fenomeno non noti a priori.

4.2.2.1 Ordinamento per copertura incrementale

Un importante punto debole del metodo appena descritto risiede nel fatto che le coperture di pattern differenti potrebbero avere importanti sovrapposizioni. Un pattern P_j potrebbe presentare un alto valore di $c(P_j)/\varepsilon(P_j)$, e risultare dunque desiderabile, ma coprire in gran parte gli stessi casi di un altro pattern P_k che precede P_j nell'ordinamento secondo tali valori.

Per superare tale problema, definiamo un nuovo valore $Cov^I(P_b)$, detto copertura incrementale. Dato un ordinamento O dei pattern di una classe, ad esempio positive, ed indicando con $P_k \prec P_b$ che P_k precede P_b in O , la copertura incrementale $Cov^I(P_b)$ è rappresentata dall'insieme di record di S^+ coperti da P_b che non erano in $Cov(P_k)$, per tutti i k tali che $P_k \prec P_b$. Ne consegue che, per il primo pattern nell'ordinamento, la copertura incrementale coincide con quella normale. Il numero incrementale di casi corretti è indicato con $c^I(P_b) = |Cov^I(P_b)|$. Dato O , definiamo in maniera simmetrica l'errore incrementale $Err^I(P_b)$, la sua cardinalità $e^I(P_b)$ e la percentuale di errore incrementale $\varepsilon^I(P_b) = 100 * e^I(P_b) / (c^I(P_b) + e^I(P_b))\%$.

Un modo per calcolare questi valori è mantenere una matrice M delle incidenze tra pattern e record: ogni elemento m_{bk} è pari a: 1 se $r_k \in Cov(P_b)$; -1 se $r_k \in Err(P_b)$; 0 altrimenti.

Ora possiamo definire un secondo criterio di preferenza, chiamato probabilità empirica disgiunta: un pattern P_b è preferito a un pattern P_k se e solo se $c^I(P_b) / \varepsilon^I(P_b) > c^I(P_k) / \varepsilon^I(P_k)$. Ciò equivale a dire che P_b definisce una categoria abbastanza ampia di individui che si comportano in modo sufficientemente uniforme e che sono abbastanza disgiunti (anche se non completamente) dalle categorie definite dai pattern che precedono P_b all'interno dell'ordinamento O . Per identificare un pattern significativo secondo questo criterio, utilizziamo la seguente *Procedura 1*, separatamente per ciascuna classe.

Procedura 1 – Identificazione di pattern significativi sufficientemente disgiunti

| | |
|----------------------------|---|
| Input | Insieme dei pattern P di una data classe; matrice M di incidenza pattern-record |
| Output | Ordinamento O^I di P sulla base della probabilità empirica disgiunta |
| 1. Inizializzazione | Ordinare i pattern per valori decrescenti di $c(P) / \varepsilon(P)$, ottenendo l'ordinamento iniziale O_0 . Porre $\pi := 1$. |
| 2. Iterazione t | <ul style="list-style-type: none"> a) Per i pattern nelle posizioni da $\pi + 1$ all'ultima nell'attuale ordinamento O_t, calcolare c^I e ε^I corrispondenti ad O_t utilizzando la matrice M; b) Dalla posizione $\pi + 1$ fino all'ultima in O_t, ordinare i pattern per valori decrescenti di c^I / ε^I, ottenendo un nuovo ordinamento O_{t+1}; c) Confrontare O_{t+1} con O_t e porre π pari all'ultima posizione in cui O_t e O_{t+1} coincidono; d) Verificare se il pattern in posizione $\pi + 1$ ha precedentemente assunto, e poi lasciato, tale posizione. Se SI, fissarlo nella posizione $\pi + 1$ e porre $\pi := \pi + 1$; e) Se $O_{t+1} \neq O_t$, allora porre $t := t + 1$ e ripetere l'Iterazione. Altrimenti, porre $O^I := O_{t+1}$ e uscire. |

Si può dimostrare che tale procedura termina, perché l'ordinamento converge sempre (Bruni *et al.*, 2017). Essa è anche fattibile da un punto di vista computazionale, dato che essa essenzialmente ricalcola le coperture incrementali e ordina i valori.

I pattern nelle prime posizioni di O^l rappresentano dei buoni compromessi tra la copertura, l'accuratezza e la disgiunzione delle coperture. Il numero di pattern da estrarre può essere scelto, ad esempio, selezionando i primi n che coprono almeno una certa porzione dell'insieme di dati, o fino a quando il loro valore per ϵ^l/ϵ^j è al di sopra di una determinata soglia (si noti che un pattern con $\epsilon^l < \epsilon^j$ aggiunge più errori che casi corretti, e dunque risulta inutile).

Come si vedrà nel seguito (par. 6.1.2), applicando tale procedura al caso in esame si passa da un insieme di diverse migliaia di pattern a poche centinaia, in grado comunque di coprire quasi tutti record del campione.

4.2.2.2 Generazione di pattern ortogonali

La Procedura 1 descritta nel paragrafo precedente mira ad ottenere pattern che corrispondono a categorie disgiunte di individui. Tuttavia, essa non può fornire un controllo sul livello effettivo di disgiunzione tra le coperture. Pertanto, presentiamo qui un'altra procedura di ordinamento, in grado di garantire la disgiunzione totale delle categorie di individui individuate dai pattern generate dall' algoritmo SLAD.

Procedura 2 – Identificazione di pattern significativi ortogonali

| | |
|----------------------------|---|
| Input | Insieme dei pattern P di una data classe; matrice M di incidenza pattern-record |
| Output | Un sottoinsieme di pattern P^o di P , internamente ortogonali |
| 1. Inizializzazione | Ordinare i pattern per valori decrescenti di $\epsilon(P)/\epsilon(P)$, selezionando il primo pattern P_1^o . |
| 2. Iterazione t | <ol style="list-style-type: none"> Rimuovere da M tutte le colonne (corrispondenti ai record) contenute in $Cov(P_1^o)$. Se le rimanenti colonne sono minori di una soglia ν, uscire; Per ogni riga di M, aggiornare il numero di casi identificati correttamente e di errori, determinando un nuovo insieme di pattern P_t; In P_t, selezionare il pattern che massimizza $\epsilon(P)/\epsilon(P)$ e denominarlo P_{t+1}; Calcolare la negazione logica del precedente pattern ortogonale P_t^o; Generare il successive pattern ortogonale $P_{t+1}^o = P_{t+1} \wedge \neg P_t^o$; porre $t := t + 1$ e ripetere l'Iterazione. |

La procedura appena descritta genera pattern finché la copertura totale dell'insieme dei pattern non è pari ad una certa quota del campione. Intuitivamente, la lunghezza dei pattern ortogonali cresce rapidamente. Tuttavia, nell'applicazione si è sperimentato che è necessario un numero davvero ristretto di pattern per coprire quasi tutto il campione (cfr. par. 6.1.2). Inoltre, anche la sequenza di pattern P_t (generate senza la negazione dei precedenti) riveste un'importanza significativa, dal punto di vista pratico, con il vantaggio di essere ancora più facile da interpretare. Infine, le condizioni definite da ciascuna P_t , per $t = 1, \dots, \tau$, quando τ è pari a poche unità, possono essere utilizzate per produrre una divisione dell'insieme di dati considerando tutte le combinazioni 2^τ delle loro affermazioni/negazioni (come effettivamente svolto nel paragrafo 6.1.2). Questo costituisce una partizione degli individui in categorie rilevanti per il fenomeno in esame. La procedura è valida da un punto di vista computazionale, in quanto essenzialmente essa aggiorna il numero di casi coperti correttamente e di errori, trova il massimo di un vettore e registra le negazioni logiche.

Nelle procedure 1 e 2 descritte possono essere applicate alcune varianti dei criteri di ordinamento.

Dato un record r , ad esempio positivo, definiamo il suo pattern di riferimento P_0 come quello con la minima percentuale di errore $\varepsilon(P_0)$, tra quelli che lo coprono. La copertura di riferimento $Ref(P_0)$ è l'insieme di record di S^+ per cui P_0 risulta essere il pattern di riferimento. Indichiamo anche $\varrho(P_0) = |Ref(P_0)|$. Utilizzando questo concetto, possiamo ordinare sulla base di $\varrho(P)$ anziché $c(P)/\varepsilon(P)$, perseguendo comunque un compromesso tra l'accuratezza e la copertura, che corrisponde approssimativamente alla ricerca di pattern con copertura ampia, ma scegliendo solo tra quelli più accurati.

Infine, la generazione di pattern può essere effettuata non solo fissando due soglie di e^+ ed e^- sul numero di errori, ma anche fissando soglie relative alla percentuale di errore, che indichiamo con ε^+ e ε^- . Un'altra variante del metodo può essere ottenuta anche fissando in iterazioni successive più soglie differenti sulla percentuale di errore, ad esempio generando prima pattern con percentuale di errore $\leq 20\%$, poi con percentuale di errore $\leq 10\%$, ecc. Chiaramente, in quest'ultimo caso, i pattern non devono essere replicati con diverse soglie di errore. Questa ultima tecnica ha lo scopo di fornire pattern situati in diverse aree del *trade-off* tra la copertura e l'accuratezza, alle quali è possibile applicare le procedure di selezione descritte finora.

4.3 Alberi di classificazione

Gli alberi di decisione sono una famiglia di algoritmi di classificazione non parametrici in grado di suddividere un insieme di dati sulla base dell'omogeneità rispetto a una data variabile di uscita, ottenendo una partizione completa nei dati (Cerioli e Grossi, 2009). Nel caso la variabile di uscita sia continua, si parla di alberi di regressione, mentre nel caso essa sia categoriale od ordinale si parla di alberi di classificazione. Essi rivestono un ruolo importante in diversi campi dalla teoria delle decisioni, al *data mining* e al *machine learning*. Di seguito sono descritti brevemente alcuni di questi algoritmi di classificazione, che è possibile implementare attraverso il software principale utilizzato nel corso della ricerca, IBM SPSS (versioni dalla 21 alla 24):

- CHAID: CHi-squared Automatic Interaction Detector;
- CART: Classification and Regression Trees;
- QUEST: Quick, Unbiased and Efficient Statistical Tree.

Essi vengono presentati di seguito nell'ordine appena esposto, che è quello di comparsa in letteratura, rispetto al solo caso delle uscite categoriali, che risponde alle necessità del progetto di ricerca. Come si vedrà, l'algoritmo QUEST risulta essere il più efficiente e consente di valutare un'uscita categorica con variabili di input categoriche, motivo per il quale esso è stato scelto per effettuare le analisi del paragrafo 6.1.3.

4.3.1 CHAID

Iniziamo questa disamina con la prima tra le tecniche a disposizione per gli alberi di classificazione, il CHAID (CHi-squared Automatic Interaction Detector) (Kass, 1980).

Essa costituisce un'estensione di una delle prime metodologie per gli alberi di classificazione introdotte in letteratura, denominata AID (Automatic Interaction Detection) (Morgan e Sonquist, 1963). Questa annoverava tra le principali limitazioni la possibilità di considerare esclusivamente output continui e input categoriali. Il criterio di suddivisione di ciascun nodo era di tipo binario, basato sulla scomposizione della varianza dell'uscita all'interno e all'esterno dei gruppi appartenenti ai sottonodi.

L'algoritmo CHAID propone invece un metodo di segmentazione dei nodi basato sul test chi-quadro e consente di considerare anche variabili di uscita categoriali. Esso utilizza un

criterio di suddivisione dei nodi simile a quello della regressione *stepwise*. Infatti, al fine di trovare la migliore variabile per la suddivisione di un dato nodo, questo viene dapprima suddiviso in due o più nodi figlio, a seconda del tipo di variabile. Il CHAID può considerare tre diverse tipologie di variabili: categorica, ordinata senza mancanti (denominata *monotonica*), e ordinata con mancanti (denominata *floating*). In questo caso, per i valori mancanti viene utilizzata una nuova categoria separata nella variabile categorica.

Dunque, se la variabile di input considerata è categorica, il nodo viene suddiviso in tanti nodi quante sono le categorie della variabile. Se essa è invece *monotonica*, l'algoritmo suddivide il nodo padre in 10 nodi figli, ognuno dei quali definito da un intervallo di valori della variabile di input. Infine, se la variabile di input è di tipo *floating*, il nodo n è suddiviso in 10 nodi figlio, più uno per i valori mancanti. Viene poi considerata la possibilità di fondere ciascuna coppia di nodi figlio adiacenti, attraverso un test di Bonferroni⁸, opportunamente modificato, che verifica la significatività delle differenze nelle medie. I nodi così ottenuti vengono dunque considerati per la suddivisione, di nuovo attraverso un test di Bonferroni.

La procedura iterativa termina necessariamente quando tutti i nodi terminali dell'albero contengono solo record perfettamente omogenei nella variabile dipendente. Alternativamente, si può stabilire un criterio di arresto, imponendo ad esempio che l'algoritmo termini quando tutti i nodi foglia contengono un numero di casi inferiore ad una data soglia.

4.3.2 CART

La metodologia CART (*Classification and Regression Trees*) è una delle più diffuse all'interno degli alberi di classificazione (Breiman *et al.*, 1984). Rispetto agli algoritmi proposti prima della sua comparsa, su l'AID (Morgan e Sonquist, 1963) e lo stesso CHAID, essa introduce una serie di elementi innovativi, quali:

- la possibilità di prevedere una variabile qualitativa, oltre quelle quantitative continue (così come nel CHAID). Nel primo caso, come si diceva, si parla per l'appunto di alberi di classificazione, nella seconda di alberi di regressione;

⁸ Per un approfondimento, si veda (Bonferroni, 1936).

- la possibilità di utilizzare all'interno della stessa classificazione variabili indipendenti sia qualitative che quantitative;
- la possibilità di effettuare gli *split* ad ogni nodo sulla base di combinazioni lineari delle variabili continue di input;
- la definizione del criterio di suddivisione come minimizzazione dell'impurità del nodo, intesa come disomogeneità dei nodi risultanti in termini di risposta;
- la generazione del numero più elevato possibile di foglie, in luogo dell'utilizzo di un criterio di arresto, con successiva ottimizzazione attraverso tecniche di potatura (*pruning*);

A fronte di queste innovazioni, la principale limitazione della tecnica CART risiede nella possibilità di effettuare suddivisioni successive esclusivamente binarie. All'interno

4.3.2.1 Criterio di suddivisione sulla base dell'impurità

All'interno di questa breve disamina si considereranno solo gli alberi di classificazione, dato che le variabili considerate sono in gran parte categoriali.

In questi casi, si supponga che la variabile di uscita sia categorica, con m modalità. Il concetto di impurità riguarda la selezione la variabile per ogni suddivisione in modo tale che ciascuno dei sottogruppi prodotti dalla ripartizione sia per l'appunto più "puro" rispetto a quello di partenza, rispetto alle modalità della variabile di risposta. Dunque, si parla di impurità nulla se gli r record di un dato gruppo sono completamente omogenei nelle variabili di uscita, presentando cioè tutti lo stesso valore per essa. Invece, l'impurità è massima se tali record hanno valori per la variabile di uscita uniformemente suddivisi sulle m modalità. Dunque, a partire dalla radice (o nodo padre) si seleziona la variabile che produce la migliore bipartizione del campione iniziale in termine di impurità prodotta all'interno dei primi due nodi figlio.

Tale concetto può essere formalizzato come segue. Indichiamo con $f_{j|n}$ la frequenza relativa della modalità j dell'uscita all'interno di un dato sottogruppo R_n che rappresenta il nodo n . Si noti che la somma delle f così definite su tutte le m modalità della variabile di risposta è per definizione unitaria.

La misura di impurità associata al nodo n può essere definita allora come funzione di tali frequenze relative:

$$imp(n) = \phi(f_{1|n}, f_{2|n}, \dots, f_{m|n})$$

Equazione 21

dove la $\phi(\cdot)$ indica una funzione non negativa tale che:

- Quando esiste una j per cui $f_{j|n}=1$ (e dunque tutte le altre sono nulle, come si diceva), allora $Imp(n)=0$;
- $Imp(n)$ risulta essere massima quando $f_{j|n}=1/m$ per ogni $j=1,2,\dots,m$ (per rispettare la condizione di massima impurità);
- La funzione è invariante rispetto all'ordinamento specifico delle m modalità dell'uscita.

Una delle forme funzionali più utilizzate in letteratura è la seguente:

$$imp(n) = 1 - \sum_{j=1}^m f_{j|n}^2 = \sum_{j \neq i}^m (f_{j|n}) * f_{i|n}$$

Equazione 22

Questa forma funzionale risulta infatti essere relativamente semplice, da un punto di vista computazionale, e offre una doppia possibilità di interpretazione. La prima uguaglianza può rappresentare la varianza di Y nel nodo n nel caso in cui si indichi con 1 la presenza della modalità j tra i record del nodo e con 0 la presenza delle altre modalità, mentre la seconda rappresenta la stima di probabilità di una errata classificazione di un'osservazione della modalità j nella modalità i , in caso di assegnazione casuale di un record al nodo n .

Si definisce poi misura del decremento del nodo n associata a un determinato *split* s la quantità:

$$\Delta imp(s, n) = imp(n) - f_1 imp(n_1) - f_2 imp(n_2)$$

Equazione 23

Dove f_1 e f_2 sono le porzioni di record che cadono nei due nodi figlio. Per costruzione, l'Equazione 23 è sempre non negativa, ed è sempre positiva a meno del caso speciale in cui le frequenze condizionate sono identiche nei nodi figli, e lo sono anche nel nodo padre di riflesso, e quindi non vi è alcun guadagno informativo nella suddivisione. A partire dalla radice, l'algoritmo procede dunque a tutte le S suddivisioni in due nodi possibili rispetto a tutte le variabili di input, e sceglie ad ogni passo lo *split* ottimo s^* per cui vale:

$$\Delta imp(s^*, n) = \max_{s \in S} \Delta imp(s, n)$$

Equazione 24

L'impurità totale dell'albero viene poi calcolata come media delle impurità dei nodi foglia dell'albero finale, pesata sulla quota del campione iniziale all'interno di ciascun nodo.

4.3.2.2 Procedura di potatura (*pruning*)

Per scegliere l'albero ottimale, è necessaria una regola di arresto nella procedura iterativa di costruzione delle foglie. Un criterio intuitivo potrebbe essere quello di fissare una soglia minima s per la diminuzione di impurità, come espressa dall'Equazione 23, al di sotto della quale l'algoritmo si arresta. La scelta della specifica soglia s influenza dunque fortemente sia i risultati restituiti dall'algoritmo, sia il relativo carico computazionale.

In particolare, scegliendo una soglia molto piccola si otterrà un albero molto profondo (cioè con un numero elevato di livelli e di foglie), peggiorando la capacità interpretativa dello stesso, mentre con una soglia elevata si rischia di arrestare l'algoritmo molto presto a un dato nodo n , anche se i relativi nodi figli potrebbero presentare riduzioni di impurità maggiori della soglia.

L'algoritmo CART introduce dunque una metodologia per la definizione del criterio di arresto, denominata di *pruning* (potatura), che procede secondo i seguenti passi:

- 1) creazione dell'albero con dimensione massima, attraverso una soglia nulla $s=0$. Per definizione, tale albero contiene foglie costituite da casi completamente omogenei sulla variabile di uscita, fino al caso limite di foglie contenenti un solo caso;
- 2) selezione di una serie di sottoalberi, ottenuti a partire dall'albero con dimensione massima, "sfrondandolo" in alcuni punti e calcolo delle relative stime del tasso di errata classificazione dei diversi sottoalberi, attraverso uno stimatore appropriato;
- 3) scelta del sottoalbero migliore in termini di errata classificazione stimata, che è dunque quello che presenta errore minimo.

In realtà, il numero di sottoalberi da valutare al passo 2 è spesso molto elevato, anche quando le foglie dell'albero con dimensione massima è contenuto. Per limitare comunque il carico computazionale dell'algoritmo, vengono dunque introdotte delle procedure di *pruning* selettivo, che generano una sequenza di sottoalberi di dimensione decrescente, fino ad arrivare al nodo radice. I sottoalberi appartenenti a tale sequenza sono ottimali, nel senso che

essi sono quelli con il minor errore di classificazione all'interno delle rispettive classi, intese come l'insieme di tutti i sottoalberi con uno stesso numero di nodi foglia.

Per un approfondimento delle metodologie di *pruning* si veda il testo citato in apertura (Cerioli e Grossi, 2009).

4.3.3 QUEST

L'ultimo algoritmo presentato è il QUEST (Quick, Unbiased and Efficient Statistical Tree) (Loh e Shih, 1997). Si tratta anche dell'algoritmo più recente tra quelli inclusi nella presente disamina.

La metodologia CART è definita esaustiva, poiché la scelta del migliore *split* è effettuata a seguito di un'analisi di tutte le possibili suddivisioni su ogni variabile di input. Soprattutto nei casi in cui le variabili di input presentano un numero elevato di modalità o sono continue, il numero di possibili *split* diventa dunque molto elevato. La complessità computazionale del CART, ed in generale dei metodi esaustivi, cresce infatti linearmente nel caso di input continui e addirittura esponenzialmente nel caso di categoriali. Inoltre, i metodi esaustivi presentano una distorsione nella selezione delle variabili, poiché tendono a selezionare le variabili di input che presentano un numero maggiore di *split* (Loh e Shih, 1997).

Il metodo QUEST è stato dunque proposto come perfezionamento di un altro metodo, il FACT (Fast Algorithm for Classification Trees) (Loh e Vanichsetakul, 1988), proprio al fine di superare tali problemi. Esso impiega metodi statistici quali l'analisi della varianza (ANOVA) e l'analisi discriminante quadratica (QDA). Esso ha la caratteristica peculiare di separare la fase di selezione delle variabili da quella di selezione dello specifico *split* di ciascun nodo.

Per quanto riguarda la prima, al fine di evitare la distorsione propria dei metodi esaustivi, la selezione delle variabili viene eseguita mediante l'impiego di un test per la verifica d'ipotesi. In particolare, se esse sono di tipo categoriale viene eseguito un test chi-quadro per verificare l'indipendenza rispetto alle modalità della variabile dipendente. Invece, nel caso di variabili di input di tipo quantitativo, l'algoritmo QUEST prevede la verifica della significatività delle differenze fra le medie calcolate sulle diverse classi dell'output, attraverso appunto un'ANOVA.

Per quanto riguarda invece la scelta dello specifico *split* ottimo, l'algoritmo proposto da Loh e Shih impiega, come si diceva, una variante dell'analisi delle discriminanti quadratica (McLachlan, 1992), utilizzando la variabile selezionata nel passo precedente come esplicativa. La limitazione della QDA è che essa è applicabile solo nel di esplicative continue, e dunque l'algoritmo QUEST effettua una trasformazione delle variabili categoriali in continue, al fine di poter analizzare anche queste.

Infine, il QUEST prevede la possibilità di utilizzare anch'esso il metodo di *pruning* dell'algoritmo CART (cfr. paragrafo 4.3.2.2), al fine di arrestare la crescita dell'albero. Al contrario, esso non consente di utilizzare la misura dell'impurità del CART, perché essa non viene definita.

Questo algoritmo porta a una sostanziale riduzione del carico computazionale, rispetto al CART, soprattutto quando sono presenti variabili categoriali con molte modalità. Dunque, esso è il metodo di elezione per le necessità di questa ricerca (cfr. paragrafo 6.1.3).

4.4 Analisi per classi latenti (LCA)

L'analisi per classi latenti (Latent Class Analysis - LCA) (Lazarsfeld, 1950; Linzer e Lewis, 2011; Oberski, 2016) è una tecnica statistica multivariata per l'analisi dei dati categoriali. Quando sono presenti più variabili categoriche osservate in relazione tra loro, all'interno di un dataset derivante da un'indagine campionaria, è spesso interessante approfondire le cause di interazione tra di esse, costruendo cluster di comportamento omogeneo. Tale interazione potrebbe essere spiegata anche dall'influenza di una variabile non osservata, che presenta tante categorie differenti quanti sono i cluster di comportamento omogeneo appena citati, denominati appunto *classi latenti*.

I modelli di classi latenti puntano allora a stratificare la tavola di classificazione incrociata delle variabili osservate (chiamate anche "manifeste"), attraverso la costruzione di una variabile detta appunto latente, non ordinata. All'interno di ciascuna categoria di questa nuova variabile, le interazioni tra le variabili manifeste saranno annullate, il che equivale a dire che, condizionati all'appartenenza a una determinata classe latente, le variabili manifeste sono assunte essere statisticamente indipendenti.

4.4.1 Formulazione matematica dell'LCA

Il modello matematico sottostante all'analisi per classi latenti può essere espresso come segue. Indicando con y_j l'elemento j della risposta y , si stabilisce una funzione indicatore $I(y_j=r_j)$ che equivale a 1 quanto la risposta alla variabile j -esima è pari a r_j e 0 altrimenti. Allora, la probabilità di osservazione di un dato vettore di risposte è pari a

$$P(\mathbf{Y}=\mathbf{y}) = \sum_{c=1}^C \gamma_c \prod_{j=1}^J \prod_{r_j=1}^{R_j} \rho_{j,r_j|c}^{I(y_j=r_j)},$$

Equazione 25

dove γ_c è la probabilità di appartenenza alla classe latente c e $\rho_{j,r_j|c}^{I(y_j=r_j)}$ rappresenta la probabilità della risposta r_j nell'elemento j , condizionata dall'appartenenza alla classe latente c . I parametri γ rappresentano un vettore di probabilità di appartenenza di classe latente con somma 1. I parametri ρ costituiscono una matrice di probabilità delle classi dell'uscita condizionate all'appartenenza alle classi latenti. I gradi di libertà sono calcolati come il numero di possibili modelli di risposta (cioè il numero di celle nella tabella di contingenza formata dall'incrocio di tutti gli elementi osservati) meno il numero di parametri liberamente

stimati meno uno. La stima dei parametri è tipicamente eseguita usando un algoritmo EM (Expectation-Maximisation) (Borman, 2004).

4.4.2 Selezione dello specifico modello

La selezione del modello all'interno della LCA può essere condotta sia attraverso una misura assoluta dell'adattamento di un particolare modello sia attraverso il confronto delle misure di adattamento di più modelli concorrenti. Una misura comune della calibrazione assoluta nel caso di un modello categoriale è la statistica G^2 di tipo chi-quadro del rapporto delle probabilità, che in questo caso esamina l'ipotesi nulla che il modello LCA specificato si adatti ai dati in maniera corretta (Agresti, 2007). Questa statistica ha appunto una distribuzione asintotica chi-quadro. Dunque, quando la dimensione del campione è sufficientemente ampia e i gradi di libertà non sono troppo elevati, questo valore può essere paragonato alla distribuzione chi-quadro con gradi di libertà forniti dal modello LCA. Un valore p significativo indica la mancanza di forma del modello in termini assoluti. È anche possibile derivare empiricamente un valore p per questa statistica utilizzando un bootstrap parametrico (Collins *et al.*, 1993).

Al fine di confrontare la misura relativa di modelli con diversi numeri di classi latenti è possibile utilizzare diversi metodi informativi: AIC (Akaike, 1974), BIC (Schwarz, 1978), CAIC (Bozdogan, 1987) e aBIC (Schlove, 1987). In tutti questi criteri di informazione, un valore inferiore suggerisce un equilibrio ottimale tra l'adattamento ai dati e la parsimonia. Per approfondimenti sulla fase di selezione del modello si veda (Lanza e Collins, 2010).

Così come avviene nel caso di altri tipi più standard di modelli con equazioni strutturali, le variabili indipendenti, che si pensa possano influire sull'appartenenza alle differenti classi latenti, vengono incluse nel modello e viene verificato se l'invarianza nella misura della variabile latente tra i gruppi persiste, assumendo dunque che il modello di misura possa essere applicato a tutti i gruppi col fine di esaminare le differenze di gruppo nelle prevalenze delle classi latenti. Per testare tale invarianza, possono essere paragonate la distribuzione chi-quadro per la differenza nei gradi di libertà dei modelli e la differenza nella statistica G^2 tra un modello con probabilità di risposta stimate in ciascun gruppo senza vincoli e un modello in cui tali probabilità vengono vincolate all'uguaglianza tra i diversi gruppi. Un valore p significativo indica prove di misurazioni diverse tra i gruppi.

Le covariate possono anche essere incluse nella LCA per esaminare l'associazione tra esse e l'appartenenza alle classi latenti. L'impatto delle covariate sono modellate utilizzando una *link function* logistica, stimando i coefficienti di regressione logistica per ciascuna di esse (Dayton e Macready, 1988). Per l'interpretazione di tali stime, si veda il paragrafo 4.1.3.

È importante notare che il modello di misurazione della classe latente sulle variabili manifeste e il modello di previsione attraverso le variabili indipendenti sono stimati simultaneamente. È rilevante osservare che non è conosciuta a priori l'appartenenza a una data classe di ogni individuo, ma l'algoritmo restituisce la probabilità di ciascun individuo di appartenere a ciascuna delle classi ipotizzate.

5 Un nuovo modello di digital divide

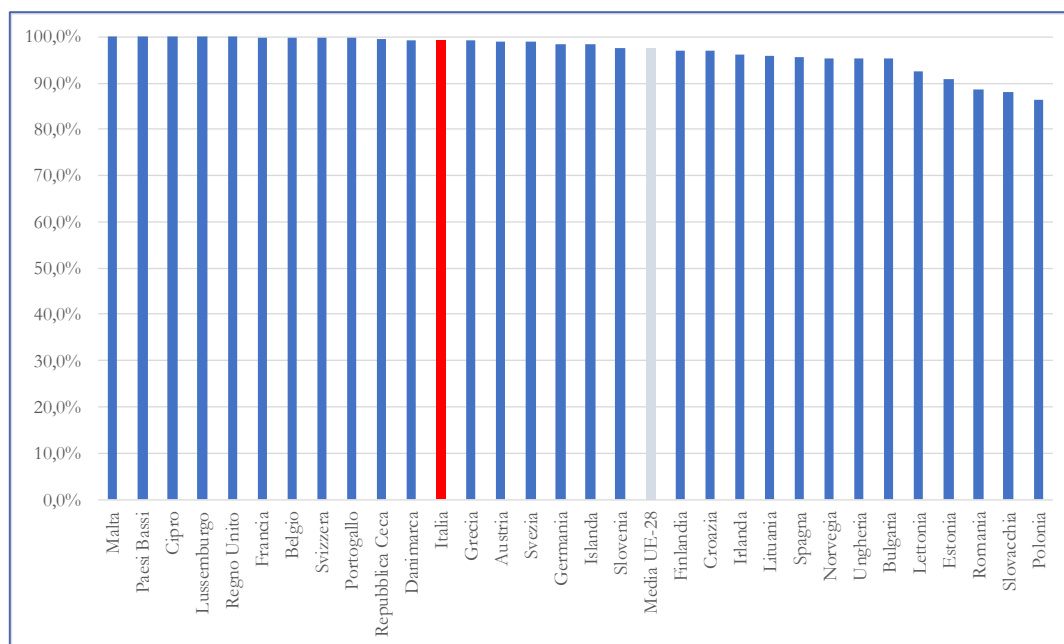
Come si diceva nell'introduzione e nel capitolo 0, l'analisi della domanda Internet, e dunque del digital divide, si estrinseca necessariamente su diversi livelli, poiché si tratta di un concetto ampio, che abbraccia una varietà di questioni legate ai diversi tassi di diffusione di Internet in diversi segmenti della popolazione.

5.1 Digital divide infrastrutturale

Per molto tempo, negli anni della prima diffusione delle reti fisse a banda larga in tutta Europa, il termine *digital divide* faceva riferimento sostanzialmente al solo divario infrastrutturale, ovvero alla mancanza di copertura dei servizi Internet a banda larga in alcune aree del paese e per la relativa popolazione. In quegli anni, in cui una buona porzione della popolazione, anche nei paesi più avanzati, non era in grado di adottare una connessione a Internet a banda larga a causa della mancanza di disponibilità, questo era il problema principale che i decisori politici si trovavano a dover affrontare.

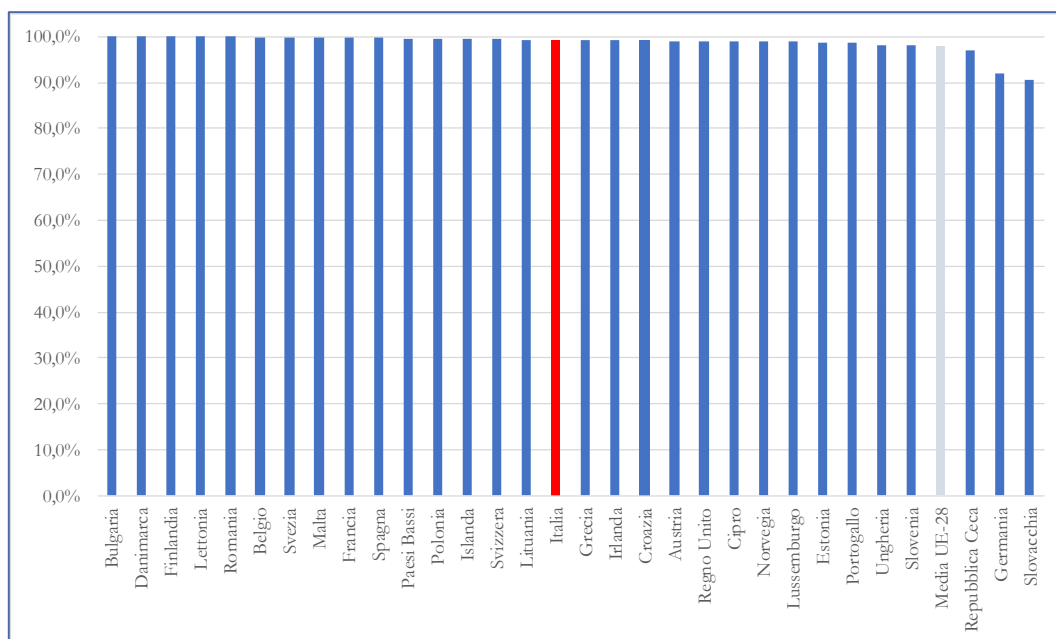
Negli ultimi anni, tuttavia, la copertura a banda larga (con velocità di almeno 2 Mbps) non può più essere considerata un problema in Europa, e nello specifico in Italia. I dati della *Digital Agenda Scoreboard* (Commissione Europea, 2017) mostrano, infatti, come nei 28 Stati Membri dell'Unione Europea il 97,5% delle famiglie è raggiunta da servizi a banda larga fissa, con un dato pari al 99,3% per il nostro paese. La mancanza di copertura continua a riguardare quote significative delle famiglie solo in alcuni paesi dell'Europa dell'Est (Polonia, Slovacchia, Romania, Estonia e Lettonia) che sono gli unici Stati Membri dell'UE con una percentuale di famiglie in *divide* da banda larga fissa oltre il 5% (Figura 22).

Figura 22 - Quota di famiglie raggiunte da banda larga fissa (anno 2016). Fonte: Commissione Europea, Digital Agenda Scoreboard



Quando si considerano i dati omologhi, riguardanti le connessioni mobili a banda larga (almeno con tecnologia HSPA), ci si trova di fronte a una situazione del tutto paragonabile. Anche in questo caso, una quota molto elevata delle famiglie europee (98,0%) risulta coperta da una connessione di questo tipo, e solo tre paesi presentano quote al di sotto del 95%: Germania, Irlanda e Slovacchia (Figura 23). Per quanto riguarda l'Italia, il valore di copertura riguarda la quasi totalità delle famiglie (99,3%), così come nel caso della banda larga fissa.

Figura 23 – Quota di famiglie raggiunte da tecnologie mobili ad alta velocità (almeno HSPA), per paese (anno 2016). Fonte: Commissione Europea, Digital Agenda Scoreboard



Ad ogni buon conto, le questioni infrastrutturali potrebbero tornare rilevanti qualora nuovi servizi Internet rendessero decisiva la disponibilità di connessioni più veloci, quali quelle a banda ultralarga. Potrebbe essere questo il caso dell'*Internet of Things* (Al-Fuqaha *et al.*, 2015; Li, Xu e Zhao, 2015; Rose, Eldridge e Chapin, 2015), con particolare riferimento alle applicazioni di domotica. Inoltre, nuove necessità di banda potrebbero derivare dal crescente consumo di video in alta definizione in simultanea da parte di più membri della famiglia, e/o dall'uso massivo di servizi *cloud*, non solo di archiviazione, ma anche di calcolo distribuito. A questo proposito, il nostro paese ha già predisposto un piano per la diffusione di reti di nuova generazione, denominato Strategia per la Banda Ultralarga (Presidenza del Consiglio dei Ministri, 2015).

5.2 Digital divide nell'adozione e nell'uso

Da quando le infrastrutture a banda larga, sia fisse che mobili, sono diventate così capillari, l'attenzione si è spostata sulle dinamiche di adozione della tecnologia, catturata dai due diversi fenomeni fortemente interrelati di cui si è parlato sin dall'introduzione, ovvero la sottoscrizione di un servizio ad Internet da parte della famiglia e la decisione di uso individuale.

In molti paesi, tra cui il nostro, in effetti il disinteresse e la mancanza di *skill* digitali, ancor più che problemi economici, fanno sì che una quota rilevante della popolazione non sottoscriva servizi di connessione o comunque non abbia mai utilizzato Internet (cfr. 3.5).

Nei casi in cui questa prima barriera ad un uso maturo di Internet viene superata, si manifestano però nuove forme di digital divide. Queste riguardano le modalità di utilizzo della Rete, in particolar modo la frequenza d'uso e le specifiche attività svolte online, che definiscono la ricchezza effettiva dell'uso (Van Deursen e Van Dijk, 2013, 2015; Van Deursen, Van Dijk e Ten Klooster, 2015).

Le analisi dimostrano che questi due aspetti sono correlati: gli utenti che accedono con più frequenza svolgono anche un numero maggiore di attività online, di crescente complessità e rilevanza per la loro vita economica e sociale. Internet diviene così uno strumento indispensabile non solo per lo svago, ma anche per la fruizione di servizi rilevanti per l'organizzazione della vita quotidiana. Queste analisi consentono infine di evidenziare contesti nei quali non è la maturità dell'utente finale, ma la carenza e la scarsa qualità dei servizi offerti, a limitarne la diffusione. In Italia ciò è evidente soprattutto per i servizi offerti dalle pubbliche amministrazioni.

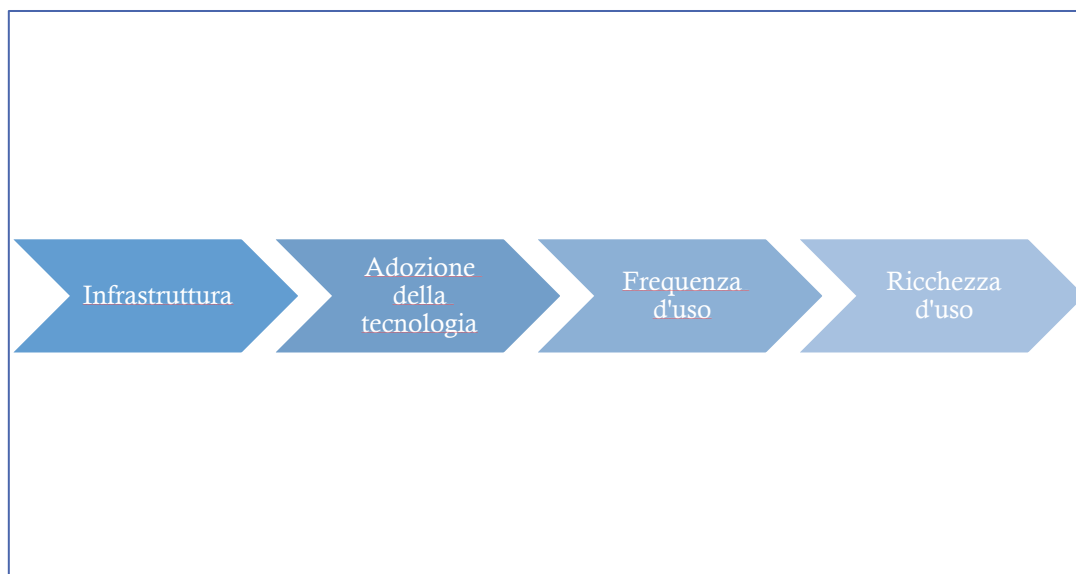
5.3 I quattro stadi del digital divide

Riassumendo il quadro appena descritto, si giunge dunque alla proposta di una nuova definizione di digital divide in quattro stadi rispetto a:

- 1) Infrastruttura: si tratta del primo stadio riguardante l'effettiva disponibilità del servizio, ormai superato per le connessioni a banda larga;
- 2) Adozione della tecnologia: l'effettiva decisione di utilizzare Internet, a livello individuale, e di sottoscrivere un servizio di connessione a banda larga, a livello familiare;
- 3) Frequenza d'uso: gli utenti che accedono ogni giorno possono essere considerati un tipo avanzato di utenza, rispetto a coloro che si collegano solo qualche volta o addirittura una sola volta a settimana;
- 4) Ricchezza d'uso: le differenze nelle specifiche tipologie di attività svolte online possono condurre a una diversa capacità di sfruttare le potenzialità di Internet.

La definizione appena descritta è riassunta in figura seguente.

Figura 24 -Rappresentazione del digital divide in quattro stadi



6 Analisi delle determinanti della domanda di Internet in Italia

In questo capitolo vengono riportate le analisi svolte seguendo il modello di digital divide in quattro stadi proposto nel capitolo precedente.

Come si diceva, al momento le caratteristiche dell'infrastruttura non sembrano svolgere un ruolo nel limitare l'adozione: i non utenti non sono in grado di apprezzare le eventuali carenze qualitative dei servizi di connettività, e per molti degli utenti le tipologie d'uso non richiedono prestazioni particolarmente elevate.

Dunque, in questa sede si procederà prima all'analisi delle determinanti dell'adozione, in termini di uso da parte degli individui e di sottoscrizione di una connessione da parte delle famiglie, che, come spiegato in altre parti del documento, risultano essere strettamente interconnesse (paragrafo 6.1). Le questioni relative alla frequenza d'uso e alla ricchezza d'uso saranno invece nella seconda parte del capitolo (paragrafo 6.2).

6.1 Uso individuale di Internet e adozione familiare della banda larga

L'analisi dello stadio di adozione è svolta inizialmente attraverso un modello di regressione logistica a più equazioni, tra i più utilizzati in letteratura laddove si ha la necessità di verificare e isolare l'impatto di una batteria (ristretta) di variabili indipendenti su di un output dicotomico o categoriale. In questo modello è anche stimato un indicatore economico che comprende diversi aspetti monitorati dall'Istat all'interno dell'indagine AVQ, di modo da dettagliare le risposte fornite dalle famiglie sul giudizio delle proprie risorse economiche (su una scala di 4 valori: ottime, buone, assolutamente insufficienti) e dagli individui sulla soddisfazione per la propria situazione economica (di nuovo su 4 valori: molto, abbastanza, poco o per niente soddisfatto). Il modello, applicato a un dataset contenente le rilevazioni AVQ dal 2005 al 2012, consente di ottenere una prima stima dell'impatto delle diverse variabili indipendenti, selezionate a partire dall'analisi della letteratura (cfr. cap. 0).

La complessità del modello logistico di uso di Internet da parte degli individui ha motivato un'analisi alternativa basata su tecniche di classificazione logica dei dati (paragrafo 6.1.2).

Infatti, sebbene il modello di regressione logistica a più equazioni sia soddisfacente in termini di bontà della calibrazione e interpretazione del fenomeno, un'analisi ulteriore dei dati può aiutare a stabilire se:

- 1) sia possibile ottenere spiegazioni semplici e differenti per diversi segmenti della popolazione;
- 2) esistano regole di classificazione alternative che possano fornire caratterizzazioni equivalenti dei dati osservati;
- 3) variabili aggiuntive rispetto a quelle considerate nel modello di regressione abbiano un qualche ruolo nella spiegazione delle variabili endogene;
- 4) variabili aggiuntive che non abbiano un effetto causale possano comunque migliorare la classificazione e la spiegazione dei fenomeni. Questo può accadere quando un input è associato a una variabile endogena per effetto di fattori latenti che influenzano entrambi.

Si tratta di un'analisi innovativa, svolta insieme agli autori dell'algoritmo SLAD (Bruni e Bianchi, 2015), basata su adattamenti a tale algoritmo, necessari per affrontare l'analisi di comportamenti socio-economici caratterizzati da scelte discrete, intrinsecamente aleatorie.

Sebbene queste procedure automatiche non forniscano un modello interpretativo vero e proprio, dato che non sono ideate per questo scopo, i pattern selezionati supportano letture alternative del fenomeno, attraverso il quale è possibile ottenere importanti spunti conoscitivi sulla segmentazione della popolazione. In particolare, essa consente di selezionare 19 variabili che hanno effettivamente impatto sull'uso di Internet. Tali variabili, opportunamente dicotomizzate sulla base delle informazioni provenienti dai pattern, in termini di impatto sull'uso di Internet, vengono poi suggerite a un algoritmo di costruzione di un albero di classificazione (paragrafo 6.1.3), che di fatto seleziona solo 7 di esse.

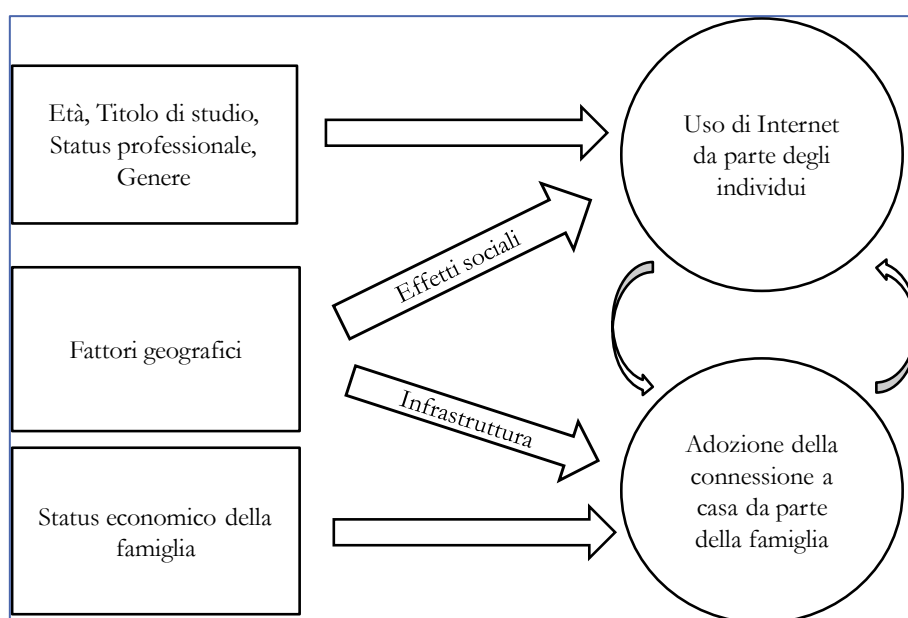
Da questa operazione si ottiene una partizione della popolazione italiana in 8 cluster, identificati con le lettere da A ad H. A partire dalle caratteristiche strutturali dei diversi segmenti di popolazione individuati, è possibile dunque studiare politiche di sostegno alla domanda Internet differenziate per cluster, e quindi più mirate a superare gli elementi ostativi che possono essere anche di natura molto diversa (culturale, di competenze, economica).

6.1.1 Analisi dell'adozione della tecnologia tramite regressioni logistiche⁹

Le prime analisi dei dati disponibili hanno come obiettivo la valutazione dell'utenza regolare, sulla base di una serie di variabili strutturali e sono svolte attraverso l'impiego della regressione logistica, applicata ai microdati dell'indagine Istat "Aspetti della Vita Quotidiana", provenienti dalle indagini degli anni 2005-2012 (cfr. par. 3.3). Il modello econometrico utilizzato è dunque del tipo logistico multinomiale, descritto all'interno del paragrafo 4.1.

Il primo modello logico di adozione ipotizzato, in cui le variabili di input sono selezionate a partire dall'analisi della letteratura (cfr. 2.1), è il seguente.

Figura 25 - Modello di adozione ipotizzato



Dunque, tale modello logico porta alla costruzione di due equazioni di regressione per i due diversi aspetti dell'adozione, a livello individuale e familiare.

L'uso di Internet da parte degli individui, almeno una volta a settimana, viene modellizzato dunque come segue.

⁹ Questo paragrafo proviene dal lavoro svolto da Claudio Leporelli (DIAG – Sapienza) e Cosimo Dolente, riportato in (Dolente e Leporelli, 2014).

$$\begin{aligned} & \text{Pr}(\text{uso Internet}) \\ & = f \left(\begin{array}{l} \text{genere, età, titolo di studio, condizione professionale,} \\ \text{posizione professionale, connessione a casa, ripartizione territoriale,} \\ \text{dimensione del comune, anno di rilevazione} \end{array} \right) \end{aligned}$$

L'uscita è codificata da una variabile dicotomica che rappresenta l'uso non solo a casa, ma in tutti i casi possibili (sul posto di lavoro, a scuola, a casa di amici, etc.).

I parametri calibrati di questa equazione di regressione sono riportati in appendice (Tabella A.1). Notiamo che i risultati confermano sostanzialmente le prime analisi svolte attraverso le statistiche descrittive (paragrafi 3.3 e 3.4), anche in un contesto in cui l'approccio multivariato ci consente di tener conto di effetti congiunti e correlazioni tra fattori.

L'età risulta essere il fattore che determina le maggiori differenze tra individui nell'uso di Internet. Dopo di essa, i fattori con la maggiore influenza sono il titolo di studio e la disponibilità di connessione a casa. Sono meno pronunciate, ma comunque significative ed in linea con le attese le influenze esercitate da condizione e posizione professionale, ripartizione geografica e dimensione del comune di residenza.

Questa prima regressione stima dunque la probabilità di utilizzo di Internet, almeno una volta alla settimana, per tutti i soggetti del campione. Per costruire un indicatore di interesse individuale verso l'utilizzo di Internet, non condizionato dalla disponibilità di una connessione, sono state poi ricalcolate le probabilità stimate dal modello in modo da farle corrispondere a quelle che il modello avrebbe previsto nel caso in cui tutti gli individui avessero avuto la disponibilità di un accesso ADSL a casa.

Il modello logico della Figura 25 prevede al secondo passo la stima dei fattori che determinano la scelta di avere una connessione da casa, da parte della famiglia. Tra gli altri fattori, è interessante verificare l'ipotesi per cui la situazione economica della famiglia giochi un ruolo importante nelle decisioni di adozione.

A tal fine, è costruito innanzitutto un indicatore della situazione economica della famiglia, che cerca di affinare le informazioni già presenti all'interno dell'indagine Istat AVQ. Ciò viene operato attraverso regressioni logistiche ordinali, variante delle regressioni logistiche multivariate per un'uscita per l'appunto ordinale, che mettono in relazione le variabili qualitative di situazione economica dell'indagine con alcune scelte di consumo effettuate ed alcune altre caratteristiche socio-demografiche di famiglie e individui.

In particolare, nella rilevazione Istat sono presenti due variabili distinte riguardanti la situazione economica. La prima, il giudizio sulle risorse economiche della famiglia, è una valutazione espressa dalla Persona di Riferimento della famiglia su una scala a 4 valori (ottime, adeguate, scarse, assolutamente insufficienti). La seconda, la soddisfazione per la propria situazione economica, è espressa invece dai singoli individui, ancora su una scala di 4 valori (molto, abbastanza, poco, per niente). Poiché le due variabili colgono aspetti differenti (giudizio della famiglia e percezione del singolo), sono dunque formulate due distinte regressioni, secondo i modelli seguenti.

$\text{Pr}(\text{giudizio sulle risorse economiche della famiglia}) =$

$$f \left(\begin{array}{l} \text{tipologia di nucleo, media della soddisfazione economica della famiglia} \\ \text{titolo di studio PR maschio,} \\ \text{titolo di studio PR femmina, condizione professionale PR maschio,} \\ \text{condizione professionale PR femmina, posizione professionale PR maschio,} \\ \text{posizione professionale PR femmina, numero di vacanze,} \\ \text{mancata vacanza per motivi economici,} \\ \text{assicurazione su infortuni o salute, assicurazione sulla vita,} \\ \text{possesso lavastoviglie, possesso antenna parabolica, possesso condizionatori,} \\ \text{numero di possessori di bancomat, numero di possessori di carte di credito,} \\ \text{problemi economici nell'abitazione, titolo di godimento dell'abitazione,} \\ \text{tipo di abitazione, distanza dell'abitazione dai servizi,} \\ \frac{h}{\text{sett}} \text{ collaboratrice domestica, consumi culturali,} \\ \text{n}^\circ \text{ di stanze, assicurazione contro i furti, n}^\circ \text{ di auto in possesso,} \\ \text{n}^\circ \text{ percettori di reddito, n}^\circ \text{ di persone mantenute dalla famiglia,} \\ \text{classe di età PR, dimensione del comune, ripartizione territoriale, anno} \end{array} \right)$$

e

Pr(soddisfazione economica) =

$$f \left(\begin{array}{l} \text{tipologia di nucleo, giudizio sulle risorse economiche della famiglia,} \\ \text{titolo di studio PR maschio,} \\ \text{titolo di studio PR femmina, condizione professionale PR maschio,} \\ \text{condizione professionale PR femmina, posizione professionale PR maschio,} \\ \text{posizione professionale PR femmina, numero di vacanze,} \\ \text{mancata vacanza per motivi economici,} \\ \text{assicurazione su infortuni o salute, assicurazione sulla vita,} \\ \text{possesso lavastoviglie, possesso antenna parabolica, possesso condizionatori,} \\ \text{numero di possessori di bancomat, numero di possessori di carte di credito,} \\ \text{problemi economici nell'abitazione, titolo di godimento dell'abitazione,} \\ \text{tipo di abitazione, distanza dell'abitazione dai servizi,} \\ \frac{h}{\text{sett}} \text{ collaboratrice domestica, consumi culturali,} \\ \text{n° di stanze, assicurazione contro i furti, n° di auto in possesso,} \\ \text{n° percettori di reddito, n° di persone mantenute dalla famiglia,} \\ \text{classe di età, dimensione del comune, ripartizione territoriale, anno} \end{array} \right)$$

Le variabili di input selezionate sono quelle socio-demografiche relative alla famiglia e alle Persone di Riferimento maschio e femmina, in modo da fornire una descrizione più possibile precisa dello status del nucleo familiare, nonché un insieme di indicatori di consumi e di possedimenti, che possono fornire un quadro più dettagliato delle reali condizioni economiche della famiglia. Di non immediata comprensione potrebbero risultare alcune variabili, che descriviamo più nel dettaglio di seguito:

- **tipologia di nucleo:** si tratta di una variabile che sintetizza la struttura della famiglia, in termini di composizione e relativi ruoli. Le modalità sono infatti: coppia con figli, coppia senza figli, monogenitore padre, monogenitore madre, nessun nucleo. Si noti che per “nessun nucleo” si intende una famiglia in cui non vi siano legami assimilabili al matrimonio e alle relazioni genitori-figli;
- **numero di vacanze:** numero di vacanze (di almeno 4 notti) della famiglia, negli ultimi 12 mesi;
- **problemi economici dell’abitazione:** si tratta di due variabili che indicano se la Persona di Riferimento ha segnalato che l’abitazione di residenza è troppo piccola o che le spese per l’abitazione sono considerate troppo alte;
- **distanza dell’abitazione dai servizi:** si tratta di una variabile che sintetizza l’insieme delle difficoltà della famiglia a raggiungere servizi quali scuole, farmacie, negozi, uffici pubblici;

- consumi culturali: si tratta di una variabile che sintetizza i consumi culturali della famiglia, in termini di fruizione nell'ultimo anno, da parte di un qualsiasi componente della famiglia, di spettacoli teatrali, cinema, lettura di quotidiani, libri, spettacoli musicali, musei, ecc.

Entrambi i modelli sono stimati tramite regressioni logistiche multinomiali ordinali, in quanto nelle variabili di output esiste un ordinamento tra le categorie, col valore più basso più favorevole. Le stime parametriche dei coefficienti della regressione logistica, a differenza di quanto avviene per la regressione lineare, non sono di facile interpretazione (cfr. par. 4.1.3). Inoltre, nella regressione ordinale la presenza di soglie di valori per le diverse categorie della risposta rendono ancora più difficile valutare l'importanza delle variabili solamente attraverso la lettura della tabella¹⁰. Essi vengono comunque riportati in appendice (Tabella A.2 e Tabella A.3).

Ad ogni modo, la calibrazione dei due modelli consentono il calcolo delle probabilità legate a ciascuno dei quattro livelli di ognuna delle due variabili. Si procede dunque alla costruzione di un unico indicatore, effettuando dapprima una media delle probabilità stimate in base alle soddisfazioni individuali dei componenti di ciascuna famiglia e poi una media tra questo valore e la probabilità stimata a partire dal giudizio della Persona di Riferimento. Infine, sono sommate da un lato le probabilità di condizione economica ottima o adeguata e, dall'altro, quelle di condizione scarsa o assolutamente insufficiente. Ne risulta un indicatore continuo tra 0 e 1, che esprime la probabilità che la famiglia abbia una situazione economica ottima o adeguata.

A questo punto, è dunque possibile definire la regressione logistica multinomiale che stima le probabilità di adozione di diverse tipologie di connessione a Internet da casa.

Rispetto alla variabile di uscita, in realtà, la domanda del questionario Istat AVQ prevede la possibilità di una risposta multipla. Per questo motivo, al fine di classificare le famiglie sulla base di categorie univoche, si procede ad assegnare ciascuna famiglia a una tipologia di connessione, secondo la gerarchia fissata di seguito:

¹⁰ Per una descrizione della regressione logistica ordinale, si veda (Agresti, 2007).

1. DSL (ADSL, SHDSL, VDSL ecc.) o connessione a banda larga via cavo (fibra ottica, rete locale, PLC, ecc.) o connessione senza fili fissa (satellite, rete pubblica WIFI, WiMax);
2. Connessione senza fili mobile con palmare o telefonino 3G (UMTS, mobile, ecc.) o connessione senza fili mobile con modem 3G (chiavetta USB, scheda);
3. Linea telefonica tradizionale o ISDN o connessione senza fili mobile a banda stretta (GPRS, 2G+ ecc.);
4. Nessuna connessione.

Questo tipo di riclassificazione è basata sul fatto che le connessioni a banda larga fissa sono solitamente erogate dagli operatori senza alcun vincolo sulla quantità di dati scambiati mensilmente, mentre le connessioni a banda larga mobile presentano normalmente questo tipo di vincolo, almeno al momento in cui viene condotta l'analisi. Seguono, nell'ordine, le connessioni *narrowband*, sia fisse che mobile, date le loro basse velocità di trasmissione, e infine le famiglie senza alcuna connessione.

Il modello logit per l'adozione della connessione a Internet da parte delle famiglie è ipotizzato essere il seguente.

$\Pr(\text{connessione internet}) =$

$$f \left(\begin{array}{l} \text{situazione economica, anno, ripartizione territoriale,} \\ \text{dimensione del comune, età della PR, numero di componenti, età minima,} \\ \text{età media, giudizio sulle risorse economiche} \\ * \text{probabilità di utilizzo internet} * \text{condizione occupazionale} \end{array} \right)$$

In particolare, l'indicatore di situazione economica è stato riportato come variabile continua e sono state aggiunte cinque variabili che riportano la probabilità di situazione economica ottima o adeguata da 0 fino a cinque soglie differenti (0,1, 0,3, 0,5, 0,7 e 0,9), e 0 altrove. L'inserimento delle variabili di soglia economica è effettuato al fine di cogliere eventuali effetti non lineari della situazione economica, che con l'introduzione di una sola variabile continua su tutto il dominio di probabilità non potrebbero essere evidenziati.

Le variabili di interazione tra giudizio sulle risorse economiche, probabilità di utilizzo Internet (calcolata dal primo modello) e condizione professionale sono state introdotte al fine di cogliere meglio il contributo derivante dal potenziale utilizzo di Internet da parte di componenti in differenti condizioni occupazionali, nei diversi contesti economici. Infatti, ci si attende che il contributo della variabile di interesse vari a seconda che i componenti siano

occupati, in cerca di occupazione, studenti o inattivi, e che tale effetto sia diverso a seconda che la famiglia abbia un giudizio sulle risorse economiche basso o alto.

Tale modello permette quindi di ottenere delle probabilità per ciascuna tipologia di connessione ad Internet per ogni famiglia, sulla base dei parametri stimati, riportati in allegato (Tabella A.4).

Dall'analisi dei risultati di questa regressione emerge dunque come l'impatto della variabile economica costruita nei passi precedenti, catturato tramite l'effetto congiunto la variabile completa e le 5 variabili di soglia, sia sicuramente importante.

Tra le variabili più importanti per l'adozione di banda larga sembrano esserci anche quelle legate all'età, nelle sue varie declinazioni. L'età minima sembra essere la più importante fra queste (minore è e meglio è, con alcune oscillazioni), indicando un contributo di curiosità e carica innovativa delle fasce d'età più giovani, nonché una probabile correlazione con l'età giovane anche della Persona di Riferimento della famiglia. L'età media, invece, sembra penalizzare le famiglie più giovani, probabilmente per questioni di reddito associate alla fase iniziale dell'attività lavorativa.

L'effetto del numero dei componenti della famiglia è positivo sull'adozione di DSL e altri tipi di banda larga, ma diminuisce al crescere della variabile, anche in questo caso, probabilmente per motivi di reddito pro capite delle famiglie più numerose.

La dimensione del comune ha un significativo effetto sul processo di adozione del DSL, in cui i comuni centro di area metropolitana sono favoriti e i comuni molto piccoli, fino a 2.000 abitanti, fortemente sfavoriti, ma anche sull'adozione dei altri tipi di banda larga, ivi compresa quella mobile, in cui i comuni più favoriti, dopo il centro delle aree metropolitane, sono proprio i comuni più piccoli. Si tratta di un risultato molto interessante, che sembrerebbe confermare l'ipotesi per cui le famiglie residenti nei comuni ancora in digital divide infrastrutturale da linea fissa si orientino verso le connessioni mobili e *fixed-wireless*, laddove possibile. Al contrario, l'effetto della ripartizione territoriale può essere considerato come trascurabile, con il Sud solo lievemente svantaggiato rispetto alle altre ripartizioni nell'adozione di DSL e altri tipi di connessioni a banda larga.

Come ci si poteva attendere, l'effetto del passare del tempo, e quindi dell'avanzare naturale del processo di diffusione tra la popolazione, misurato dalla variabile "Anno", è positivo sull'adozione della banda larga.

Infine, l'effetto dell'interazione tra somma delle probabilità di utilizzo dei componenti, condizione professionale e giudizio sulle risorse economiche della famiglia (dicotomizzato) è importante per tutte e tre le tipologie di connessione, con effetti molto simili sulle varie modalità delle variabili categoriali in interazione. Tuttavia, l'interesse ha un impatto massimo sull'adozione di DSL nel caso in cui esso provenga da inattivi in famiglie poco abbienti. Tale risultato è molto interessante, poiché implica che qualora l'interesse delle persone inattive e poco abbienti, presso cui Internet è meno diffusa, dovesse salire di una qualche misura, l'adozione di DSL potrebbe aumentare molto velocemente.

Il modello appena esposto ha il principale svantaggio di contenere la circolarità tra uso di Internet e adozione della connessione da parte della famiglia, propria del fenomeno. In statistica, questa circolarità per cui, per l'appunto, alcune variabili (in questo caso due) sono inserite a volte come dipendenti e a volte come indipendenti all'interno del modello, è nota come endogeneità. Ciò introduce un problema metodologico, perché potrebbe portare a inconsistenza nelle stime (Heckman, 1978; Nerlove, 1978; Bascle, 2008). C'è da osservare che il ruolo preponderante nella stima dell'utilizzo è svolto dalle variabili socio-economiche e demografiche, chiaramente esogene, mentre il possesso di una connessione a casa modifica solo marginalmente la propensione all'uso da parte degli utenti meno motivati. D'altra parte, dai dati emerge che, anche se la famiglia dispone di una connessione ad Internet, i componenti svantaggiati, culturalmente o per età, spesso non utilizzano ugualmente il servizio.

6.1.2 Analisi esplorativa con classificazione SLAD¹¹

Data la complessità del modello logistico di uso di Internet da parte degli individui, e al fine di ottenere delle informazioni aggiuntive sulle variabili contenute nel database Istat AVQ, viene qui utilizzata, in maniera alternativa all'uso più diffuso di questo tipo di algoritmi, la tecnica di classificazione SLAD, così come definito nel corso del paragrafo 4.2.

Si tratta di un'analisi innovativa, svolta insieme agli autori dell'algoritmo SLAD (Bruni e Bianchi, 2015), basata su adattamenti a tale algoritmo, necessari per affrontare l'analisi di comportamenti socio-economici caratterizzati da scelte discrete, intrinsecamente aleatorie.

¹¹ Questo paragrafo deriva dal lavoro svolto dal gruppo composto da Renato Bruni (DIAG – Sapienza), Giampiero Bianchi (Istat), Claudio Leporelli (DIAG – Sapienza) e Cosimo Dolente. Esso è contenuto all'interno dell'articolo (Bruni *et al.*, 2017).

Sebbene, infatti, il modello di regressione logistica a più equazioni del paragrafo 6.1.1 sia soddisfacente in termini di bontà della calibrazione e interpretazione del fenomeno, un'analisi ulteriore dei dati può aiutare a stabilire se:

- 1) sia possibile ottenere spiegazioni semplici e differenti per diversi segmenti della popolazione, cioè se una classificazione dei casi in cluster può migliorare la spiegazione delle variabili endogene in ogni cluster; ad esempio, in alcuni segmenti della popolazione, Internet non viene utilizzato a casa di mancanza di interesse o di capacità, mentre in altri sono le barriere economiche ad impedire l'adozione di persone interessate;
- 2) esistano regole di classificazione alternative che possano fornire caratterizzazioni equivalenti dei dati osservati; per esempio può essere utilizzato l'effetto congiunto della condizione e della posizione professionale, anziché una variabile qualitativa sulla soddisfazione dell'individuo per il suo status economico;
- 3) variabili aggiuntive rispetto a quelle considerate nel modello di regressione abbiano un qualche ruolo nella spiegazione delle variabili endogene;
- 4) variabili aggiuntive che non abbiano un effetto causale possano comunque migliorare la classificazione e la spiegazione dei fenomeni. Questo può accadere quando un input è associato a una variabile endogena per effetto di fattori latenti che influenzano entrambi. Esempi rilevanti potrebbero essere rappresentati dall'associazione tra l'uso di Internet e l'utilizzo di carte di credito o la pensione al turismo o al consumo culturale.

Inoltre, risulta di particolare interesse verificare se l'approccio SLAD può aiutare a identificare il miglior sottoinsieme di queste variabili basati su criteri di precisione, generalità e parsimonia. Infatti, ci si aspetta che esso ignori le variabili fortemente correlate con quelle selezionate.

Il database utilizzato è ancora quello contenente i microdati dell'indagine "Aspetti della Vita Quotidiana" di Istat, relativi all'anno 2012. Da esso, è stato estratto un numero di variabili superiore a quello utilizzato nel modello precedente (dove per spiegare l'uso individuale di Internet si utilizzano 9 variabili), per poter perseguire al meglio gli obiettivi appena esposti. Sono considerate, infatti, 39 variabili, che descrivono una pluralità di possibili determinanti socio-economiche e culturali dell'uso di Internet. L'uscita, che all'interno di questo metodo

è denominata anche classe o etichetta, è nuovamente l'uso di Internet regolare (almeno una volta a settimana). Il campione di analisi è limitato ai soli rispondenti tra 15 e 74 anni, perché per i minori di 15 anni non sono rilevate le variabili relative all'occupazione, mentre per gli ultra settantaquattrenni alcune variabili perdono di senso, poiché le loro caratteristiche risultano essere fortemente polarizzate sulle variabili di segmentazione.

Sono stati dunque eseguiti quattro diversi esperimenti utilizzando l'algoritmo SLAD, con e senza le procedure 1 e 2 proposte (cfr. 4.2.2.1 e 4.2.2.2), per valutarne i vantaggi, anche dal punto di vista metodologico.

I risultati sono riassunti in Tabella 13, mostrando per ciascun test i parametri di ingresso e i seguenti indicatori di prestazioni dell'output, per i pattern di ciascuna classe:

- il numero totale di pattern selezionati, che risulta essere l'indicatore più rilevante dei problemi di ridondanza e frammentazione;
- il numero di pattern che forniscono una copertura incrementale positiva rispetto ai precedenti;
- la cardinalità massima della copertura di un unico pattern, che indica se è stato effettivamente ottenuto almeno un pattern con una certa generalità;
- la cardinalità media della copertura incrementale, calcolata sull'insieme di pattern selezionati, che indica la capacità dell'algoritmo di limitare la ridondanza dei pattern generati;
- il numero massimo di condizioni che appaiono in un unico pattern, utile per valutare la generalità e la leggibilità dei pattern selezionati;
- la percentuale di record nel campione coperto da tutta la serie di pattern selezionati, al fine di fornire una misura dell'idoneità dell'algoritmo nell'applicazione proposta;
- il numero medio di pattern che coprono correttamente un singolo record, indicatore utile della ridondanza dei pattern selezionati.

Da un punto di vista metodologico, si noti che quando si utilizzano le tecniche SLAD di base (test 1 e 2), la copertura incrementale di ciascun modello viene calcolata ex-post, dopo aver ordinato i modelli sulla base della copertura totale discendente. Quando si usano invece le procedure 1 e 2 proposte (cfr. 4.2.2.1 e 4.2.2.2, rispettivamente), i pattern senza copertura incrementale vengono invece esclusi direttamente dall'algoritmo.

Tabella 13 - Riassunto dei risultati nei diversi test della metodologia SLAD

| | | Test 1 Standard SLAD | Test 2 SLAD con errore | Test 3 Procedura 1 | Test 4 Procedura 2 |
|-------------------------------|--|----------------------------|------------------------------|--------------------------|--------------------------|
| Parametri di input | ϵ^+ | 10 | 5 | 10 | 10 |
| | ϵ | 10 | 10 | 10 | 10 |
| | ϵ^+ | 0% | 20% | 20% | 40% |
| | ϵ^- | 0% | 10% | 20% | 40% |
| Pattern positivi | Totale selezionati | 6.65 | 6.761 | 198 | 7 |
| | Con copertura incrementale non nulla | 1.314 | 420 | 198 | 7 |
| | Copertura incrementale media | 1,1 | 2,6 | 87 | 2.057 |
| | Copertura massima | 164 | 6.544 | 8.096 | 8.096 |
| | Numero massimo di condizione | 10 | 10 | 6 | 2 |
| | % di record coperti (positivi) | 42% | 96% | 96% | 80% |
| | Numero medio di pattern che copre ciascun record | 21 | 148 | 85 | 1 |
| Pattern negativi | Totale selezionati | 72.682 | 10.856 | 243 | 7 |
| | Con copertura incrementale non nulla | 2.621 | 1.034 | 243 | 7 |
| | Copertura incrementale media | 0,2 | 1,3 | 65 | 1.872 |
| | Copertura massima | 1.673 | 4.175 | 8.144 | 7.236 |
| | Numero massimo di condizione | 6 | 13 | 8 | 3 |
| | % di record coperti (positivi) | 90% | 86% | 96% | 80% |
| | Numero medio di pattern che copre ciascun record | 341 | 114 | 98 | 1 |

Nel test 1 i pattern sono generati senza la possibilità di ammettere errori, cioè casi coperti di classe errata. In realtà, all'interno della nostra applicazione, ciò significa ignorare la natura probabilistica del fenomeno, cercando sottoinsiemi del campione caratterizzati integralmente

dallo stesso valore della classe di output. In questo modo, si ottiene un numero elevato di pattern, che hanno al contempo una copertura massima e media abbastanza ridotta. E, ancor più importante, solo 7.641 dei 18.077 record positivi del campione (circa il 42%) sono effettivamente coperti, mentre il numero sale sensibilmente per quel che riguarda i record negativi (cioè i non utenti), pari a 14.762 su 16.378 (circa il 90%). Si tratta di un risultato interessante, che suggerisce che è molto più facile trovare categorie omogenee di persone che non utilizzano affatto Internet, rispetto a quelle di utenti. I pattern selezionati, in particolare quelli negativi, presentano poi elevate sovrapposizioni: ogni record positivo (negativo) identificato è coperto da 21 (rispettivamente, 341) pattern, in media. Ciò può essere dovuto principalmente alla correlazione tra variabili esplicative diverse. I risultati di questa prima prova non sono chiaramente soddisfacenti: il grande numero di pattern, anche limitandosi a quelli che forniscono una copertura incrementale non nulla, rende difficile ottenere informazioni addizionali sul fenomeno che siano di un qualche valore esplicativo.

Per migliorare i risultati appena esposti, è possibile modificare alcuni parametri, utilizzando ancora solo la generazione standard di pattern dell'algorithm SLAD. In primo luogo, viene concesso ai pattern di selezionare una data percentuale massima di record della classe opposta, detta di errore. In secondo luogo, vengono utilizzati parametri asimmetrici, sotto l'assunzione che i record negativi sono più omogenei, derivante dai risultati del primo test. In particolare, il test 2 viene svolto impostando $\epsilon^+ = 5$, $\epsilon^+ \leq 20\%$, $\epsilon^- = 10$, $\epsilon^- \leq 10\%$. Si ottengono così 6.761 pattern positivi e 10.856 negativi. Sia le coperture massime che le medie sono più elevate che nel test 1, per entrambe le tipologie di pattern. Inoltre, c'è un miglioramento significativo della percentuale di record positivi coperti. Tuttavia, il numero di pattern risultanti è ancora troppo grande per fornire approfondimenti utili sul fenomeno. Questi risultati, derivanti da applicazioni dell'algorithm SLAD già presente in letteratura, evidenziano la necessità di procedure che siano commisurate alle specificità dei fenomeni di scelta discreta, e consentano dunque di esplorare i compromessi tra diversi obiettivi:

- il potere di discriminazione tra le due classi dell'uscita (utenti di Internet e non utenti), considerando al contempo che il comportamento probabilistico implica che ci si occupa di sottoinsiemi per loro natura non omogenei, e che le distribuzioni

delle classi di output in ciascuna regione dello spazio non sono conosciute in anticipo;

- la rilevanza, che indica che i pattern dovrebbero coprire effettivamente un numero elevato di osservazioni;
- la semplicità, che indica che i pattern dovrebbero essere costituiti dal minor numero possibile di condizioni;
- la non ridondanza, vale a dire che i pattern dovrebbero avere piccole sovrapposizioni (procedura 1) o nessuna sovrapposizione (procedura 2) tra loro.

Il test 3 è eseguito dunque impiegando la Procedura 1, che utilizza la copertura incrementale come criterio di ordinamento (cfr. 4.2.2.1). L' algoritmo seleziona così 198 pattern positivi e 243 pattern negativi, con un evidente miglioramento rispetto alle migliaia di pattern ottenuti nei primi due test. Inoltre, per aumentare la leggibilità dei risultati e quindi la comprensione del fenomeno, possono essere impostati alcuni criteri ex-post per ridurre ancora, e in maniera importante, il numero di pattern.

Se si scelgono pattern con una copertura incrementale di almeno 200 record e con un errore incrementale massimo inferiore al 45%, si ottengono allora solamente 8 pattern positivi che identificano correttamente l'84% degli utenti Internet nel campione e 5 pattern negativi che identificano correttamente il 75% dei non utenti. Il dettaglio di tali pattern selezionati è riportato nelle seguenti tabelle, iniziando dai pattern positivi (Tabella 14).

Le variabili che danno luogo alle condizioni sono state già introdotte o hanno un significato sufficientemente intuitivo. L'indice economico è quello derivante dalle elaborazioni effettuate con regressioni logistiche, riportate all'interno del paragrafo precedente. La variabile "consumi culturali" rappresenta un indice sintetico delle attività culturali *offline* svolte dall'individuo negli ultimi 12 mesi, quali ad esempio la partecipazione a concerti, eventi sportivi, visite a musei, al cinema, la lettura di libri, ed è qui dicotomizzata in due livelli: alti (almeno 2 attività) o bassi (1 o nessuna attività).

Tabella 14 - Pattern positivi prodotti nel test 3 dalla Procedura 1 della metodologia SLAD (anno 2012)

| Pattern | | | | | | | | |
|-------------------------|---|---|---|---|---|----|----|---|
| Condizioni | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| <i>Carta di credito</i> | | | | | | Si | Si | |

| Pattern | | | | | | | | |
|--|------------------------|-----------------|------------------------|-----------------|-----------|-----------|------------------------|------------------------|
| Condizioni | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| <i>Stato civile</i> | | Celibe o nubile | | Celibe o nubile | | | | |
| <i>Condizione professionale della PR femmina</i> | | | | | Occupata | | | Occupata |
| <i>Livello di istruzione</i> | Superiori o università | | Superiori o università | | | | Superiori o università | Superiori o università |
| <i>Età</i> | | | | <= 34 | | | | |
| <i>Condizione professionale</i> | | | Occupato | | | Occupato | | |
| <i>Consumi culturali</i> | Alti | Alti | | | Alti | | | |
| <i>Indice economico</i> | | | | | | >= 0,465 | | |
| <i>Ore lavorate a settimana</i> | | | >21 | | | | | |
| Copertura pesata | 10.991.933 | 8.787.901 | 9.524.593 | 8.522.194 | 8.182.859 | 5.898.051 | 7.497.830 | 8.256.792 |
| Copertura incrementale pesata | 10.991.933 | 2.981.266 | 2.978.311 | 1.365.471 | 950.800 | 653.613 | 283.305 | 305.607 |
| Errore % | 14,4% | 14,6% | 16,5% | 16,6% | 15,5% | 15,3% | 14,1% | 16,5% |
| Errore % incrementale | 14,4% | 23,8% | 27,6% | 36,3% | 39,7% | 38,0% | 44,4% | 42,2% |

Si noti come la copertura incrementale diminuisce abbastanza rapidamente, soprattutto dopo i primi 2-3 pattern nell'ordinamento. Al contempo, l'errore incrementale tende a salire, poiché vengono esclusi i casi comuni con i pattern precedenti, che hanno comunque bassi tassi di errore, e dunque, intuitivamente, la copertura incrementale viene decurtata in maniera più rilevante che non l'errore. Il pattern più ampio comprende circa 11.000.000 di utenti nella popolazione reale, ed è identificato dai consumi culturali elevati uniti a un titolo di studio di scuola superiore o universitario. L'errore, che come si diceva deve essere letto in questo caso come fisiologico tasso di non utenti in un dato cluster di popolazione, è qui del 14,4%, per cui si ha l'85,6% di utenza.

Passando ai pattern negativi prodotti dal test 3 e selezionati sulla base dei criteri esposti pocanzi, il dettaglio è riportato in Tabella 15.

Tabella 15 - Pattern negativi prodotti nel test 3 dalla Procedura 1 della metodologia SLAD (anno 2012)

| Pattern | | | | | |
|------------------------------|--------------|----------|--------------|--------------|--------------|
| Condizioni | 1 | 2 | 3 | 4 | 5 |
| <i>Livello di istruzione</i> | Medie o meno | | Medie o meno | Medie o meno | Medie o meno |

| Pattern | 1 | 2 | 3 | 4 | 5 |
|--------------------------------------|------------|-----------|-----------|-----------|-----------|
| Condizioni | | | | | |
| <i>Carta di credito</i> | No | No | | | No |
| <i>Consumi culturali</i> | Bassi | | Bassi | | |
| <i>Vacanze</i> | | | No | | No |
| <i>Età</i> | | >= 55 | | >= 55 | |
| <i>Stato civile</i> | | | | | Coniugato |
| Copertura pesata | | | | | |
| Errore % incrementale | 10.130.689 | 8.568.241 | 9.091.408 | 8.495.235 | 6.623.126 |
| Copertura incrementale pesata | 10.130.689 | 2.787.261 | 932.471 | 679.315 | 512.196 |
| Errore % | 17,0% | 13,7% | 16,7% | 11,2% | 15,2% |
| Errore % incrementale | 17,0% | 27,3% | 32,1% | 34,7% | 41,1% |

In questo caso, la copertura incrementale si riduce molto rapidamente, e già dal terzo pattern nell'ordinamento essa è pari a circa il 10% di quella totale. Il pattern più ampio di non utenti, che comprende più di 10.000.000 di individui (come nel caso degli utenti), è dunque rappresentato da persone che non possiedono una carta di credito, con livello di istruzione basso e bassi consumi culturali. Si noti che, a meno della carta di credito, si tratta di un pattern complementare al pattern più grande individuato per gli utenti. All'interno del pattern, gli individui sono non utenti di Internet nell'83% dei casi. Il secondo pattern è costituito invece da individui che non possiedono carta di credito e che hanno almeno 55 anni. Se si considera solo la copertura incrementale, si identificano dunque 2,8 milioni di persone che hanno o un titolo elevato o consumi culturali elevati, più di 54 anni e non possiedono la carta di credito, e che sono non utenti nel 76,7% dei casi. È un dato interessante, se si considera che un titolo elevato o dei consumi culturali elevati sono in genere associati all'utilizzo della Rete. Ci si trova qui di fronte a uno dei vantaggi di utilizzare questa metodologia, che permette di considerare le coperture incrementali: la possibilità di identificare in segmenti specifici della popolazione interazioni tra variabili non prevedibili a priori, che possono rendere più chiaro il fenomeno nelle sue determinanti.

Si noti come, in entrambi i casi, l'età non è considerata dai primissimi pattern, ma compaia nel secondo pattern negativo, come si diceva in associazione col non possesso della carta di credito, e nel quarto pattern positivo, da sola. Questo evidenzia ancora una volta la differenza

con altri tipi di metodologie laddove, in questo caso, il poter considerare l'interazione tra diverse variabili porta a letture originali del fenomeno, non ottenibili facilmente attraverso altre metodologie più tradizionali.

In media, utilizzando la procedura 1, e senza realizzare selezioni ex-post del tipo appena descritto, ogni record selezionato è ancora coperto da circa 3 pattern diversi. Dunque, abbiamo ancora una certa ridondanza, che può essere eliminata utilizzando la procedura 2, come avviene all'interno del test 4.

In questa procedura, un pattern aggiunto nella fase *i* della procedura seleziona record che non sono stati selezionati nei precedenti *i-1* passi. Quindi, ogni pattern viene generato solo se è strettamente necessario per coprire un minimo numero predefinito di record aggiuntivi. Inoltre, si può logicamente definire la regione dello spazio aggiunto al passo *i* come congiunzione delle condizioni nel pattern *i* con la negazione logica di tutti i precedenti *i-1* pattern (cfr. paragrafo 4.2.2.2). In linea di principio, questo consente di identificare le espressioni logiche di una partizione dell'insieme di tutti i record selezionati. Nella pratica, però, le espressioni risultanti possono essere particolarmente complesse. Ad ogni modo, la disgiunzione logica dei pattern identificati è in grado di fornire un'espressione leggibile per l'insieme dei record coperti, per ciascuna classe dell'output.

I risultati per i pattern positivi ottenuti all'interno del test 4, tramite l'uso della procedura 2, sono mostrati nella seguente tabella.

Tabella 16 - Pattern positivi prodotti nel test 4 dalla Procedura 2 della metodologia SLAD (anno 2012)

| Pattern | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Totale |
|--|------------------------------|-----------|-----------|----------------------|----------|----------|------------------------------|---------------|
| Condizioni | | | | | | | | |
| <i>Carta di credito</i> | | | Si | | | Si | | |
| <i>Fonte di reddito</i> | | | | Lavoro dipendente | | | | |
| <i>Condizione professionale</i> | | | Occupato | | | | | |
| <i>Possesso condizionatore</i> | | | | | Si | | | |
| <i>Livello di istruzione</i> | Superiori o università | | | | | | Superiori o università | |
| <i>Età</i> | | <= 34 | | | | | | |
| <i>Consumi culturali</i> | Alti | | | | | | | |
| Copertura incrementale pesata | 10.991.933 | 5.034.679 | 2.279.560 | 753.382 | 215.027 | 112.605 | 124.319 | 19.511.505 |
| Errore % | 14% | 26% | 24% | 37% | 37% | 37% | 37% | 21% |

Il numero totale di utenti coperti da questo insieme di pattern positivi è 14.399 su 18.077, pari all'80% del campione. In particolare, il primo pattern mostra come quasi 11 milioni di utenti (con un errore del 14%) sono descritti dalla congiunzione di due soli condizioni: il livello di istruzione almeno pari alla scuola superiore e un alto livello di consumi culturali. È interessante notare che il secondo pattern riguarda invece gli utenti con meno di 35 anni che non sono coperti dal pattern precedente, cioè non presentano un livello elevato di istruzione o di consumi culturali. Questo pattern è comunque abbastanza ampio, dato che copre 5 milioni di persone, con un'incidenza (o errore) del 26,2% di non utenti.

Applicando le leggi di De Morgan¹², è possibile semplificare la disgiunzione dei pattern positivi, ottenendo così la seguente espressione:

$$(Età \leq 34) \vee (Fonte \ di \ reddito = Lavoro \ salariato) \vee (Possesso \ di \ un \ condizionatore = sì) \vee (Carta \ di \ credito = sì) \vee (Istruzione = Scuola \ superiore \ o \ università)$$

Si tratta di un risultato interessante, che dimostra che l'uso di Internet è influenzato dalla giovane età, dalla fonte di reddito e dal livello di istruzione. L'algoritmo sceglie poi le ultime due variabili (il possesso di un condizionatore a casa e della carta di credito), probabilmente come le *proxy* più parsimoniose delle abitudini economiche degli individui. In particolare, in buona parte della popolazione italiana l'aria condizionata potrebbe ancora essere considerata un bene di lusso, mentre il possesso di una carta di credito potrebbe ancora rappresentare una *proxy* sia del reddito che di un atteggiamento positivo verso l'innovazione.

Passando ai pattern negativi, i risultati del test 4 sono riportati nella tabella seguente.

Tabella 17 - Pattern negativi prodotti nel test 4 dalla Procedura 2 della metodologia SLAD (anno 2012)

| Pattern | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Totale |
|---|---------------------|---------------------|----|-------|-----------|-----------|---------------------|--------|
| Condizione professionale | | | | | Casalinga | Casalinga | | |
| Numero di membri con reddito nella famiglia | | | | | | | 0-1 | |
| Area geografica | | | | | Sud | | | |
| Carta di credito | No | | No | | | No | | |
| Età | | | | >= 55 | | | | |
| Livello di istruzione | Scuola media o meno | Scuola media o meno | | | | | Scuola media o meno | |
| Parentela con la PR | | | | | Coniuge | Coniuge | | |

¹² Si tratta di una serie di leggi sull'equivalenza degli operatori logici.

| | | | | | | | | |
|--------------------------------------|-----------|-----------|-----------|-----------|--------|--------|---------|------------|
| <i>Stato civile</i> | Sposato | | | | | | | |
| <i>Consumi culturali</i> | | Basso | | | | | | |
| <i>Auto medie per membro</i> | | | | | | | <=0.585 | |
| <i>Vacanze negli ultimi 12 mesi</i> | | | No | | | | | |
| Copertura incrementale pesata | 9.044.929 | 4.499.515 | 1.823.726 | 1.021.704 | 38.329 | 30.888 | 19.226 | 16.478.317 |
| Errore % | 18% | 26% | 36% | 36% | 32% | 32% | 39% | 24% |

In questo caso, 13.102 su 16.378 non utenti sono coperti dai pattern selezionati, di nuovo l'80% del campione. Il più grande (7.236 record, pari a 9 milioni circa di persone) e il più omogeneo (18% di errore, cioè di utenti) è costituito da persone che non possiedono una carta di credito, hanno un livello di istruzione basso, e sono attualmente sposati. Altri 4,5 milioni di persone, inclusi nel secondo pattern, mostrano bassi livelli di istruzione e di consumi culturali, ma possiedono una carta di credito o sono attualmente sposati. I pattern negativi coinvolgono un numero complessivo maggiore di condizioni (11 vs 7) e non consentono il livello di semplificazione della classe precedente. Comunque, considerando solo i primi quattro modelli, che coprono la maggior parte dei casi nella partizione (13.032 su 13.102), otteniamo questa importante espressione:

$$(Et\grave{a} \geq 55) \vee (Edu = Medio\ o\ basso \wedge Cons. Culturale = basso) \vee (Carta\ di\ credito = No) \wedge \\ Vacanze = No) \vee (Edu = Medio\ o\ basso \wedge Carta\ di\ credito = No \wedge Stato\ civile = Sposato)$$

In conclusione, l'analisi logica dei dati genera un gran numero di pattern in grado di descrivere i dati. Attraverso il lavoro effettuato durante il presente progetto, e qui presentato, sono state messe a punto procedure valide, da un punto di vista computazionale, per ottenere pattern che possano essere significativi, da un punto di vista pratico, per le analisi socio-economiche (cfr. par. 4.2.2.1 e 4.2.2.2). Tali procedure hanno lo scopo di individuare i migliori compromessi tra accuratezza e copertura, e di consentire di ricavare descrizioni delle regioni dello spazio dei dati che non abbiano alcuna intersezione, evitando la ridondanza e generando sostanzialmente una partizione.

Sebbene queste procedure automatiche non forniscano un modello interpretativo appropriato, dato che non sono ideate per questo scopo, i pattern selezionati supportano letture alternative del fenomeno, attraverso il quale è possibile ottenere importanti informazioni atte all'identificazione di diverse categorie di utenti e non utenti. Di seguito, è

mostrata un'importante applicazione dei risultati ottenuti con le tecniche SLAD, attraverso l'uso di alberi di classificazione.

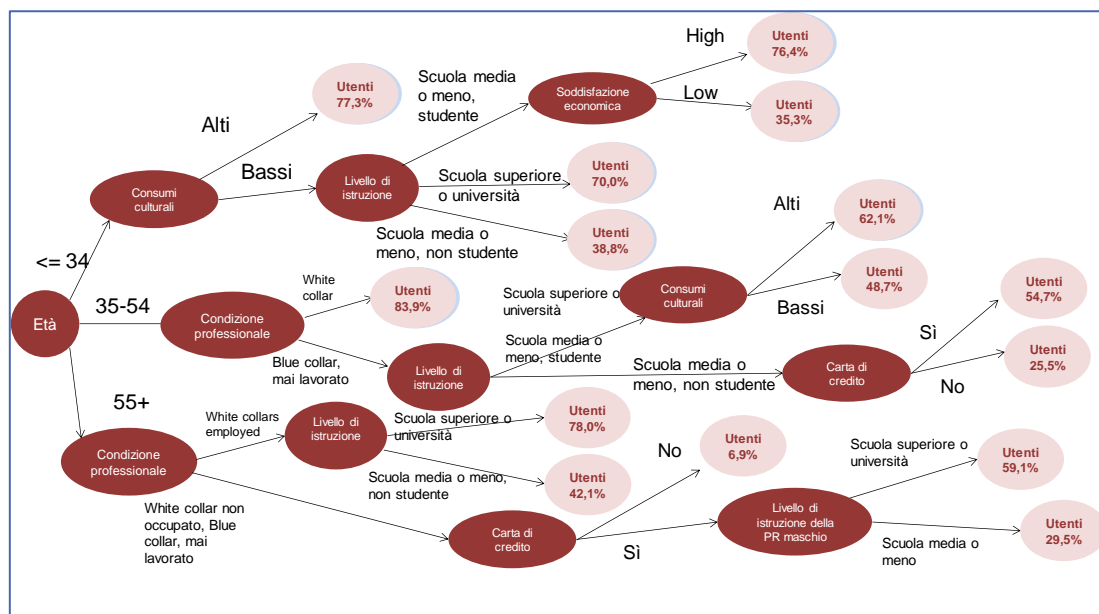
6.1.3 Partizione della popolazione italiana attraverso gli alberi di classificazione

A valle dei risultati della SLAD, al fine di costruire una partizione significativa di tutta la popolazione italiana rispetto all'uso di Internet, viene effettuata un'analisi con alberi di classificazione, attraverso l'uso dell'algoritmo QUEST (cfr. par. 4.3).

In particolare, l'analisi SLAD consente di selezionare 19 variabili che hanno effettivamente impatto sull'uso di Internet. Tali variabili, opportunamente dicotomizzate sulla base delle informazioni provenienti dai pattern, in termini di impatto sull'uso di Internet, vengono dunque suggerite all'algoritmo di costruzione di un albero di classificazione.

Il risultato dell'algoritmo QUEST è un albero di 5 livelli massimi e 28 nodi totali, di cui 15 foglie, mostrato in figura seguente. Si noti come, a partire dalle 19 variabili in input, l'algoritmo seleziona 7 variabili di segmentazione.

Figura 26 – Albero risultante dall'applicazione dell'algoritmo QUEST (anno 2012)



I nodi terminali dell'albero costituiscono per costruzione una partizione della popolazione italiana (di 6 anni o più, per i quali viene rilevata la variabile) rispetto all'uso di Internet. Tuttavia, alcuni di questi nodi terminali identificano porzioni molto ridotte della popolazione. Si decide dunque, in questi casi, di risalire l'albero e prendere come nodo finale della partizione il nodo padre, invece dei nodi figli.

Da questa operazione si ottiene una partizione di 8 differenti cluster della popolazione italiana, riportati in tabella seguente, identificati con le lettere da A ad H.

Tabella 18 - Partizione risultante dall'aggregazione dei nodi dell'albero (anno 2012)

| Nodo | Condizioni | % della popolazione | % di utenti nel nodo |
|------|--|---------------------|----------------------|
| A | Età <=34 | 23,1% | 77,3% |
| | Consumi culturali = Alti | | |
| B | Età <=34 | 3,4% | 40,1% |
| | Consumi culturali = Bassi | | |
| | Livello di istruzione = Scuola media o meno, in formazione | | |
| C | Età <=34 | 2,2% | 70,0% |
| | Consumi culturali = Bassi | | |
| | Livello di istruzione = Scuola superiore o università | | |
| D | Età <=34 | 2,4% | 38,8% |
| | Consumi culturali = Bassi | | |
| | Livello di istruzione = Scuola media o meno, NON in formazione | | |
| E | 35<= Età <=54 | 12,0% | 83,9% |
| | Stato occupazionale = White collar (occupato o meno) | | |
| F | 35<= Età <=54 | 20,0% | 39,0% |
| | Stato occupazionale = Blue collar (occupato o meno) o non ha mai lavorato | | |
| G | Età >=55 | 3,2% | 69,0% |
| | Stato occupazionale = White collar (occupato) | | |
| H | Età >=55 | 33,7% | 12,3% |
| | Stato occupazionale = White collar (non occupato), Blue collar (occupato o meno) o non ha mai lavorato | | |

Si può notare come il cluster H sia al contempo il più ampio (33,7% della popolazione) e quello con meno utenti di Internet (12,3% del cluster). Si tratta di individui maturi, con più di 55 anni, che non hanno mai lavorato, o sono stati *white collar* ma non lavorano più, oppure sono *blue collar* (occupati o meno). In definitiva, si tratta di anziani che non svolgono al momento mansioni di intelletto, e dunque sembra essere evidente la loro marginalità sociale e culturale, rispetto al mondo attuale, che li porta a utilizzare pochissimo la Rete. Il loro tasso di utilizzo di Internet è di gran lunga inferiore a quello di qualunque altro cluster della popolazione.

Infatti, il penultimo cluster per utilizzo di Internet è il D, che però presenta più del triplo degli utenti del nodo H, in percentuale sul totale degli individui del cluster, fermandosi comunque a un insufficiente 38,8%. Si tratta di individui giovani (con meno di 35 anni), tra i

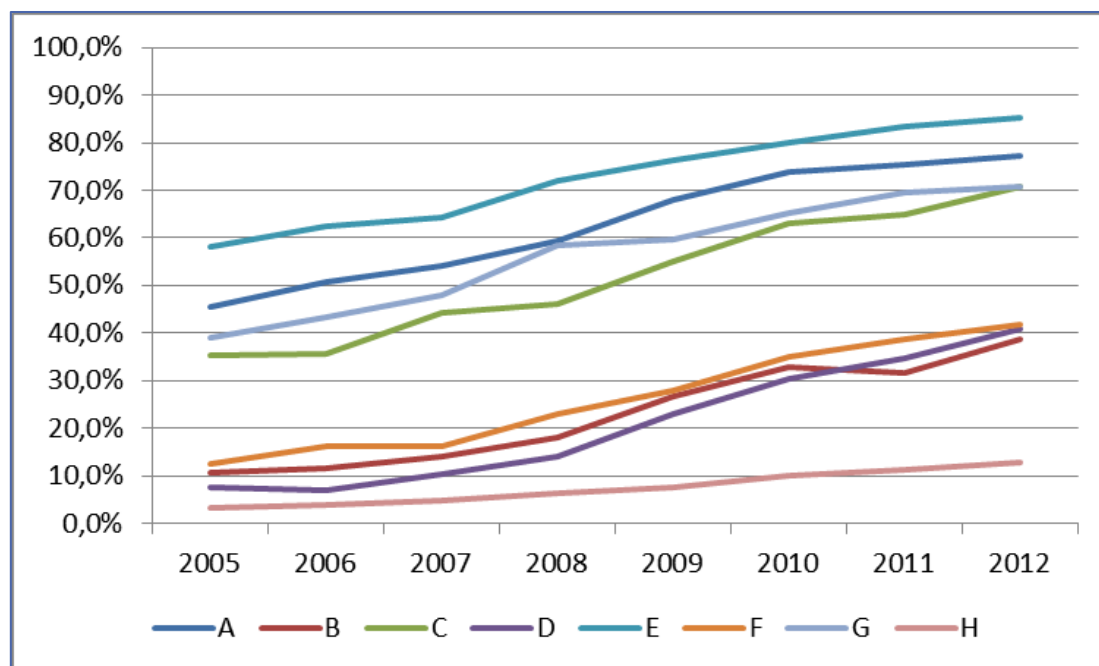
quali i bassi consumi culturali uniti a un basso titolo di studio, essendo peraltro usciti dai circuiti formativi, producono una scarsa attitudine all'uso delle ICT. Per quanto tale dato sia preoccupante, trattandosi appunto di un segmento di giovani, essi sono in numero ridotto tra la popolazione (2,4%).

Dall'altro lato, il cluster che mostra una maggiore quota di utenti Internet (83,9%) è quello composto da individui tra 35 e 54 anni, che hanno svolto o svolgono ancora mansioni di intelletto (white collar). Essi costituiscono il 12% della complessiva popolazione italiana (nodo E). In questo cluster è dunque particolarmente evidente l'effetto congiunto di un'età favorevole, ma non giovanissima, e dell'attività lavorativa, anche se non necessariamente in corso. Tale attività, questa volta con la condizione più stringente di essere ancora in corso, sembra favorire fortemente l'uso anche tra i più anziani (nodo G), laddove gli occupati white collar con 55 anni o più mostrano un tasso di utenti di quasi 57 punti superiore a quello dei loro coetanei del nodo H. Si tratta comunque di una porzione abbastanza ridotta della popolazione (3,2%), di circa 10 volte inferiore a quella che compone l'altro cluster di ultra cinquantacinquenni.

Infine, è interessante notare come anche gli individui giovani con consumi culturali elevati presentino un tasso di utenti di Internet elevato (77,3%). Si tratta del secondo cluster (nodo A) per dimensione, che contiene il 23% circa della popolazione italiana. Dunque, questi risultati mostrano come l'effetto dell'età, che resta un fattore fondamentale nell'uso (come mostrato in 3.3 e 6.1.1), sia di volta in volta mitigato o viceversa potenziato da altri fattori, quali principalmente i consumi culturali, il titolo di studio e lo status occupazionale.

Nella figura seguente è invece possibile osservare l'evoluzione del tasso di utenti per ciascun nodo, negli anni 2005-2012.

Figura 27 - Evoluzione del tasso di utenza di Internet in ogni nodo della partizione (anni 2005-2012)



Come si può notare, le diverse sottopopolazioni individuate dalla partizione hanno avuto un andamento nell'adozione di Internet molto differente. In particolare, il cluster di coloro che usano meno Internet, identificato dalle persone di 55 anni o più che non sono occupate, non lo sono mai state o lo sono ma in lavori manuali (nodo H), è anche quello che ha sperimentato l'incremento di gran lunga più lento negli anni.

Inoltre, vi è un'evidente separazione tra i cluster A, E, G e C da un lato, che contengono combinazioni favorevoli di età, titolo di studio, occupazione e consumi culturali, e gli altri cluster dall'altro, in cui vi sono combinazioni evidentemente sfavorevoli di tali variabili.

Infine, è di particolare rilievo il risultato che non vi è stata inversione nell'ordinamento dei cluster per quota di utenti, fatta eccezione per i cluster B e D, che mostrano comunque valori molto simili. Dunque, non c'è stato alcun recupero nella diffusione nei cluster più svantaggiati, nel periodo considerato.

A partire dalle caratteristiche strutturali dei diversi segmenti di popolazione individuati, è possibile dunque studiare politiche di sostegno alla domanda Internet differenziate per cluster e quindi più mirate a superare gli elementi ostativi che possono essere anche di natura molto diversa (culturale, di competenze, economica).

Si noti comunque che in alcuni segmenti della popolazione il tasso di adozione risulta essere basso a causa di problemi strutturali non facilmente risolvibili nel breve tempo, anche con azioni ad-hoc. Si pensi ad esempio ai settori più anziani della popolazione, in cui abbiamo visto risiedere buona parte dei non utenti e delle famiglie senza connessione a banda larga. Agire su questi segmenti è difficile, sia per ragioni culturali che, per l'appunto, di età (cfr. successivo capitolo 7).

6.2 Frequenza d'uso e ricchezza d'uso

Nel corso del precedente paragrafo 6.1 sono stata indagata le determinanti del secondo stadio del modello di digital divide proposto nel capitolo 5. Tale secondo stadio è rappresentato dall'adozione di Internet, che come abbiamo visto si estrinseca in due diversi fenomeni fortemente interrelati, la sottoscrizione di un servizio ad Internet da parte della famiglia e la decisione di uso individuale.

In questa seconda sezione del capitolo vengono dunque indagati i restanti due stadi del digital divide, la frequenza d'uso e le attività online, *proxy* di una certa ricchezza d'uso. Anche in questo caso, come nel caso del duplice aspetto di uso e connessione, si è però di fronte a un'influenza reciproca di questi due fenomeni.

L'ipotesi è: chi accede più spesso a Internet apprezza maggiormente tutte le sue potenzialità, e ne fa dunque un uso più ricco. D'altro canto, una persona interessata a utilizzare una pluralità di diversi servizi online, accederà anche ad Internet con maggiore frequenza.

Entrambi i fenomeni possono dunque essere visti come espressione di una variabile latente (nel senso di non misurata) che potrebbe essere definita come "interesse per Internet". L'interesse può infatti portare ad utilizzare una serie di servizi, conoscerne sempre di più e accedere ogni giorno a Internet.

L'ipotesi derivante dall'analisi di regressione effettuata nel in particolare, è, inoltre, che vi sia una relazione stretta fra le attività svolte online e quelle svolte offline.

Inoltre, i risultati dell'analisi di regressione logistica sull'uso e dell'analisi esplorativa SLAD suggeriscono una stretta relazione fra l'uso di Internet e le attività culturali svolte *offline*. Insieme ai consumi culturali, dunque, l'interesse per Internet potrebbe essere l'espressione di un'altra variabile latente, la "cultura" personale, o meglio il capitale culturale di una persona (Bourdieu, 1979, 1986). Essa è intesa come la ricchezza e l'ampiezza degli interessi e del know-how di un individuo, derivate dal percorso formativo, dalle preferenze personali, dalle esperienze lavorative e personali, dalle reti sociali.

Di seguito, la verifica di tali associazioni tra variabili d'uso di Internet e attività culturali è oggetto di un'analisi per classi latenti.

6.2.1 Classi latenti di utenza di Internet¹³

Questa analisi ha dunque il fine di indagare l'uso di Internet (in termini di frequenza e ricchezza), ma anche il non uso, in relazione a diverse caratteristiche degli individui, e al contempo considerare una serie di aspetti riguardanti le attività culturali svolte, quali il recarsi al teatro, al cinema, leggere libri e quotidiani, recarsi a visitare musei, mostre e monumenti ecc. Per provare a stimare questa variabile latente, come si diceva pocanzi, è possibile utilizzare il metodo dell'analisi per classi latenti, descritto nel paragrafo 4.4.

Questo metodo consente di costruire automaticamente una nuova variabile categoriale, suddivisa su di un certo numero predefinito di classi, che è considerata latente, nel senso di non misurata o misurabile. Tale costruzione avviene a partire da un certo numero di variabili manifeste, che dunque sono considerate *proxy* della variabile latente, e da esse causate o concausate, sulla base di un altro insieme di variabili di input o indipendenti, che si considera possano influenzare sia la variabile latente che quelle manifeste, così come avviene all'interno della regressione logistica per quel che riguarda la relazione tra variabili di ingresso e di uscita.

È necessario, in questa sede, svolgere analisi separate per utenti e non utenti, poiché per costruzione queste due categorie presentano differenti variabili manifeste. In particolare, mentre per gli utenti possono essere considerate la frequenza di uso, la connessione a casa e il tipo di attività svolte online, per i non utenti vengono considerate, insieme alle altre variabili *offline*, i motivi del non uso. Entrambi i tentativi di formulazione rispondono comunque all'esigenza di catturare un'unica variabile di interesse per Internet, a partire dai dati disponibili nel dataset di analisi, che è quello dell'indagine Istat AVQ del 2014. L'analisi è svolta sui soli individui tra 15 e 74 anni poiché, per i minori di 15 anni, le variabili relative all'occupazione non sono censite, e gli ultra settantaquattrenni perdono senso alcune variabili, poiché le loro caratteristiche risultano essere fortemente polarizzate sulle variabili di segmentazione.

Per gli utenti, quindi, come variabili manifeste sono stati considerati 56 indicatori dei comportamenti di chi dichiara di aver utilizzato Internet nei 3 mesi precedenti la somministrazione del questionario, tra cui i più rilevanti sono quelli relativi a:

¹³ Una versione ridotta e divulgativa di questa analisi è stata pubblicata da Claudio Leporelli (DIAG -Sapienza) e Cosimo Dolente, in diversi articoli, sulla rivista online AgendaDigitale.eu (<http://www.agendadigitale.eu>).

- la frequenza di utilizzo di Internet, in generale, a casa, al lavoro e in mobilità;
- l'aver svolto o meno ciascuna di 10 attività culturali, in senso lato, compresa la lettura di libri e quotidiani, alcune di maggiore interesse per la popolazione giovanile (ad esempio spettacoli di musica pop e di sport) altri per la popolazione matura (teatro, musica classica, etc.);
- il possesso di una carta di credito;
- il giudizio sul proprio benessere economico personale percepito (in quattro livelli);
- l'aver svolto o meno ciascuna di 36 attività su Internet, classificabili come attività di accesso all'Informazione (*Inf* nelle successive tabelle), di Comunicazione (*Com*), di Fruizione di Contenuti (*Fr Cont*), di partecipazione a social network (*Soc*), di fruizione di Servizi (*Serv*), legate al Software e a servizi Cloud (*Sw & Cl*), di interazione con la Pubblica Amministrazione (*PA*) e, infine di Gaming (*Gam*) (Van Deursen e Van Dijk, 2013);
- alcuni indicatori di sintesi sulle attività Internet svolte (numero di attività video, abilità percepita in base all'utilizzo di attività di nicchia, etc.);
- un indice di capacità informatiche, sintetizzato a partire dalle 10 presenti nel dataset attraverso un modello di Rasch¹⁴, utilizzato per capirne l'importanza relativa.

Per i non utenti, invece, le variabili manifeste incluse nel modello sono in numero inferiore (24), poiché non sono presenti ovviamente le 36 diverse attività svolte online e le altre variabili relative all'uso dell'ICT. Sono invece incluse (cfr. Tabella A.6 in allegato) le variabili relative a:

- i 13 motivi del non uso;
- le stesse variabili relative all'aver svolto o meno ciascuna delle attività culturali, utilizzate anche per gli utenti;
- il possesso di una carta di credito;
- la soddisfazione per la propria situazione economica personale.

Le variabili di input considerate ricalcano invece, in gran parte, quelle utilizzate all'interno della regressione logistica (cfr. 6.1.1), per entrambe le categorie della popolazione:

¹⁴ Si tratta di una tipologia di analisi che mira a stabilire la difficoltà relativa delle risposte a un questionario a partire da un insieme di risposte (Rasch, 1961; Fischer e Molenaar, 2012).

- Classe di età;
- Ripartizione geografica;
- Condizione professionale;
- Posizione professionale;
- Dimensione del comune di residenza;
- Ripartizione geografica;
- Livello di istruzione;
- Attività di istruzione in corso;
- Giudizio sulle risorse economiche della famiglia.

Per motivi computazionali, è particolarmente difficoltoso svolgere questo tipo di analisi consentendo un numero elevato di classi latenti. Tuttavia, l'analisi SLAD (cfr. paragrafo 6.1.2) mostra come si possa ipotizzare che gli utenti siano molto disomogenei, laddove i non utenti sono più facilmente individuabili.

Per questo motivo, a seguito di diverse iterazioni nell'applicazione dell'algoritmo POLca con i software R e SPSS (Linzer e Lewis, 2011), si è deciso di fissare a 6 il numero di classi latenti per gli utenti e a 3 quelle dei non utenti. Per quanto tale scelta rappresenti solo una delle possibili soluzioni, essa produce risultati facilmente interpretabili, coerentemente con la valutazione più ampia del fenomeno finora esposto.

In Tabella 19 sono riportate le consistenze delle diverse classi, in termini di quota di utenti nella classe, e le frequenze di uso di Internet di ciascuna di esse. In allegato sono invece riportati i valori relativi alle distribuzioni delle variabili manifeste sulle classi latenti e alla stima dei parametri delle variabili indipendenti.

Tabella 19 - Classi latenti di utenti per frequenza d'uso di Internet (anno 2014)

| | | Classi latenti | | | | | |
|--|-------------------------------------|----------------|-------|-------|-------|-------|-------|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| Individui nella classe (milioni) | | 4,753 | 4,723 | 3,774 | 4,487 | 5,725 | 4,544 |
| Quota di utenti totali nella classe | | 17,0% | 16,9% | 13,5% | 16,0% | 20,4% | 16,2% |
| | <i>Tutti i giorni</i> | 97,9% | 85,9% | 11,2% | 1,6% | 97,3% | 97,1% |
| | <i>Almeno una volta a settimana</i> | 2,0% | 12,9% | 87,9% | 75,8% | 2,6% | 2,8% |

| | | Classi latenti | | | | | |
|-------------------------------------|--------------------------------------|----------------|------|------|-------|------|------|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| Frequenza di uso di Internet | <i>Meno di una volta a settimana</i> | 0,0% | 1,2% | 0,9% | 22,6% | 0,0% | 0,1% |

La quota di utenti appartenenti a ciascuna delle sei classi non differisce di molto (dal 13,5% al 20,4%). Per quanto riguarda invece la frequenza di Internet, si può notare come le classi 1, 5 e 6 siano composte nella quasi totalità dei casi da individui che usano Internet tutti i giorni, così come la classe 2, anche se in misura minore, dato la quota comunque rilevante di utenti che lo utilizzano almeno una volta a settimana, ma non tutti i giorni. Le classi 3 e 4, invece, sono composte in gran parte da utenti di quest'ultimo tipo. La quota di utenti che utilizzano addirittura meno di una volta settimana è rilevante, e anzi consistente (22,6%), solo all'interno della classe 4.

Passando alle attività online, la Tabella 20 riporta la probabilità stimata che un utente assegnato a ciascuna classe abbia svolto, nei tre mesi precedenti la somministrazione del questionario, quell'attività. Le righe sono ordinate in modo decrescente rispetto al numero delle classi in cui almeno il 50% degli utenti svolge l'attività e rispetto alla probabilità aggregata di svolgimento dell'attività. In alto sono dunque riportate le attività più ampiamente svolte, mentre in basso si trovano quelle di nicchia.

Tabella 20 - Probabilità stimata di utilizzo dell'applicazione per classe di appartenenza (anno 2014)

| Gruppo | Applicazioni | Classi latenti | | | | | | Totale |
|---------|---------------------------------|----------------|-------|-------|-------|-------|-------|--------------|
| | | 1 | 2 | 3 | 4 | 5 | 6 | |
| Com | E Mail | 99,5% | 78,8% | 86,7% | 58,5% | 99,1% | 86,8% | 85,3% |
| Inf | Uso di siti wiki | 94,7% | 36,3% | 62,3% | 29,9% | 79,3% | 71,2% | 62,8% |
| Inf | Leggere informazioni online | 91,0% | 44,5% | 55,7% | 33,1% | 79,8% | 53,6% | 60,4% |
| Com | Messaggistica istantanea | 90,1% | 32,1% | 50,4% | 16,1% | 52,0% | 78,8% | 53,1% |
| Soc | Social network | 90,9% | 45,1% | 58,6% | 24,1% | 48,7% | 88,7% | 58,8% |
| Inf | Informazioni su acquisti | 89,1% | 35,3% | 55,9% | 31,8% | 80,0% | 46,7% | 57,3% |
| Soc | Blog | 91,3% | 30,9% | 51,9% | 14,4% | 46,7% | 81,1% | 52,2% |
| Inf | Informazioni Sanitarie | 72,1% | 28,8% | 45,4% | 28,1% | 66,8% | 33,0% | 46,4% |
| Fr Cont | Download film, immagini, musica | 85,8% | 11,2% | 44,9% | 7,8% | 32,4% | 77,3% | 42,4% |
| Serv | Servizi per viaggi e soggiorni | 74,7% | 20,8% | 37,9% | 16,3% | 62,2% | 31,9% | 41,5% |
| Serv | Servizi bancari | 73,0% | 26,8% | 31,6% | 14,5% | 71,7% | 19,4% | 41,0% |
| Inf | Inform Istruzione | 74,7% | 14,5% | 35,9% | 10,8% | 50,8% | 33,5% | 37,1% |
| Fr Cont | Caricare Contenuti Multimedia | 73,0% | 15,0% | 31,7% | 6,5% | 30,7% | 53,4% | 34,8% |

| Gruppo | Applicazioni | Classi latenti | | | | | | Totale |
|---------|----------------------------|----------------|-------|-------|-------|-------|-------|--------|
| | | 1 | 2 | 3 | 4 | 5 | 6 | |
| Fr Cont | Streaming Video | 83,6% | 4,6% | 31,1% | 4,1% | 20,3% | 64,2% | 33,9% |
| Gam | Giochi | 52,0% | 18,0% | 30,6% | 12,6% | 20,9% | 57,9% | 31,4% |
| Fr Cont | Streaming Film | 66,3% | 2,3% | 21,0% | 1,9% | 8,4% | 54,2% | 24,9% |
| Com | Videochiamate | 71,7% | 21,8% | 33,7% | 14,0% | 39,9% | 47,5% | 38,1% |
| Sw & Cl | Download Software | 74,6% | 5,0% | 16,3% | 1,9% | 31,4% | 29,9% | 26,8% |
| Sw & Cl | Salvare Su Cloud | 64,7% | 10,3% | 21,7% | 5,3% | 29,3% | 27,8% | 26,6% |
| Fr Cont | Ascolto Radio | 57,8% | 9,8% | 23,8% | 5,5% | 23,9% | 39,9% | 26,6% |
| Serv | Comprare Beni | 57,3% | 10,1% | 15,2% | 4,6% | 38,3% | 18,5% | 24,7% |
| Fr Cont | Guardare Tv | 62,3% | 4,5% | 20,7% | 2,8% | 16,0% | 40,4% | 24,0% |
| Soc | Condividere Opin Soc. Pol. | 59,9% | 7,9% | 18,9% | 3,0% | 19,2% | 28,2% | 22,7% |
| Sw & Cl | Condividere Su Cloud | 57,3% | 6,5% | 16,4% | 3,5% | 21,8% | 21,7% | 21,2% |
| PA | Informazioni Da SITI PA | 48,0% | 6,9% | 21,2% | 5,8% | 43,5% | 7,8% | 23,0% |
| Serv | Cercare Lavoro | 40,1% | 12,1% | 21,5% | 9,8% | 11,8% | 27,6% | 20,0% |
| PA | Ottenere Moduli Pa | 42,3% | 3,9% | 15,5% | 2,8% | 38,0% | 4,0% | 18,6% |
| Fr Cont | Lettura Download Ebook | 47,5% | 3,0% | 10,1% | 1,7% | 20,4% | 12,4% | 16,1% |
| Serv | Vendere | 38,2% | 6,6% | 11,8% | 3,5% | 18,0% | 12,6% | 15,3% |
| Gam | Giochi In Rete | 33,0% | 6,4% | 12,5% | 2,1% | 8,0% | 30,4% | 15,1% |
| Serv | Linkedin E Simili | 41,8% | 1,7% | 6,3% | 0,7% | 16,0% | 8,5% | 12,8% |
| PA | Spedire Moduli Pa | 29,6% | 2,4% | 9,2% | 1,7% | 25,0% | 1,9% | 12,2% |
| Soc | Votaz Online Soc E Pol | 36,7% | 2,2% | 6,9% | 0,9% | 14,5% | 7,2% | 11,6% |
| Serv | Appuntamento Medico | 21,6% | 2,9% | 7,9% | 2,7% | 16,8% | 2,4% | 9,4% |
| Fr Cont | Corso Online | 21,6% | 1,8% | 4,6% | 1,1% | 11,8% | 3,4% | 7,6% |
| Sw & Cl | Creare Siti Web | 19,9% | 0,9% | 2,4% | 0,4% | 3,7% | 5,8% | 5,5% |

In verde i valori $\geq 50\%$

I comportamenti osservati differiscono fortemente, tra le diverse classi. Gli utenti con profili d'uso più limitati sono quelli appartenenti alla classe 4 ed alla classe 2: è solo una (l'utilizzo di e-mail) l'attività svolta da almeno il 50% degli utenti assegnati a ciascuna di queste classi. Il numero di attività svolte da almeno il 50% degli utenti (caselle evidenziate in verde) cresce fino ad un massimo di 23 per la classe 1.

In Tabella 21, per ciascuno dei gruppi di attività definiti nella prima colonna della tabella precedente, sono valutate, per ciascuna classe e per il complesso degli utenti, il numero di attività di ciascun gruppo svolte da almeno il 50% degli utenti assegnati ad una classe (o il 50% del totale degli utenti) e, similmente, il numero atteso di attività. La tabella evidenzia che solo alcune attività di comunicazione, accesso ad informazioni e partecipazione a *social network* sono svolte da almeno la metà del complesso degli utenti, mentre le attività di fruizione di contenuti (prevalentemente multimediali) sono significativamente presenti diffusamente solo

nella prima e sesta classe, quelle di utilizzo di servizi nella prima e nella quinta e quelle connesse al software e al *cloud* solo nella prima.

Tabella 21 - Numero di attività svolte per gruppo di attività, per ciascuna delle classi latenti individuate (anno 2014)

| | Classi di utenti | | | | | | Totale utenti |
|--|------------------|-----|-----|-----|------|-----|---------------|
| | 1 | 2 | 3 | 4 | 5 | 6 | |
| N° di attività svolte da almeno il 50% degli utenti | | | | | | | |
| <i>Informazione</i> | 5 | - | 3 | - | 5 | 2 | 3 |
| <i>Fruizione Contenuti</i> | 6 | - | - | - | - | 4 | - |
| <i>Comunicazione</i> | 3 | 1 | 2 | 1 | 2 | 2 | 2 |
| <i>Social network</i> | 3 | - | 2 | - | - | 2 | 2 |
| <i>Servizi</i> | 3 | - | - | - | 2 | - | - |
| <i>Software & Cloud</i> | 3 | - | - | - | - | - | - |
| <i>Gaming</i> | - | - | - | - | - | - | - |
| <i>PA</i> | - | - | - | - | - | - | - |
| Totale | 23 | 1 | 7 | 1 | 9 | 10 | 7 |
| N° atteso di attività | | | | | | | |
| <i>Informazione</i> | 4,2 | 1,6 | 2,6 | 1,3 | 3,6 | 2,4 | 2,6 |
| <i>Fruizione Contenuti</i> | 0,5 | 0,0 | 0,1 | 0,0 | 0,2 | 0,1 | 0,2 |
| <i>Comunicazione</i> | 2,6 | 1,3 | 1,7 | 0,9 | 1,9 | 2,1 | 1,8 |
| <i>Social network</i> | 2,8 | 0,9 | 1,4 | 0,4 | 1,3 | 2,1 | 1,5 |
| <i>Servizi</i> | 3,5 | 0,8 | 1,3 | 0,5 | 2,3 | 1,2 | 1,6 |
| <i>Software & Cloud</i> | 0,2 | 0,0 | 0,0 | 0,0 | 0,0 | 0,1 | 0,1 |
| <i>Gaming</i> | 0,8 | 0,2 | 0,4 | 0,1 | 0,3 | 0,9 | 0,5 |
| <i>PA</i> | 1,2 | 0,1 | 0,5 | 0,1 | 1,1 | 0,1 | 0,5 |
| Totale | 15,8 | 5,0 | 8,0 | 3,4 | 10,7 | 9,0 | 8,7 |

La prima classe emerge quindi come quella cui vengono assegnati gli utenti più sofisticati, sia per attività legate al tempo libero, che al lavoro o alla fruizione di servizi. La sesta classe condivide con la prima le attività ricreative (in particolare di *gaming*) ma non quelle di utilizzo di servizi. La quinta condivide con la prima l'utilizzo di servizi utili a persone mature, ma non le attività ricreative multimediali. Considerando il numero medio di attività attese per ciascun utente, è in particolare la classe 4 quella che mostra un uso più povero, seguita dalla 2, mentre la classe 3, con 8 attività, è vicina al valore di 9 della classe 6.

Al di là delle attività svolte online, al fine di inquadrare le diverse tipologie di utenza, è utile verificare quali siano le caratteristiche socioeconomiche e demografiche degli utenti assegnati alle diverse classi (Tabella 22).

Tabella 22 - Caratteristiche socio-demografiche delle classi latenti di utenti (anno 2014)

| | Classe latente | | | | | |
|---|----------------|---------|---------|---------|---------|---------|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Età media in anni | 34,0 | 43,8 | 39,2 | 46,2 | 46,8 | 27,0 |
| Classe di età mediana | 30 - 34 | 40 - 44 | 35 - 39 | 45 - 49 | 45 - 49 | 20 - 24 |
| % Persona di riferimento | 37,4% | 49,5% | 38,7% | 48,3% | 58,0% | 18,3% |
| % Coniuge o convivente | 13,8% | 28,1% | 26,8% | 34,8% | 29,4% | 10,1% |
| % Figlio | 46,0% | 20,7% | 30,8% | 14,2% | 11,1% | 66,5% |
| % Maschi | 59% | 53% | 51% | 48% | 52% | 54% |
| % Laureati | 34,0% | 11,4% | 15,2% | 8,5% | 37,6% | 7,5% |
| % Laureati, diplomati o studenti delle scuole superiori | 93,0% | 65,2% | 74,7% | 56,1% | 90,2% | 81,8% |
| % Occupati | 60,1% | 62,7% | 55,3% | 56,7% | 80,1% | 34,1% |
| % In cerca di nuova occupazione | 10,7% | 10,9% | 12,0% | 10,6% | 4,8% | 14,7% |
| % In cerca di prima occupazione | 3,9% | 2,4% | 5,0% | 3,0% | 0,2% | 7,6% |
| % Studenti | 21,9% | 6,4% | 12,5% | 3,1% | 2,4% | 38,8% |
| % Pensionati, casalinghe e inabili | 3,4% | 17,6% | 15,2% | 26,6% | 12,5% | 4,7% |
| % White Collars | 50,5% | 41,4% | 36,9% | 32,7% | 75,3% | 14,9% |

Risulta confermata l'intuizione che scaturiva dall'analisi delle attività svolte: le classi che vedono la presenza preponderante di giovani sono la 1 (più maturi e professionalmente ben inseriti) e la 6 (con una età media inferiore, preponderanza di studenti e inserimento professionale meno brillante per chi ha completato gli studi); la classe 5 contiene prevalentemente *white collar* maturi, spesso persone di riferimento della famiglia, occupati, con alta incidenza di laureati. Le classi 2, 3 e 4 contengono invece gli utenti con uso limitato, e sono meno distinguibili tra loro di quanto non lo fossero in base alle attività svolte, presentando mediamente situazioni sociali più deboli: maggiore incidenza di pensionati e casalinghe, minore di laureati e *white collar*.

Venendo alle variabili manifeste non ancora analizzate, è significativo che la classe 1 sia anche quella con il più alto numero atteso di consumi culturali dichiarati, diversi dalla lettura (2,7, su un massimo teorico di 8), seguita ancora una volta dalla 5 e dalla 6. All'estremo opposto, con 1,1, le classi con uso limitato, 2 e 4. Quasi il 60% degli utenti assegnati alle classi 2 e 4, e oltre il 50% di quelli della classe 6, inoltre, non legge libri. Quasi due terzi degli utenti assegnati alla classe 5 e oltre metà di quelli della classe 1 utilizza una carta di credito; questo utilizzo è invece molto basso nella classe 6 (si tratta di giovani che vivono per due terzi in famiglia) e comunque inferiore ad un terzo nelle classi 2, 3 e 4.

Il risultato relativo alle attività culturali pare confermare la fondatezza dell'ipotesi per cui le attività online e offline siano fortemente connesse, e mostra che la suddivisione in classi latenti cattura effettivamente questa correlazione.

Particolarmente interessante è notare che, sulla base dei parametri calibrati dalla regressione, le variabili di input non sembrano discriminare in maniera sostanzialmente diversa tra le classi 2, 3 e 4. Questo pare essere un risultato fondamentale, poiché indica che i comportamenti nell'uso di Internet non sono stimabili in maniera affidabile sulla base delle sole variabili di segmentazione. Ciò implica, in altri termini, che persone della stessa classe di età, che condividono lo stesso livello di istruzione, condizione professionale e posizione professionale, mostrano spesso differenti profili d'uso della Rete.

Dopo la verifica della stretta relazione tra attività online e offline, si tratta di un altro passo verso la conferma dell'ipotesi per cui ciò che spinge i differenti usi della Rete sia quella cultura personale di cui si parlava pocanzi, non catturabile attraverso l'uso di variabili oggettive.

Per quanto riguarda i non utenti, le 3 classi stimate, in aggiunta alle 6 appena esposte, divengono le classi 7, 8 e 9 della classificazione complessiva degli individui. La classificazione prende in esame non solo le motivazioni del non uso, ma anche, come si diceva, i consumi culturali, l'utilizzo di servizi finanziari e la soddisfazione per la propria situazione economica. Questo permette di affiancare alle motivazioni del mancato utilizzo una caratterizzazione socio-economica degli individui.

La classe 7 (2,9 milioni di persona, il 17,5% del totale dei non utenti considerati) è composta da individui di età media elevata (56 anni) e in maggioranza inattivi: solo il 37% è occupato o studente. Tuttavia, tra coloro che hanno lavorato o lavorano ancora, circa la metà è *white collar*, e il 52% possiede almeno un diploma di scuola superiore. In questa classe, il livello di soddisfazione per la propria situazione economica è più elevato rispetto alle altre classi di non utenti, i consumi culturali sono più alti e l'utilizzo di carta di credito è maggiore. Nonostante questa collocazione sociale favorevole, il 60% circa di questi soggetti si dichiara non interessata all'uso di Internet, e il 29% dichiara di non saperlo usare. È significativo notare che in questa classe il 48% degli individui vive in una famiglia che dispone di un accesso ad Internet a larga banda fissa.

La classe 8, di gran lunga la più numerosa riguardo i non utenti (10,5 milioni di persone, circa il 64,6% del totale), comprende soggetti anch'essi caratterizzati da età elevata (età media 58

anni), ma anche da significativa marginalità sociale: solo il 23% degli individui è occupato o studente, mentre il 63% è ritirato dal lavoro o casalinga; l'82% di coloro che hanno lavorato è o è stato blue collar; infine, solo il 17% degli individui della classe possiede almeno un diploma di scuola superiore. In questo caso, la quota di individui che vive in famiglie con connessione ad Internet a larga banda fissa scende al 32% circa.

Infine, la classe 9 è composta di 2,9 milioni di persone (circa il 19% dei non utenti considerati), e si differenzia dalle due precedenti per l'età media inferiore (41 anni, un valore inferiore a quello di tutta la popolazione italiana tra 15 e 74 anni) e per aver dichiarato nel 60% dei casi di non usare Internet per motivi legati al costo della connessione o degli strumenti necessari. Cresce la quota di occupati e studenti (50% circa della classe), mentre la quota di blue collar tra coloro che hanno lavorato è la più elevata, 86%. Il livello di istruzione è intermedio rispetto alle due classi precedenti, con il 31% circa di diplomati o laureati. In linea con le motivazioni espresse per il mancato uso, il livello di soddisfazione per la propria situazione economica è il più basso tra le tre classi di non utenti. Anche i consumi culturali ne risentono.

Si noti che questo esercizio conduce a risultati in parte diversi, anche se non contrastanti, rispetto all'analisi delle motivazioni del non uso, basate direttamente sulle risposte ai questionari (cfr. paragrafo 3.5). La classificazione prende in esame non solo le motivazioni del non uso, ma anche, come si diceva, i consumi culturali, l'utilizzo di servizi finanziari e la soddisfazione per la propria situazione economica. Questo permette di affiancare alle motivazioni del mancato utilizzo una caratterizzazione socio-economica degli individui.

7 Considerazioni di policy

La formulazione di indicazioni di policy atte a favorire la diffusione di Internet all'interno del nostro paese, al di là del "naturale" processo di diffusione tra la popolazione, proprio di ciascuna tecnologia (Bass, 1969), è stato tra gli obiettivi del presente progetto di ricerca. Le analisi effettuate nel corso del capitolo 6 hanno dunque come ricaduta pratica l'individuazione dei segmenti di utenza e, soprattutto, di non utenza. A partire da questi, è poi possibile immaginare politiche di sostegno alla domanda di Internet che siano efficaci, perché rispondenti alle diverse esigenze e motivazioni di questi differenti cluster.

7.1 Classificazione delle famiglie sulla base dell'adozione della tecnologia¹⁵

Una prima classificazione, utile ad indentificare i cluster oggetto di politiche di sostegno, si può ottenere a partire dai risultati dei modelli di regressione logistica (cfr. paragrafo 6.1.1) che, come si diceva in fase di descrizione della metodologia, consentono di stimare le probabilità dell'uscita per ciascun record del campione. In particolare, nel seguito verranno condotte le stime riguardo l'adozione addizionale di connessione a banda larga, fissa o mobile, di qualsiasi tipo da parte delle famiglie, all'interno di specifici scenari di policy.

In Tabella 23 sono riportate le probabilità di adozione di banda larga, al 2012, per ciascun cluster individuato dall'incrocio dei decili della probabilità di avere una situazione economica ottima o adeguata (denominata nelle tabelle "indice economico", per brevità) e dei decili della somma delle probabilità di tutti i componenti della famiglia (cioè componenti di 6 anni o più) di essere utenti di Internet, sulla base della prima regressione logistica.

Si noti come quest'ultima variabile rappresenti sia un interesse medio della famiglia, che un indice di massa critica, in quanto misura per l'appunto, in base a caratteristiche demografiche, culturali e lavorative, il numero atteso potenziale di utenti frequenti di Internet tra i componenti di 6 anni e più della famiglia.

¹⁵ La prima parte di questo paragrafo, relativa ai risultati delle regressioni logistiche, proviene dal lavoro svolto da Claudio Leporelli (DIAG – Sapienza) e Cosimo Dolente (DIMA – Spaienza, FUB), riportato in (Dolente e Leporelli, 2014).

Tabella 23 - Probabilità di adozione di una qualsiasi connessione a banda larga nel caso base, in percentuale, per cluster di utenza potenziale e situazione economica (anno 2012)

| Indice economico | Somma delle probabilità di utenza dei componenti della famiglia | | | | | | | | | | Totale |
|------------------|---|-------|-------|-------|-------|-------|-------|-------|-------|------|--------|
| | ≤0,5 | 0,5-1 | 1-1,5 | 1,5-2 | 2-2,5 | 2,5-3 | 3-3,5 | 3,5-4 | 4-4,5 | >4,5 | |
| 0-0,1 | 4,7 | 24,8 | 34,3 | 47,5 | 60,9 | 69,9 | 78,3 | 82,9 | 85,2 | 96,0 | 31,0 |
| 0,1-0,2 | 5,5 | 36,0 | 48,8 | 64,8 | 72,6 | 80,3 | 85,8 | 87,4 | 91,3 | 92,5 | 38,3 |
| 0,2-0,3 | 5,8 | 41,7 | 55,0 | 71,3 | 79,8 | 85,0 | 89,0 | 91,5 | 91,6 | - | 42,2 |
| 0,3-0,4 | 6,9 | 44,9 | 60,6 | 74,4 | 81,6 | 86,6 | 91,4 | 92,2 | 94,5 | 96,6 | 47,9 |
| 0,4-0,5 | 7,4 | 47,8 | 60,9 | 75,9 | 82,6 | 87,6 | 90,4 | 92,7 | 93,5 | 92,7 | 50,3 |
| 0,5-0,6 | 8,4 | 48,8 | 61,7 | 77,5 | 83,7 | 88,1 | 89,9 | 91,5 | 94,4 | 96,0 | 50,8 |
| 0,6-0,7 | 7,9 | 49,0 | 65,8 | 77,7 | 83,3 | 88,3 | 91,0 | 91,8 | 93,8 | 94,4 | 50,8 |
| 0,7-0,8 | 10,7 | 54,1 | 66,4 | 80,4 | 85,4 | 89,6 | 91,9 | 93,6 | 94,6 | 95,5 | 55,3 |
| 0,8-0,9 | 12,7 | 58,2 | 68,5 | 84,3 | 88,8 | 91,7 | 92,9 | 94,8 | 95,0 | 96,0 | 66,7 |
| 0,9-1 | 19,8 | 59,4 | 74,0 | 87,9 | 90,1 | 94,1 | 94,7 | 95,4 | 95,7 | 96,7 | 81,8 |
| Totale | 7,7 | 47,1 | 58,4 | 76,3 | 81,1 | 86,9 | 90,0 | 92,8 | 92,9 | 95,3 | 50,7 |

La tabella evidenzia l'effetto sinergico dei due indicatori sulle scelte di adozione. In particolare, le probabilità di adozione sono minime quando l'utenza potenziale è bassa e massime quando essa risulta essere molto alta. Tuttavia, quando tale valore è alquanto basso, ma la situazione economica è comunque abbastanza buona, si ottengono probabilità di adozione rilevanti. Ad esempio, in famiglie nel decile economicamente più sfavorito, la probabilità di adozione supera il 50% se il numero atteso di utenti potenziali è superiore a 2, mentre nel decile economicamente più favorito, perché ciò avvenga è sufficiente che in famiglia ci sia almeno un utente atteso.

Se guardiamo alla numerosità di adottanti in ciascuna cella, invece, si ottiene una mappa in cui i valori più alti si concentrano nelle aree che presentano maggiore numerosità nella popolazione e livelli di adozione intorno alla media.

Tabella 24 - Numero di famiglie (in migliaia) adottanti una qualsiasi connessione a banda larga nel caso base, per cluster di utenza potenziale e situazione economica (anno 2012)

| Indice economico | Somma delle probabilità di utenza dei componenti della famiglia | | | | | | | | | | Totale |
|------------------|---|-------|-------|-------|-------|-------|-------|-------|-------|------|--------|
| | ≤0,5 | 0,5-1 | 1-1,5 | 1,5-2 | 2-2,5 | 2,5-3 | 3-3,5 | 3,5-4 | 4-4,5 | >4,5 | |
| 0-0,1 | 25 | 100 | 99 | 96 | 102 | 70 | 43 | 20 | 8 | 5 | 568 |
| 0,1-0,2 | 49 | 188 | 134 | 231 | 171 | 125 | 68 | 21 | 25 | 9 | 1.022 |
| 0,2-0,3 | 44 | 179 | 127 | 188 | 160 | 123 | 75 | 15 | 12 | - | 924 |
| 0,3-0,4 | 43 | 195 | 156 | 233 | 156 | 113 | 75 | 35 | 9 | 7 | 1.023 |
| 0,4-0,5 | 41 | 200 | 129 | 257 | 164 | 120 | 70 | 24 | 8 | 5 | 1.019 |
| 0,5-0,6 | 52 | 213 | 148 | 267 | 174 | 114 | 70 | 39 | 9 | 5 | 1.090 |

| Indice economico | Somma delle probabilità di utenza dei componenti della famiglia | | | | | | | | | | Totale |
|------------------|---|-------|-------|-------|-------|-------|-------|-------|-------|--------|--------|
| | $\leq 0,5$ | 0,5-1 | 1-1,5 | 1,5-2 | 2-2,5 | 2,5-3 | 3-3,5 | 3,5-4 | 4-4,5 | $>4,5$ | |
| 0,6-0,7 | 49 | 217 | 132 | 270 | 155 | 132 | 78 | 45 | 9 | 5 | 1.092 |
| 0,7-0,8 | 86 | 374 | 180 | 391 | 211 | 186 | 148 | 55 | 8 | 13 | 1.653 |
| 0,8-0,9 | 63 | 486 | 217 | 624 | 286 | 281 | 131 | 98 | 15 | 10 | 2.211 |
| 0,9-1 | 4 | 80 | 65 | 253 | 74 | 78 | 35 | 81 | 5 | 16 | 692 |
| Totale | 457 | 2.233 | 1.388 | 2.809 | 1.653 | 1.343 | 793 | 435 | 108 | 75 | 11.293 |

Le variabili su cui può essere più agevole agire per un decisore pubblico, tra quelle emerse come rilevanti a seguito delle analisi di regressione del paragrafo 6.1.1, sono quelle di carattere economico e infrastrutturale, che sono catturate da:

- la probabilità di avere una situazione economica ottima o adeguata (cioè l'indice economico);
- la dimensione del comune, poiché i problemi di copertura infrastrutturale più gravi riguardano ormai quasi esclusivamente i comuni più piccoli del paese.

Infatti, per quanto la presenza di un interesse dei componenti della famiglia verso Internet sia di fondamentale importanza, agire su di esso, attraverso policy mirate, è sicuramente un compito molto difficile. Questo perché si tratta di una variabile culturale che è frutto di diversi fattori che vanno dall'età media e minima in famiglia, ai titoli di studio dei componenti della famiglia, ad altre variabili socio-demografiche, non facilmente modificabili dall'esterno.

Invece, intervenire sulla situazione economica, nell'ottica dell'adozione di Internet, è relativamente più semplice. Si potrebbe fornire un sussidio in forma diretta o indiretta alle famiglie con determinati requisiti, in modo da aumentarne la disponibilità a pagare. L'effetto di politiche di questo tipo, all'interno del modello proposto, può essere simulato attraverso una modifica della probabilità che la situazione economica sia ottima o adeguata, per diverse categorie di famiglie. In particolare, sono stati individuati due scenari di intervento in questo senso.

Il primo incentivo ipotizzato si riferisce a tutti coloro che vivono situazioni di ristrettezze economiche, individuati nelle famiglie con $p(\text{situazione economica}) \leq 0,5$. Tale incentivo è stato simulato tramite un incremento della stessa di 0,1, per tutte le famiglie oggetto di interesse.

Tabella 25 - Probabilità incrementali di adozione di banda larga per lo scenario 1, in punti percentuali, per cluster di utenza potenziale e situazione economica (anno 2012)

| Indice economico | Somma delle probabilità di utenza dei componenti della famiglia | | | | | | | | | | |
|------------------|---|-------|-------|-------|-------|-------|-------|-------|-------|------|--------|
| | ≤0,5 | 0,5-1 | 1-1,5 | 1,5-2 | 2-2,5 | 2,5-3 | 3-3,5 | 3,5-4 | 4-4,5 | >4,5 | Totale |
| 0-0,1 | 6,6 | 21,4 | 22,8 | 21,3 | 17,6 | 14,5 | 10,3 | 8,4 | 6,6 | 1,3 | 15,6 |
| 0,1-0,2 | 7,3 | 23,1 | 21,8 | 16,6 | 12,9 | 9,1 | 6,2 | 5,6 | 3,4 | 2,6 | 13,8 |
| 0,2-0,3 | 7,5 | 22,6 | 20,3 | 14,0 | 9,7 | 6,8 | 4,6 | 3,4 | 3,1 | 0,0 | 12,6 |
| 0,3-0,4 | 8,9 | 21,9 | 18,4 | 12,5 | 8,9 | 6,1 | 3,6 | 3,1 | 2,1 | 1,2 | 12,9 |
| 0,4-0,5 | 9,4 | 21,4 | 18,2 | 12,0 | 8,2 | 5,5 | 3,8 | 2,8 | 2,3 | 2,4 | 12,6 |
| 0,5-0,6 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 0,6-0,7 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 0,7-0,8 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 0,8-0,9 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 0,9-1 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| Totale | 4,5 | 10,3 | 10,7 | 6,0 | 5,6 | 3,5 | 2,3 | 1,2 | 1,8 | 0,7 | 6,5 |

Dunque, l'effetto maggiore si ha per utenza potenziale medio-bassa, da 0,5 a 2, in tutte le situazioni economiche su cui interviene l'azione di policy simulata. L'adozione addizionale derivante da questo scenario è stimato essere del 6,5% delle famiglie italiane, quantificabili in ben 1.428.650 famiglie.

Nel secondo scenario, si ipotizza invece che venga fornito un incentivo economico a tutte le famiglie in ristrettezze economiche con un componente tra 14 e 24 anni, poiché in tale fascia d'età si trovano le persone più inclini all'utilizzo di Internet, e, probabilmente, all'alfabetizzazione digitale dei propri famigliari. Queste famiglie sono individuate come quelle che presentano una $p(\text{situazione economica}) < 0,5$ e hanno, per l'appunto, almeno un componente tra 14 e 24 anni. L'incentivo è simulato modificandone la $p(\text{situazione economica})$, ponendola pari a 0,5.

Tabella 26 - Probabilità incrementali di adozione di banda larga per lo scenario 2, in punti percentuali, per cluster di utenza potenziale e situazione economica (anno 2012)

| Indice economico | Somma delle probabilità di utenza dei componenti della famiglia | | | | | | | | | | |
|------------------|---|-------|-------|-------|-------|-------|-------|-------|-------|------|--------|
| | ≤0,5 | 0,5-1 | 1-1,5 | 1,5-2 | 2-2,5 | 2,5-3 | 3-3,5 | 3,5-4 | 4-4,5 | >4,5 | Totale |
| 0-0,1 | 0,0 | 3,5 | 12,5 | 16,0 | 27,3 | 23,8 | 18,7 | 10,5 | 13,1 | 3,2 | 9,0 |
| 0,1-0,2 | 0,0 | 2,4 | 10,2 | 9,4 | 14,2 | 13,6 | 10,3 | 10,5 | 7,0 | 4,4 | 5,3 |
| 0,2-0,3 | 0,0 | 0,8 | 4,4 | 5,2 | 6,9 | 7,0 | 7,0 | 5,8 | 5,5 | 0,0 | 2,6 |
| 0,3-0,4 | 0,0 | 1,1 | 2,5 | 3,4 | 5,3 | 4,8 | 3,9 | 3,9 | 2,7 | 1,7 | 2,0 |
| 0,4-0,5 | 0,0 | 0,2 | 0,8 | 1,5 | 1,8 | 2,2 | 1,8 | 1,6 | 1,4 | 1,5 | 0,8 |
| 0,5-0,6 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 0,6-0,7 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |

| Indice economico | Somma delle probabilità di utenza dei componenti della famiglia | | | | | | | | | | |
|------------------|---|-------|-------|-------|-------|-------|-------|-------|-------|------|--------|
| | ≤0,5 | 0,5-1 | 1-1,5 | 1,5-2 | 2-2,5 | 2,5-3 | 3-3,5 | 3,5-4 | 4-4,5 | >4,5 | Totale |
| 0,7-0,8 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 0,8-0,9 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| 0,9-1 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| Totale | 0,0 | 0,7 | 3,4 | 2,6 | 5,3 | 4,1 | 3,4 | 1,6 | 3,2 | 1,0 | 1,9 |

Si può notare come questa volta l'effetto maggiore si ottiene per utenza potenziale medio-alta, da 1,5 a 3,5, e principalmente per situazioni economiche molto gravi (da 0 a 0,3 di probabilità di avere una situazione economica ottima o adeguata). L'adozione addizionale complessiva calcolata per questo scenario di intervento è pari all'1,9% delle famiglie italiane, quantificabili in 426.325 famiglie con almeno un componente tra i 14 e i 24 anni.

Per stimare poi quale sarebbe l'adozione incrementale delle famiglie dei centri più piccoli nel caso in cui si riuscisse ad ovviare ai problemi di digital divide infrastrutturale (ma non solo) propri di queste realtà, si è condotta una simulazione modificando la probabilità d'adozione delle famiglie che risiedono nei centri sotto i 50.000 abitanti, tramite l'uso del coefficiente stimato dalla prima regressione del paragrafo 6.1.1 per la dimensione comunale oltre i 50.000 abitanti (scenario 3).

Tabella 27 - Probabilità incrementali di adozione di banda larga per lo scenario 3, in punti percentuali, per cluster di utenza potenziale e situazione economica (anno 2012)

| Indice economico | Somma delle probabilità di utenza dei componenti della famiglia | | | | | | | | | | |
|------------------|---|-------|-------|-------|-------|-------|-------|-------|-------|------|--------|
| | ≤0,5 | 0,5-1 | 1-1,5 | 1,5-2 | 2-2,5 | 2,5-3 | 3-3,5 | 3,5-4 | 4-4,5 | >4,5 | Totale |
| 0-0,1 | 0,4 | 1,0 | 2,0 | 2,6 | 2,4 | 1,9 | 1,5 | 0,8 | 2,4 | 0,6 | 1,3 |
| 0,1-0,2 | 0,5 | 1,7 | 2,7 | 2,3 | 2,6 | 2,5 | 2,0 | 1,8 | 1,0 | 1,3 | 1,6 |
| 0,2-0,3 | 0,7 | 2,1 | 2,8 | 2,5 | 2,4 | 2,0 | 1,7 | 0,9 | 1,4 | 0,0 | 1,7 |
| 0,3-0,4 | 0,8 | 2,5 | 3,2 | 2,8 | 2,2 | 1,9 | 0,9 | 1,4 | 0,9 | 0,5 | 1,9 |
| 0,4-0,5 | 1,2 | 2,5 | 3,2 | 2,3 | 2,6 | 2,1 | 2,3 | 1,1 | 1,6 | 2,3 | 2,1 |
| 0,5-0,6 | 1,2 | 2,7 | 4,0 | 2,6 | 2,2 | 2,1 | 2,1 | 1,6 | 0,8 | 0,2 | 2,3 |
| 0,6-0,7 | 0,9 | 2,3 | 2,7 | 2,7 | 3,0 | 2,0 | 1,6 | 2,3 | 1,8 | 1,2 | 2,0 |
| 0,7-0,8 | 1,4 | 2,6 | 3,0 | 2,1 | 2,3 | 1,7 | 1,6 | 1,2 | 1,6 | 0,8 | 2,0 |
| 0,8-0,9 | 2,1 | 2,7 | 3,6 | 2,0 | 2,0 | 1,7 | 1,5 | 0,8 | 0,8 | 1,0 | 2,3 |
| 0,9-1 | 1,6 | 3,1 | 1,9 | 1,6 | 1,8 | 0,9 | 1,2 | 1,3 | 1,3 | 0,7 | 1,8 |
| Totale | 1,0 | 2,3 | 3,0 | 2,3 | 2,3 | 1,9 | 1,7 | 1,3 | 1,2 | 0,9 | 1,9 |

In questo caso, l'effetto della simulazione è molto più distribuito che nei casi precedenti lungo le due variabili di utenza potenziale e situazione economica, perché più distribuito su queste due variabili è anche l'intervento ipotizzato. L'adozione addizionale complessiva

calcolata per questo scenario di intervento è di nuovo di circa l'1,9% delle famiglie italiane, quantificabili questa volta in 424.127 famiglie residenti in comuni sotto i 50.000 abitanti.

Infine, è stata condotta una simulazione che riunisce tutti e tre gli interventi menzionati (scenario 4). Per quanto riguarda la modifica della probabilità di avere una situazione economica ottima o adeguata, quando una stessa famiglia rientrava nei criteri di selezione di entrambi gli scenari di incentivazione economica, si è adottato il criterio della probabilità maggiore tra le due risultanti dalla simulazione.

Tabella 28 - Probabilità incrementali di adozione di banda larga per lo scenario 4, in punti percentuali, per cluster di utenza potenziale e situazione economica (anno 2012)

| Indice economico | Somma delle probabilità di utenza dei componenti della famiglia | | | | | | | | | | Totale |
|------------------|---|-------|-------|-------|-------|-------|-------|-------|-------|------|--------|
| | ≤0,5 | 0,5-1 | 1-1,5 | 1,5-2 | 2-2,5 | 2,5-3 | 3-3,5 | 3,5-4 | 4-4,5 | >4,5 | |
| 0-0,1 | 7,4 | 25,1 | 32,0 | 31,7 | 32,8 | 26,6 | 19,7 | 13,4 | 13,6 | 3,3 | 21,7 |
| 0,1-0,2 | 8,5 | 26,3 | 29,1 | 22,4 | 20,5 | 16,1 | 11,7 | 10,9 | 7,3 | 5,5 | 17,7 |
| 0,2-0,3 | 8,9 | 25,1 | 24,3 | 17,6 | 13,5 | 10,5 | 8,2 | 6,1 | 6,1 | 0,0 | 15,2 |
| 0,3-0,4 | 10,5 | 24,6 | 21,3 | 15,0 | 11,3 | 8,2 | 4,9 | 4,8 | 3,2 | 2,0 | 15,1 |
| 0,4-0,5 | 11,6 | 23,8 | 20,6 | 13,5 | 9,8 | 6,9 | 5,3 | 3,6 | 3,3 | 4,0 | 14,6 |
| 0,5-0,6 | 1,2 | 2,7 | 4,0 | 2,6 | 2,2 | 2,1 | 2,1 | 1,6 | 0,8 | 0,2 | 2,3 |
| 0,6-0,7 | 0,9 | 2,3 | 2,7 | 2,7 | 3,0 | 2,0 | 1,6 | 2,3 | 1,8 | 1,2 | 2,0 |
| 0,7-0,8 | 1,4 | 2,6 | 3,0 | 2,1 | 2,3 | 1,7 | 1,6 | 1,2 | 1,6 | 0,8 | 2,0 |
| 0,8-0,9 | 2,1 | 2,7 | 3,6 | 2,0 | 2,0 | 1,7 | 1,5 | 0,8 | 0,8 | 1,0 | 2,3 |
| 0,9-1 | 1,6 | 3,1 | 1,9 | 1,6 | 1,8 | 0,9 | 1,2 | 1,3 | 1,3 | 0,7 | 1,8 |
| Totale | 5,9 | 13,1 | 15,1 | 9,0 | 9,6 | 6,6 | 5,0 | 2,9 | 4,1 | 1,9 | 9,2 |

L'effetto congiunto dei 3 scenari di intervento produce una adozione aggiuntiva ancora più alta soprattutto nei segmenti in cui si concentrava l'effetto dell'incentivo ipotizzato per lo scenario 1. L'adozione addizionale complessiva calcolata per l'insieme dei 3 scenari di intervento è di circa il 9,2% delle famiglie italiane, quantificabili in ben 2.025.909 famiglie. Tale numero è inferiore alla somma degli adottanti dei 3 casi singoli perché con ogni probabilità alcune famiglie rientrano nei criteri di intervento di più di uno scenario.

Una seconda classificazione interessante, per la valutazione di politiche di incentivazione della domanda, deriva dalla partizione delle famiglie effettuata attraverso gli alberi di classificazione. In questo senso, è possibile classificare le famiglie italiane a seconda dell'appartenenza dei loro componenti ai diversi nodi della partizione operata nel paragrafo 6.1.3. In particolare, si propongono le seguenti categorie:

- 5) Famiglie con componenti appartenenti ai nodi che risultano essere in gran parte utenti (quota di utenti maggiore o uguale al 69% della popolazione), segnatamente i nodi A, C, E, G;
- 6) Famiglie non appartenenti alla classe precedente, ma i cui componenti comprendono dei bambini 0-14 anni, che sono potenzialmente utenti entro pochi anni;
- 7) Famiglie non appartenenti alle due classi precedenti, ma i cui componenti ricadono in tutto o in parte nei nodi B, D ed F, che presentano tassi di utenza intorno al 40%;
- 8) Famiglie non appartenenti alle precedenti, i cui componenti sono dunque interamente ricompresi nel nodo H, che presenta un tasso di utenza molto basso, pari al 12,9%.

Si noti che mentre nella classificazione della popolazione operata dall'albero i bambini tra i 6-14 anni ricadono in diversi nodi (in particolare i nodi da A a D), qui essi vengono tutti ricompresi nella seconda classe, a meno che un altro componente della famiglia non appartenga alle classi A, C, E o G. Tale scelta è effettuata al fine di isolare l'effetto di tali componenti giovanissimi, data la loro condizione di cittadini, e di utenti, in divenire.

Nella tabella seguente è riportata la quota delle famiglie italiane in ciascuna delle classi esposte, per gli anni 2005-2012.

Tabella 29 – Quota delle famiglie italiane all'interno di ciascuna classe sulla base del nodo di appartenenza dei componenti (anni 2005 – 2012)

| | | Anno | | | | | | | |
|---------------------------------|---|-------|-------|-------|-------|-------|-------|-------|-------|
| | | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 |
| Classificazione famiglia | <i>Con componenti appartenenti ai nodi A,C,E e G</i> | 55,1% | 53,8% | 53,6% | 51,6% | 52,1% | 52,6% | 52,2% | 50,8% |
| | <i>Senza componenti dei nodi A,C,E e G, ma con bambini</i> | 6,5% | 6,6% | 6,5% | 7,1% | 6,9% | 6,7% | 6,3% | 6,8% |
| | <i>Senza componenti nei nodi A,C,E e G e senza bambini, ma con componenti nei nodi B,D ed F</i> | 8,9% | 9,6% | 10,1% | 10,6% | 10,6% | 10,6% | 11,3% | 11,6% |
| | <i>Con soli componenti del nodo H</i> | 29,5% | 29,9% | 29,8% | 30,7% | 30,3% | 30,1% | 30,1% | 30,8% |

Si noti come le due classi di gran lunga più consistenti sono: la prima, che rappresenta le famiglie con componenti appartenenti ai nodi di utenti A, C, E e G, con quote sempre superiori al 50% delle famiglie italiane, anche se in lieve diminuzione; l'ultima, che

rappresenta le famiglie con i soli componenti del nodo H, in gran parte non utenti, che si attesta intono al 30% delle famiglie italiane.

Per quanto riguarda la prima classe, essa è per costruzione già ampiamente formata da utenti, cui non dovrebbe essere necessario rivolgere politiche di incentivazione della domanda.

La classe di soli componenti del nodo H rappresenta invece uno “zoccolo duro” di famiglie in stragrande maggioranza composte di non utenti anziani, che vivono in famiglie composte esclusivamente di persone mature e che non hanno (più) accesso a posizioni lavorative che abilitino all’uso di Internet. Si tratta dunque di famiglie che è difficile raggiungere attraverso policy di incentivazione della domanda di Internet, perché risultano fortemente penalizzate nell’uso per motivi strutturali (età elevata, titolo di studio basso, condizione e posizione professionale sfavorevoli), non modificabili con questo tipo di politiche.

7.2 Classificazione delle famiglie sulla base delle classi latenti di uso¹⁶

Un’altra classificazione derivante dalle analisi del capitolo precedente, utilizzabile per individuare le famiglie oggetto di policy, è basata sulle classi latenti dei componenti della famiglia, descritte nel paragrafo 6.2.1. Tali classi sono stimate per individui tra 15 e 74 anni, tuttavia, per completezza, in questa classificazione sono considerati anche i bambini tra 0 e 14 anni e gli individui di 75 anni e più, nonché un piccolo numero di individui non classificati a causa di dati mancanti.

Ne deriva la seguente partizione delle famiglie italiane, in cui ognuna delle classi è costruita per differenza con la precedente:

- 1) Famiglie con almeno un utente appartenente alle classi di utenti più evoluti, prevalentemente giovani, cioè la 1 e la 6;
- 2) Famiglie senza utenti delle classi 1 e 6, ma con utenti esperti, prevalentemente maturi, della classe 5;
- 3) Famiglie senza utenti delle classi di uso più completo (1, 5 e 6), ma di cui fanno parte utenti meno sofisticati (appartenenti alle classi 2, 3 e 4);

¹⁶ Un’analisi simile, di tipo divulgativo, è stata pubblicata da Claudio Leporelli (DIAG -Sapienza) e Cosimo Dolente sulla rivista online AgendaDigitale.eu (<http://www.agendadigitale.eu>).

- 4) Famiglie senza utenti classificati, ma in cui sono presenti non utenti della classe 9, che segnalano spesso motivi economici per il non uso;
- 5) Famiglie non appartenenti alle classi precedenti, in cui sono però presenti bambini (0-14 anni);
- 6) Famiglie in cui sono presenti non utenti delle classi 7 o 8, non appartenenti alle classi precedenti;
- 7) Famiglie composte esclusivamente da persone di 75 anni o più;
- 8) Famiglie esclusivamente composte da individui non classificati, cui non appartengano utenti, non utenti della classe 9 o bambini.

La consistenza delle famiglie in ciascuno di questi gruppi, al 2014, è riportata in tabella seguente.

Tabella 30 – Consistenza delle famiglie italiane, secondo la classificazione sulla base delle classi latenti (anno 2014)

| Classificazione delle famiglie | Numero di famiglie | Quota sulle famiglie | Numero di individui | Quota sulla popolazione |
|--|---------------------------|-----------------------------|----------------------------|--------------------------------|
| <i>Con individui della classe 1 o 6</i> | 6.967.078 | 27,6% | 21.536.563 | 35,6% |
| <i>Con individui della classe 5</i> | 3.089.653 | 12,2% | 7.575.979 | 12,5% |
| <i>Con individui della classe 2, 3 o 4</i> | 5.683.742 | 22,5% | 14.623.685 | 24,2% |
| <i>Con individui della classe 9 e senza utenti</i> | 1.340.809 | 5,3% | 3.436.457 | 5,7% |
| <i>Senza utenti o individui della classe 9, ma con bambini</i> | 423.933 | 1,7% | 1.702.954 | 2,8% |
| <i>Con solo non utenti delle classi 7 e 8</i> | 4.047.673 | 16,0% | 6.586.266 | 10,9% |
| <i>Con solo anziani di 75 o più</i> | 3.254.214 | 12,9% | 4.107.378 | 6,8% |
| <i>Con individui non classificati</i> | 422.701 | 1,7% | 911.687 | 1,5% |
| Totale | 25.229.804 | 100,0% | 60.480.969 | 100,0% |

Al fine di comprendere la possibile utilità di tale classificazione, viene riportata un'analisi delle caratteristiche delle classi di famiglie individuate. Nel seguito, al fine di facilitare la lettura delle tabelle, esse sono accorpate in 5 gruppi, a seconda che contengano:

- 1) un qualche utente avanzato (classi 1,5 e 6);
- 2) un qualche utente di altro tipo (classi 2,3 e 4);
- 3) nessun utente ma almeno un non utente per motivi economici (classe 9);
- 4) nessun utente o non utente di classe 9, ma almeno un bambino (0-14);
- 5) solo altre classi di non utenti e/o persone di 75 anni o più.

Tale raggruppamento è operato sulla base della permeabilità di queste tipologie di famiglie alla banda larga, come si evince anche dalla seguente tabella sulla connessione posseduta dalla famiglia.

Tabella 31 - Suddivisione delle famiglie per classe latente dei componenti e connessione ad Internet da casa (anno 2014)

| | | Tipologia di connessione ad Internet della famiglia | | |
|---------------------------------|--|---|--------------------|-------------------------|
| | | Banda larga fissa | Banda larga mobile | Banda stretta o nessuna |
| Classificazione famiglia | <i>Con individui della classe 1, 5 o 6</i> | 75,4% | 21,2% | 3,4% |
| | <i>Con individui della classe 2, 3 o 4</i> | 52,7% | 34,3% | 13,0% |
| | <i>Con individui della classe 9 e senza utenti</i> | 10,4% | 5,5% | 84,1% |
| | <i>Senza utenti o individui della classe 9, ma con bambini</i> | 30,0% | 17,5% | 52,6% |
| | <i>Con solo non utenti delle classi 7 e 8 o anziani (75 o più)</i> | 7,3% | 2,2% | 90,5% |

Si noti infatti che per la prima categoria, le famiglie che contengono utenti avanzati, la connessione a banda larga è quasi sempre presente a casa (97% circa dei casi). Per il secondo gruppo di famiglie, la situazione è solo lievemente meno rosea, con l'87% delle famiglie che dispone già di una connessione a banda larga. Tuttavia, è sensibilmente più bassa, rispetto al gruppo precedente, la quota di famiglie con connessione a larga banda fissa. Come si diceva in altre parti del documento, questa forma di connessione è considerata migliore di quella mobile, per le famiglie, perché normalmente non limitata in termini di traffico mensile.

Il terzo gruppo possiede Internet solo in una quota limitata dei casi, meno del 16%, ma ciò costituisce comunque un'anomalia, dato che esso non è composto da famiglie con componenti classificati come utenti. Si tratta infatti di nuclei che non hanno al loro interno utenti Internet tra 15 e 74 anni, ma che hanno almeno un membro che appartiene alla classe di coloro che dichiarano in gran parte di non utilizzare Internet a causa degli alti costi degli strumenti o del collegamento. La spiegazione dell'anomalia di questo gruppo può essere del tutto simile a quella specificata nel seguito per il quarto gruppo, anch'esso anomalo.

A questo appartengono le famiglie senza utenti né non utenti della classe 9, ma con bambini. Si noti la particolarità di questa categoria, in cui, pur non essendo presenti utenti tra 15 e 74 anni, un 47% delle famiglie possiede una connessione a banda larga, fissa o mobile, a casa. Tale apparente anomalia, molto più pronunciata rispetto a quella già segnalata per il terzo gruppo, può essere dovuta a due cause. In parte, i bambini tra 0 e 14 anni presenti in famiglia,

non classificati, sono comunque utenti della Rete, e possono dunque aver spinto le proprie famiglie all'adozione. Inoltre, poiché si tratta di un numero contenuto di osservazioni del campione, è possibile che esso presenti errori di risposta non sia comunque affidabile da un punto di vista statistico.

Le famiglie del quinto e ultimo gruppo sono quelle in cui si riscontrano i tassi più bassi di adozione di una connessione a banda larga (inferiori al 10%). Al contempo, esse sono anche quelle a maggiore marginalità sociale, poiché composte in gran parte di individui anziani, che non sono “contagiabili” da utenti o giovani nello stesso nucleo. Come si vedrà meglio nel seguito, pensare a politiche di incentivo per questo tipo di famiglie è quindi difficile, tenendo conto della suddetta marginalità sociale e del fatto che un semplice incentivo economico non dovrebbe essere in grado di far fronte alle problematiche di interesse e di alfabetizzazione digitale degli individui che ne fanno parte.

È possibile analizzare le caratteristiche delle famiglie nelle macro classi appena proposte incrociando l'appartenenza a queste classi con altre informazioni. Ad esempio, per quanto riguarda la ripartizione territoriale di appartenenza, si ottiene la seguente tabella.

Tabella 32 - Suddivisione delle famiglie nelle diverse ripartizioni geogarfiche di residenza, per classe latente dei componenti (anno 2014)

| | | Classificazione famiglia | | | | |
|---|-------------------|--|--|--|--|--|
| | | <i>Con individui della classe 1, 5 o 6</i> | <i>Con individui della classe 2, 3 o 4</i> | <i>Con individui della classe 9 e senza utenti</i> | <i>Senza utenti o individui della classe 9, ma con bambini</i> | <i>Con solo non utenti delle classi 7 e 8 o anziani (75 o più)</i> |
| Ripartizione geografica di residenza | <i>Nord-Ovest</i> | 43% | 22% | 3% | 1% | 29% |
| | <i>Nord-Est</i> | 43% | 23% | 3% | 1% | 29% |
| | <i>Centro</i> | 42% | 22% | 5% | 2% | 27% |
| | <i>Sud</i> | 33% | 23% | 9% | 3% | 30% |
| | <i>Isole</i> | 35% | 24% | 8% | 2% | 30% |

Il Sud e le Isole appaiono le aree con la quota più alta di famiglie incentivabili all'adozione attraverso sussidi economici (terza colonna). In maniera speculare, le aree del Nord e del Centro sono quelle che presentano tassi più elevati (fino a 10 punti in più) di famiglie cui appartengono utenti avanzati. La variabilità nelle altre categorie è invece molto bassa. In particolare, la categoria delle famiglie più escluse da un punto di vista sociale, l'ultima, è presente su tutto il territorio nazionale con quote comprese tra il 27% e il 30% del totale,

anche nelle aree che presentano più alte quote di famiglie con utenti evoluti. Ciò è probabilmente dovuto, in primo luogo, all'elevato tasso di vecchiaia anche di queste zone, come di tutto il paese, che non può essere controbilanciato dalla maggiore dinamicità del loro tessuto produttivo, e dal reddito medio più elevato che al Sud.

Passando all'età, i dati della seguente tabella analizzano l'appartenenza ai cinque raggruppamenti di famiglia per le diverse fasce di età individuali. Essi gettano diverse ombre sulla possibilità che i giovani risolveranno autonomamente, col tempo, i loro problemi di esclusione digitale

Tabella 33 – Suddivisione degli individui per fasce di età, sulle diverse categorie di famiglie per classi latenti dei componenti (anno 2014)

| Fascia di età | Classificazione famiglia | | | | |
|---------------|-------------------------------------|-------------------------------------|---|---|---|
| | Con individui della classe 1, 5 o 6 | Con individui della classe 2, 3 o 4 | Con individui della classe 9 e senza utenti | Senza utenti o individui della classe 9, ma con bambini | Con solo non utenti delle classi 7 e 8 o anziani (75 o più) |
| 0-14 | 50,5% | 32,3% | 8,5% | 8,7% | 0,0% |
| 15-19 | 74,7% | 19,0% | 2,9% | 1,8% | 0,9% |
| 20-29 | 67,9% | 21,5% | 6,1% | 1,5% | 0,8% |
| 30-39 | 54,7% | 30,0% | 8,1% | 3,8% | 2,0% |
| 40-49 | 71,5% | 20,1% | 4,2% | 2,0% | 0,8% |
| 50-64 | 52,4% | 24,9% | 5,1% | 1,0% | 14,5% |
| 65 e più | 15,9% | 14,5% | 2,8% | 0,5% | 64,2% |

Il dato positivo è che una porzione consistente dei quattordicenni, una volta che compie i 15 anni e viene dunque classificata, va ad incrementare la quota degli utenti tra 15 e 19 anni appartenenti alle famiglie con utenti avanzati (che arriva così al 74,7%). In questa fascia di età resta però una quota del 19% di utenti superficiali (e che vivono in famiglie senza utenti avanzati) e un 5,6% di non utenti. In altri termini, il contesto familiare sembrerebbe impedire a circa un quarto dei nativi digitali tra 15 e 19 anni di diventare utenti Internet in modo ricco e sofisticato.

Questa analisi è confermata dalle differenze tra famiglie con utenti avanzati e famiglie con utenti superficiali per quanto riguarda il tipo di connessione e gli strumenti utilizzati per connettersi. Abbiamo già visto che le famiglie con utenti superficiali (classi 2, 3 e 4) dispongono di connessione fissa nel 52,7% dei casi, contro il 75,4% delle famiglie con utenti avanzati. Ebbene, alla scelta del tipo di connessione si accompagna una scelta dello strumento

utilizzato, che fa ritenere che la famiglia abbia talvolta scarsa dimestichezza con gli strumenti informatici. Ciò è rilevabile nelle due tabelle che seguono. In esse si propone separatamente, per le famiglie che hanno una connessione fissa, e per quelle che hanno solo una connessione mobile, una partizione in base alla tipologia del più sofisticato e flessibile strumento utilizzato per connettersi, partendo dal PC, passando per smartphone e tablet e arrivando a televisori connessi e console per giochi.

Tabella 34 – *Suddivisione delle famiglie all'interno delle diverse categorie per classi latenti, per strumento di connessione ad Internet. Famiglie con connessione fissa a banda larga (anno 2014)*

| | | Strumento di connessione della famiglia | | |
|-------------------------|---|---|---|---|
| | | PC | Dispositivo mobile (ma non tramite PC) | TV o console (ma non tramite PC o dispositivo mobile) |
| Classe familiare | Con individui della classe 1 o 6 | 99,0% | 0,9% | 0,0% |
| | Con individui della classe 5 e senza classi 1 o 6 | 98,6% | 1,4% | 0,0% |
| | Con individui della classe 2, 3 0 4 e senza classi 1, 5 o 6 | 98,6% | 1,3% | 0,1% |

Tabella 35 – *Suddivisione delle famiglie all'interno delle diverse categorie per classi latenti, per strumento di connessione ad Internet. Famiglie con sola connessione mobile a banda larga (anno 2014)*

| | | Strumento di connessione della famiglia | | |
|-------------------------|---|---|---|---|
| | | PC | Dispositivo mobile (ma non tramite PC) | TV o console (ma non tramite PC o dispositivo mobile) |
| Classe familiare | Con individui della classe 1 o 6 | 78,0% | 21,2% | 0,0% |
| | Con individui della classe 5 e senza classi 1 o 6 | 79,7% | 19,6% | 0,0% |
| | Con individui della classe 2, 3 0 4 e senza classi 1, 5 o 6 | 66,5% | 30,1% | 0,1% |

Le famiglie che dispongono di connessione fissa utilizzano quasi sempre il Personal Computer. Invece, per le famiglie che si dotano soltanto di connessioni mobili, non è trascurabile la quota di famiglie che non utilizzano il PC.

Combinando le informazioni su connessione e strumento otteniamo che circa un quarto degli utenti superficiali non utilizza a casa un PC. Per tutte le famiglie l'utilizzo di televisori connessi e console per giochi non svolge un ruolo incrementale rispetto a PC, smartphone e tablet. Questi dati si prestano a valutazioni in chiaroscuro: è positivo che il PC non sia più

un collo di bottiglia per l'accesso a Internet, ma sembra indubbio che chi ne evita l'uso sia destinato a rimanere vincolato ad una utilizzazione superficiale, che ne impedisce probabilmente lo sviluppo di *skill* valorizzabili nel mondo del lavoro. Ciò è particolarmente grave per le fasce di età più giovani.

D'altra parte i dati segnalano una opportunità che non appare ancora sfruttata: il possibile aumento delle connessioni attraverso *connected TV* da parte di coloro che ancora oggi non usano Internet. Se questo non avverrà, circa il 30% delle famiglie Italiane potrebbe restare senza connessione, rendendo problematico, ancora per molti anni, il raggiungimento degli obiettivi dell'Agenda Digitale Europea (Commissione Europea, 2014).

Passando alle considerazioni più strettamente di policy, i diversi gruppi sociali individuati dalle classi latenti e dalla classificazione appena esposta agiscono in contesti profondamente diversi. Come diffondere l'uso di Internet, e una cultura completa della Rete, in tali contesti?

Si noti che l'incentivo economico può in questo caso avere effetto su un segmento specifico di famiglie, quelle della quarta categoria, in cui sono presenti componenti della classe 9, il cui uso è in gran parte impedito da problemi di natura economica. Si tratta di 1,3 milioni di famiglie, cui appartengono circa 3,5 milioni di persone. Qualora, attraverso un piano mirato, si riuscisse ad incentivare l'adozione di una larga parte di tali famiglie, si otterrebbe dunque un incremento significativo nella diffusione di Internet. Tuttavia, una gran parte dei non utenti non sarebbe incentivata da un tale tipo di intervento, e probabilmente parte degli attuali individui nella classe 9 diverrebbero comunque utenti poco maturi, ricadendo nelle classi 2, 3 e 4, che, per i motivi esposti durante l'analisi della classificazione, sono comunque utenti poco esperti.

Valutare invece politiche adatte a superare i problemi relativi a disinteresse, mancanza di capacità e uso superficiale della Rete, che colpiscono i non utenti delle classi 7 e 8 e le relative famiglie del gruppo 6, ma anche gli anziani del gruppo 7 (in totale, più di 10 milioni di non utenti), non è per nulla agevole. Si tratta infatti, perlopiù, di una questione culturale, come ampiamente dimostrato nel corso delle analisi del capitolo 6.

Quella parte di non utenti, ma anche di utenti superficiali, che sia ancora parte attiva della società (occupati, in cerca di occupazione, studenti), può essere più facilmente raggiunta da interventi formativi nei rispettivi contesti. Gli studenti possono essere oggetto di interventi

di formazione mirati, nelle scuole e nelle università. Particolarmente importante, in questo caso, non è tanto lo sviluppo di competenze pratiche nell'uso della rete e dei relativi strumenti, quanto apprendere i limiti, i pericoli e le potenzialità del mondo digitale, formando meta-competenze per un uso consapevole di Internet (Bagnara e Matarazzo, 2015).

Un altro contesto in cui è relativamente agevole attivare iniziative di formazione è rappresentato dai luoghi di lavoro. È possibile ipotizzare, infatti, politiche di incentivo alle imprese che prevedano corsi di formazione digitale per i propri dipendenti, sullo stile di quanto avviene, ad esempio, per quello che riguarda i corsi obbligatori sulla sicurezza aziendale.

Per i disoccupati, in cerca di prima o nuova occupazione, si potrebbe fare largo uso di sessioni sulla cultura digitale nei corsi professionalizzanti operati dai centri per l'impiego, con insegnamenti pratici, misti a informazioni di carattere più generale sull'economia digitale e sulla possibilità di utilizzare le competenze digitali all'interno del moderno mercato del lavoro.

Una grossa fetta dei non utenti è però in condizioni di inattività: casalinghe e pensionati in primis. Si tratta dei quasi 11 milioni di membri delle 8 milioni di famiglie dell'ultimo gruppo. Raggiungere queste tipologie di persone, che non possono essere intercettate tipicamente in luoghi di aggregazione (se non in casi specifici, quali i centri anziani), rappresenta una sfida importante. Sicuramente, vale qui il paradigma per cui si impara con più volontà, e dunque più facilmente, ciò che si ritiene essere più utile. Non a caso, una buona porzione di non utenti giudica non interessante, o non utile la Rete. Si potrebbe quindi tentare di rendere disponibili online, in maniera semplificata, alcuni servizi che le persone più anziane potrebbero ritenere utili. Dato il basso grado di competenze e di consapevolezza di questo tipo di utenti, che si affaccerebbero per la prima volta al mondo di Internet, sarebbe necessario anche un importante sforzo per la protezione dalle frodi e da altri attacchi malevoli in cui essi potrebbero incorrere. Inoltre, sarebbe opportuno sviluppare una rete di punti di aiuto ai cittadini più anziani o comunque poco esperti nell'uso della Rete.

Conclusioni

Alla luce delle analisi svolte all'interno di questo progetto di ricerca, si può affermare che, con il diffondersi dell'uso di Internet tra la popolazione, la questione della domanda e dell'effettivo superamento di ogni divario nell'uso della Rete è ben lontana dall'essere risolta. Al contrario, il digital divide si rivela essere un tema molto più complesso e articolato di quanto potesse apparire in passato. La questione è particolarmente rilevante, poiché un uso diffuso e maturo della Rete da parte della quasi totalità della popolazione costituisce una precondizione affinché il nostro paese colga appieno i vantaggi offerti dal mondo digitale.

In particolar modo, già dalle prime elaborazioni descrittive è emerso un quadro che vede una certa distanza dell'Italia dal resto dei paesi europei, e in particolar modo da quelli con popolazione comparabile: Gran Bretagna, Germania, Francia e Spagna. Tale distanza è evidente sia in termini di uso di Internet da parte degli individui, che di diffusione delle connessioni, con la parziale eccezione della banda larga mobile (cfr. capitolo 1).

Al fine di indagare al meglio le ragioni di questo ritardo, il presente lavoro si è dunque concentrato sull'analisi delle determinanti della domanda di Internet in Italia, e sulle ragioni del non uso.

Come sostrato delle analisi, si è cercato dapprima di fornire un nuovo *framework* interpretativo del fenomeno, che fosse in grado di catturarne la complessità in maniera più appropriata, rispetto alla più diffusa bipartizione tra adottanti e non adottanti. A partire dall'analisi della letteratura, si è dunque formulata una definizione di digital divide in quattro stadi, che riguardano aspetti diversi dell'adozione e dell'uso di Internet (cfr. capitolo 5).

In particolare, il primo di questi stadi, riguardante l'infrastruttura, sembrerebbe al momento ricoprire un ruolo di minor rilievo, soprattutto considerando che i dati Eurostat indicano una copertura quasi totale delle reti a banda larga fissa e mobile, con velocità di almeno 2 Mbit/s, nel nostro paese (cfr. paragrafo 5.1). Il digital divide infrastrutturale potrebbe riemergere con forza nel prossimo futuro, e probabilmente lo farà, man mano che servizi ad alto consumo di banda si diffonderanno tra la popolazione di utenti. Alcuni dei servizi di questo tipo esistono già, come ad esempio: la fruizione contemporanea di più video in streaming ad alta qualità, attraverso una stessa connessione; l'uso di applicativi *cloud* avanzati, che comprendano lo *storage*, come anche l'elaborazione remota; la domotica avanzata resa possibile dall'*Internet of Things*. Altri servizi ad alto consumo di banda si affacceranno

sicuramente sul mercato nel futuro, grazie alla diffusione delle reti NGAN, come la storia di Internet ci insegna.

Tuttavia, pur considerando il digital divide infrastrutturale come risolto o comunque affrontato con ricchezza di finanziamenti pubblici, o in taluni casi in *stand-by* in attesa degli sviluppi appena descritti, i dati del caso italiano testimoniano un importante secondo divide, relativo all'effettiva adozione (*take-up*) dei servizi Internet (cfr. capitolo 3).

Uno degli obiettivi principali nel nostro paese deve essere dunque quello di cercare di spingere le ancora corpose porzioni della popolazione di non utenti ad adottare la Rete. Per far ciò, è però necessario individuare correttamente le caratteristiche di tali segmenti, con le relative motivazioni per il mancato utilizzo, in modo da disegnare politiche di sostegno davvero efficaci.

A questo scopo è rivolta una parte consistente del presente lavoro, che mostra, attraverso l'analisi dei microdati provenienti dall'indagine Istat "Aspetti della Vita Quotidiana" su diversi anni, come il problema della mancata adozione, sia a livello individuale che familiare, sia dovuto a un insieme di problematiche in larga parte culturali, associate all'alto indice di vecchiaia della popolazione, ai bassi tassi di istruzione, all'alto tasso di inattivi (cfr. paragrafo 6.1). Inoltre, queste variabili sono tra loro fortemente associate, a coppie o gruppi più ampi. Ad esempio, nella classe dei più anziani si riscontrano anche livelli di istruzione bassi e alti tassi di inattività e di posizioni professionali manuali. La complessità del fenomeno è poi sensibilmente incrementata dalla presenza di interazioni forti tra decisioni individuali d'uso e decisioni familiari di sottoscrizione di servizi a banda larga da casa, che creano una circolarità tra le due, comportando anche alcuni problemi metodologici.

Al fianco del divide relativo all'adozione della tecnologia, è necessario però tener conto di alcuni nuovi ostacoli nella diffusione omogenea dell'uso di Internet tra la popolazione. Dai dati emerge infatti come si stiano effettivamente manifestando alcune nuove forme di digital divide, anche all'interno dell'ampia e variegata categoria degli utenti di Internet. In particolare, è necessario rivolgere l'attenzione verso gli aspetti di uso della Rete, su tutti la frequenza d'uso e la ricchezza d'uso, approssimata dalle attività specifiche svolte online. Attraverso l'uso del metodo delle classi latenti sugli utenti (cfr. paragrafo 6.2), si ottiene una netta suddivisione di questi in diverse categorie molto differenti tra loro, di cui solo alcune

sembrano cogliere appieno, già da adesso, le potenzialità fornite dalla Rete. Le tre classi di utenti “avanzati” sono costituite rispettivamente da: utenti giovanissimi per lo più in formazione (classe 6); giovani adulti istruiti, benestanti, e ben inseriti professionalmente (classe 1); utenti maturi ancora in età lavorativa, con alto tasso di attività e occupazione in posizioni qualificate (classe 5). Tutti questi utenti svolgono un buon numero di attività culturali anche *offline*. Altre categorie di utenti, che utilizzano con meno frequenza la Rete, con età media inferiore a quella degli utenti della classe 5 ma superiore a quella degli utenti della classe 1, sono meno scolarizzate, prevalentemente ancora attive sul mercato del lavoro, anche se con ruoli meno qualificati e maggiore disoccupazione rispetto alla classe 5. Esse mostrano anche un utilizzo nettamente più povero, votato a poche attività sostanzialmente di comunicazione, non sufficientemente evolute da fare di questi individui degli utenti avanzati.

Queste forme di divide sono senza dubbio di natura diversa rispetto a quelle del mancato accesso. Ad ogni modo, anche gli utenti meno avanzati subiscono limitazioni che non li mettono in grado di sfruttare appieno le possibilità fornite da Internet.

Passando a considerare i problemi posti dai non utenti, risulta essere largamente minoritaria, sebbene ancora rilevante in termini numerici, la porzione di individui e di famiglie che non adottano Internet per motivazioni economiche. Questo limita l'efficacia degli interventi di sostegno economico della domanda, sia pubblici attraverso incentivi all'adozione (come i voucher), sia privati, basati su politiche di prezzo (si pensi ad esempio agli schemi zero-rating). Un possibile ruolo dei fattori di costo è tuttavia riscontrabile nell'esistenza di una quota significativa di famiglie che dispongono solo della connessione mobile a banda larga (il 22% nel 2016). Tale scelta è con ogni probabilità dovuta a una ridotta disponibilità a pagare, che porta a dotarsi un servizio fortemente penalizzato in termini di volume di traffico consentito, a fronte di una spesa mensile sensibilmente inferiore, e di minori vincoli contrattuali.

La gran parte dei divari nell'adozione sono invece dovuti a problemi nella struttura stessa della popolazione italiana, in termini principalmente di età, di livelli di istruzione e di coinvolgimento attivo nella vita sociale del paese. Inoltre, si evidenzia come ampi segmenti della popolazione di non utenti sono emarginati socialmente, perché non attivi sul mercato

del lavoro o nello studio. Gli effetti negativi delle variabili appena citate sono poi acuite dall'impossibilità di un contagio (almeno nel contesto familiare), in quella parte della popolazione che vive nelle famiglie in cui non sono presenti utenti, o comunque membri con caratteristiche migliori in una o più delle dimensioni evidenziate. Si tratta di 11 milioni di persone, che vivo in quasi 8 milioni di famiglie.

Queste problematiche sono difficilmente superabili nel breve e nel medio periodo, in quanto intrinseche della struttura della popolazione. Il quadro è aggravato dalla palese difficoltà nell'intervenire in questo tipo di dinamiche attraverso politiche pubbliche o piani di marketing di imprese private.

Al fianco dell'importanza delle questioni culturali, un altro risultato rilevante del presente lavoro è che le differenze emerse tra gli stessi utenti non sono spiegabili in termini "classici", ovvero attraverso l'uso delle sole variabili di segmentazione. I profili di uso differenti dovrebbero dunque essere indagati con l'aiuto di dati più fini. Ad esempio, i dati disponibili sul titolo di studio non distinguono tra lauree e diplomi diversi, per contenuto disciplinare, né misurano i livelli di apprendimento acquisiti attraverso la formazione, che potrebbero differire fortemente, come mostrano le indagini PISA dell'OCSE relative ai quindicenni (OECD, 2016).

Sembra probabile, comunque, che esista una eterogeneità di fondo nei comportamenti non spiegabile attraverso variabili socio-economiche osservabili, perché legata a fattori culturali individuali. Si tenga presente poi che le analisi quantitative presuppongono spesso la razionalità degli agenti economici, intesa come intrinseca coerenza nelle scelte (Robbins, 1932). Diversi studi indicano invece come gli stessi individui non compiono le scelte in maniera del tutto razionale (Simon, 1955, 1959; Kahneman e Tversky, 1979; Read, 2007). Le metodologie utilizzate, quale la regressione logistica, tengono presente questa razionalità limitata attraverso l'inserimento di variabili aleatorie, ma ciò potrebbe non essere comunque sufficiente.

In definitiva, il messaggio che emerge dalle differenti analisi dei dati sulla popolazione italiana sul periodo di 10 anni, tra il 2005 e il 2014, è che non si possa più parlare di digital divide al

singolare, ma di una pluralità di situazioni di disuguaglianza su indicatori chiave dell'emarginazione sociale.

Il quadro risultante evidenzia profonde spaccature nella società italiana, rispetto all'uso di Internet. Da una parte, le caratteristiche strutturali della popolazione italiana dividono nettamente, e, stando alle analisi multi-periodo, sempre di più, gli utenti dai non utenti, e le famiglie connesse da quelle non connesse. Dall'altra, anche all'interno di queste categorie, esistono differenze importanti, legate sostanzialmente alle classi sociali e alla cultura personale dei singoli individui, che non sono spiegabili solo attraverso le "classiche" variabili strutturali. Tutto questo dovrebbe essere ulteriormente indagato con l'ausilio di diverse fonti di dati, col fine ultimo di supportare politiche pubbliche e iniziative private di sostegno alla domanda Internet che siano davvero efficaci, perché basate sulla reale comprensione del fenomeno *dei* digital divide italiani.

Bibliografia

- Agcom (2017) *Relazione Annuale sull'attività svolta e sui programmi di lavoro*.
- Agid e Team Digitale (2016) *Piano triennale per l'informatica nella pubblica amministrazione 2017-2019*. Available at: <http://pianotriennale-ict.readthedocs.io/it/latest/> (Consultato: 24 ottobre 2017).
- Agresti, A. (2007) *An Introduction to Categorical Data Analysis, Statistics*. doi: 10.1002/0471249688.
- Akaike, H. (1974) «A new look at the statistical model identification», *IEEE Transactionson Automatic Control*, 19(6), pagg. 716–723. doi: 10.1109/TAC.1974.1100705.
- Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M. e Ayyash, M. (2015) «Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications», *IEEE Communications Surveys and Tutorials*, 17(4), pagg. 2347–2376. doi: 10.1109/COMST.2015.2444095.
- Alexe, G., Alexe, S., Bonates, T. O. e Kogan, A. (2007) «Logical analysis of data – the vision of Peter L. Hammer», *Annals of Mathematics and Artificial Intelligence*, 49(1–4), pagg. 265–312. doi: 10.1007/s10472-007-9065-2.
- Andergassen, R., Nardini, F. e Ricottilli, M. (2017) «Innovation diffusion, general purpose technologies and economic growth», *Structural Change and Economic Dynamics*. Elsevier B.V., 40, pagg. 72–80. doi: 10.1016/j.strueco.2016.12.003.
- Bagnara, S. e Matarazzo, G. (2015) «La didattica per nativi digitali? Una didattica per la riflessione», in Campione, V. (a c. di) *La didattica nell'era digitale*. Roma: Astrid.
- Bascle, G. (2008) «Controlling for endogeneity with instrumental variables in strategic management research», *Strategic Organization*, 6, pagg. 285–327. doi: 10.1177/1476127008094339.
- Bass, F. M. (1969) «A New Product Growth for Model Consumer Durables», *Management Science*, 15(5), pagg. 215–227. doi: 10.1287/mnsc.15.5.215.
- Basu, S. e Fernald, J. G. (2008) «Information and communications technology as a general purpose technology: evidence from U.S. industry data», *Economic Review*, 8(2), pagg. 1–15.

- Bonates, T. O., Hammer, P. L. e Kogan, A. (2008) «Maximum patterns in datasets», *Discrete Applied Mathematics*, 156(6), pagg. 846–861. doi: 10.1016/j.dam.2007.06.004.
- Bonferroni, C. E. (1936) *Teoria statistica delle classi e calcolo delle probabilità*. Firenze: Regio Istituto Superiore di Scienze Economiche e Commerciali.
- Borman, S. (2004) «The Expectation Maximization Algorithm: A short tutorial».
- Boros, E., Crama, Y., Hammer, P. L., Ibaraki, T., Kogan, A. e Makino, K. (2011) «Logical analysis of data: Classification with justification», *Annals of Operations Research*, 188(1), pagg. 33–61. doi: 10.1007/s10479-011-0916-1.
- Boros, E., Hammer, P. L., Ibaraki, T. e Kogan, A. (1997) «Logical analysis of numerical data», *Mathematical Programming*, 79(1–3), pagg. 163–190. doi: 10.1007/BF02614316.
- Boros, E., Hammer, P. L., Ibaraki, T., Kogan, A., Mayoraz, E. e Muchnik, I. (2000) «An implementation of logical analysis of data», *IEEE Transactions on Knowledge and Data Engineering*, 12(2), pagg. 292–306. doi: 10.1109/69.842268.
- Bourdieu, P. (1979) «La Distinction Critique sociale du jugement – Fiche de lecture», *Pensée sociologique. XXème siècle. –100 Fiches de lectures: Les livres qui ont marqué le XXème siècle*. doi: 10.2307/3684493.
- Bourdieu, P. (1986) «The forms of capital», in *Handbook of Theory and Research for the Sociology of Education*, pagg. 241–58.
- Bozdogan, H. (1987) «Model selection and Akaike’s Information Criterion (AIC): The general theory and its analytical extensions», *Psychometrika*, 52(3), pagg. 345–370. doi: 10.1007/BF02294361.
- Breiman, L., Friedman, J. H., Olshen, R. A. e Stone, C. J. (1984) *Classification and Regression Trees, The Wadsworth statistics probability series*. doi: 10.1371/journal.pone.0015807.
- Bresnahan, T. F., Brynjolfsson, E. e Hitt, L. M. (2002) «Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence», *The Quarterly Journal of Economics*. Oxford University Press, 117(1), pagg. 339–376. doi: 10.1162/003355302753399526.
- Bresnahan, T. F. e Trajtenberg, M. (1992) «General Purpose Technologies: “Engines of growth”», *NBER Working Paper Series*, (August), pagg. 1–43.

Bruni, R. e Bianchi, G. (2015) «Effective Classification Using a Small Training Set Based on Discretization and Statistical Analysis», *IEEE Transactions on Knowledge and Data Engineering*, 27(9), pagg. 2349–2361. doi: 10.1109/TKDE.2015.2416727.

Bruni, R., Bianchi, G., Dolente, C. e Leporelli, C. (2017) «Logical Analysis of Data as a Tool for the Analysis of Probabilistic Socio-Economic Behavior», *Preprint submitted to Computers & Operations Research*.

Büchi, M., Just, N. e Latzer, M. (2016) «Modeling the second-level digital divide: A five-country study of social differences in Internet use», *New Media & Society*. SAGE PublicationsSage UK: London, England, 18(11), pagg. 2703–2722. doi: 10.1177/1461444815604154.

Cardona, M., Schwarz, A., Yurtoglu, B. B. e Zulehner, C. (2009) «Demand estimation and market definition for broadband Internet services», *Journal of Regulatory Economics*, 35(1), pagg. 70–95. doi: 10.1007/s11149-008-9076-x.

Castelli, F. e Leporelli, C. (1993) «Critical mass of users versus critical mass of services in a multiproduct information service system», *Information Economics and Policy*, 5(4), pagg. 331–355. doi: 10.1016/0167-6245(93)90005-2.

Cerioli, A. e Grossi, L. (2009) «Gli alberi di classificazione», in Zani, S. e Cerioli, A. (a c. di) *Analisi dei dati e data mining per le decisioni aziendali*. Milano: Giuffrè editore, pagg. 513–572.

Cerno, L. e Pérez Amaral, T. (2006) «Demand for internet access and use in Spain», *Contributions to Economics*, pagg. 333–353. doi: 10.1007/3-7908-1746-5-18.

Chaudhuri, A., Flamm, K. e Horrigan, J. (2005) «An analysis of the determinants of internet access», *Telecommunications Policy*, 29(9–10), pagg. 731–755. doi: 10.1016/j.telpol.2005.07.001.

Choudrie, J. e Dwivedi, Y. K. (2005) «the Demographics of Broadband Residential Consumers in a British Local Community: the London Borough of Hillingdon», *The Journal of Computer Information Systems*, 45(4), pagg. 93–101.

Collins, L. M., Fidler, P. L., Wugalter, S. E. e LONG, J. D. (1993) «Goodness-of-fit testing for latent class models», *Multivariate Behavioral Research*, 28(3), pagg. 375–389. doi: 10.1207/s15327906mbr2803_4.

Commissione Europea (2010) *Un'Agenda Digitale Europea*. Commissione Europea.

Commissione Europea (2014) *Digital Agenda for Europe - Rebooting Europe's economy*, European Commission. doi: 10.2775/41229.

Commissione Europea (2017) *Connectivity - Broadband market developments in the EU*.

Cover, T. e Hart, P. (1967) «Nearest neighbor pattern classification», *IEEE Transactions on Information Theory*, 13(1), pagg. 21–27. doi: 10.1109/TVT.1967.1053964.

Crama, Y. e Hammer, P. L. (2011) *Boolean functions: theory, algorithms, and applications*. Cambridge University Press.

Crama, Y., Hammer, P. L. e Ibaraki, T. (1988) «Cause-effect relationships and partially defined Boolean functions», *Annals of Operations Research*, 16(1), pagg. 299–325. doi: 10.1007/BF02283750.

Dayton, C. M. e Macready, G. B. (1988) «Concomitant-variable latent-class models», *Journal of the American Statistical Association*, 83(401), pagg. 173–178. doi: 10.1080/01621459.1988.10478584.

Van Deursen, A. J. e Van Dijk, J. A. (2013) «The digital divide shifts to differences in usage», *New Media & Society*, 16(3), pagg. 507–526. doi: 10.1177/1461444813487959.

Van Deursen, A. J. e Van Dijk, J. A. (2015) «Toward a Multifaceted Model of Internet Access for Understanding Digital Divides: An Empirical Investigation», *The Information Society*, 31(5), pagg. 379–391. doi: 10.1080/01972243.2015.1069770.

Van Deursen, A. J., Van Dijk, J. A. e Ten Klooster, P. (2015) «Increasing inequalities in what we do online: A longitudinal cross sectional analysis of Internet activities among the Dutch population (2010 to 2013) over gender, age, education, and income», *Telematics and Informatics*. Elsevier Ltd, 32(2), pagg. 259–272. doi: 10.1016/j.tele.2014.09.003.

van Deursen, A. J. e Helsper, E. J. (2015) «The Third-Level Digital Divide: Who Benefits Most from Being Online?», in, pagg. 29–52. doi: 10.1108/S2050-206020150000010002.

Van Dijk, J. A. G. M. (2005) *The deepening divide: inequality in the information society*, Communication. Thousand Oaks CA, London, New Delhi, Sage Publications. doi: 10.4135/9781452229812.

Van Dijk, J. A. (2012) «The evolution of the digital divide: The digital divide turns to inequality of skills and usage», in *Digital Enlightenment Yearbook 2012*, pagg. 57–75. doi: 10.3233/978-1-61499-057-4-57.

Van Dijk, J. e Hacker, K. (2003) «The Digital Divide as a Complex and Dynamic Phenomenon», *Information Society*, pagg. 315–326. doi: 10.1080/01972240309487.

DiMaggio, P. e Hargittai, E. (2001) *From the «Digital Divide» to «Digital Inequality»: Studying Internet Use As Penetration Increases.*

Dolente, C. e Leporelli, C. (2014) «La domanda di connettività delle famiglie italiane», in *Prospettive di sviluppo delle reti a banda ultralarga in Italia. Dalle parole ai fatti?* Roma: Edizione Efesto, pagg. 97–150.

Dutton, W. H. e Blank, G. (2014) «The emergence of next generation internet users», *International Economics and Economic Policy*, 11(1–2), pagg. 29–47. doi: 10.1007/s10368-013-0245-8.

Eurostat (2017a) *Being young in Europe today - demographic trends - Statistics Explained.* Available at: http://ec.europa.eu/eurostat/statistics-explained/index.php/Being_young_in_Europe_today_-_demographic_trends (Consultato: 22 settembre 2017).

Eurostat (2017b) *Eurostat Database, European Commission; Online statistical database.* Available at: <http://ec.europa.eu/eurostat/data/database> (Consultato: 15 settembre 2017).

Falck, O., Heimisch, A. e Wiederhold, S. (2016) «Returns to ICT Skills», *OECD Education Working Papers*, (134), pagg. 1–62. doi: DOI:

Fischer, G. H. E. e Molenaar, I. W. E. (2012) *Rasch models: Foundations, recent developments, and applications, Rasch Models Foundations Recent Developments and Applications.* doi: 10.1007/978-1-4612-4230-7.

Flamm, K. e Chaudhuri, A. (2007) «An analysis of the determinants of broadband access», *Telecommunications Policy*, 31(6–7), pagg. 312–326. doi: 10.1016/j.telpol.2007.05.006.

Guo, C. e Ryoo, H. S. (2012) «Compact MILP models for optimal and Pareto-optimal LAD patterns», *Discrete Applied Mathematics*, 160(16–17), pagg. 2339–2348. doi: 10.1016/j.dam.2012.05.006.

H. Stock, J. e W. Watson, M. (2010) *Introduction to Econometrics (3rd Edition) (Addison-Wesley Series in Economics)*, Addison-Wesley.

Hammer, P. L., Kogan, A., Simeone, B. e Szedmák, S. (2004) «Pareto-optimal patterns in

logical analysis of data», *Discrete Applied Mathematics*, 144(1–2), pagg. 79–102. doi: 10.1016/j.dam.2003.08.013.

Hastie, T., Tibshirani, R., Friedman, J. e Franklin, J. (2005) *The elements of statistical learning: data mining, inference and prediction*, *The Mathematical Intelligencer*. doi: 10.1007/BF02985802.

Heckman, J. J. (1978) «Dummy endogenous variables in a simultaneous equation system», *Econometrica*, 46(4), pagg. 931–959.

Helpman, E. (1998) *General purpose technologies and economic growth*. MIT Press.

Helpman, E. e Trajtenberg, M. (1996) «Diffusion of General Purpose Technologies», *NBER Working Paper Series*, pagg. 1–46. doi: 10.3386/w5773.

Hitt, L. e Tambe, P. (2007) «Broadband adoption and content consumption», *Information Economics and Policy*, 19(3–4), pagg. 362–378. doi: 10.1016/j.infoecopol.2007.04.003.

Hosmer, D. W. e Lemeshow, S. (2004) «Applied Logistic Regression», in *John Wiley and Sons*, pag. 373.

Istat e FUB (2014) *Internet@Italia 2013. La popolazione Italiana e l'uso di Internet*. Roma.

Istat e FUB (2015) *Internet@Italia 2014. L'uso di Internet da parte di cittadini e imprese*. Roma.

ITU, (International Telecommunication Union) (2016) *Measuring the Information Society Report 2016*, ITU: Geneva. doi: 10.3359/oz0303157.

Kahneman, D. e Tversky, A. (1979) «Prospect Theory: An Analysis of Decision under Risk», *Econometrica*, 47(2), pagg. 263–292. doi: 10.1111/j.1536-7150.2011.00774.x.

Kass, G. V. (1980) «An Exploratory Technique for Investigating Large Quantities of Categorical Data», *Applied Statistics*, 29(2), pag. 119. doi: 10.2307/2986296.

Katz, R. L. (2012) *Impact of Broadband on the Economy*, ITU *Universe of Broadband*. Ginevra.

Köksal, E. e Anil, B. (2015) «The Determinants of Broadband Access and Usage in Turkey : Do Regions Matter ?», *Topics in Middle Eastern and African Economies*, 17(1), pagg. 113–132. doi: 10.2139/ssrn.2362376.

Lanza, S. T. e Collins, L. M. (2010) *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. New York: Wiley. doi: 10.1002/9780470567333.

Lazarsfeld, P. F. (1950) «The logical and mathematical foundation of latent structure

analysis», in Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. F., Star, S. A., e Clausen, J. A. (a c. di) *Measurement and prediction*. Princeton, NJ, US: Princeton University Press.

LECG (2009) *Economic Impact of Broadband: An Empirical Study*.

Lera-López, F., Billon, M. e Gil, M. (2011) «Determinants of internet use in Spain», *Economics of Innovation and New Technology*, 20(2), pagg. 127–152. doi: 10.1080/10438590903378017.

Li, S., Xu, L. Da e Zhao, S. (2015) «The internet of things: a survey», *Information Systems Frontiers*. Elsevier B.V., 17(2), pagg. 243–259. doi: 10.1007/s10796-014-9492-7.

Linzer, D. A. e Lewis, J. B. (2011) «poLCA: An R Package for Polytomous Variable Latent Class Analysis», *Journal of Statistical Software*, 42(10), pagg. 1–29. doi: 10.18637/jss.v042.i10.

Loh, W.-Y. e Shih, Y.-S. (1997) «Split Selection Methods for Classification Trees», *Statistica Sinica*, 7(4), pagg. 815–840. doi: 10.2307/24306157.

Loh, W. Y. e Vanichsetakul, N. (1988) «Tree-structured classification via generalized discriminant analysis», *Journal of the American Statistical Association*, 83(403), pagg. 715–725. doi: 10.1080/01621459.1988.10478652.

Lyons, S. (2014) «Timing and determinants of local residential broadband adoption: Evidence from Ireland», *Empirical Economics*, 47(4), pagg. 1341–1363. doi: 10.1007/s00181-013-0790-6.

Madden, G. e Simpson, M. (1997) «Residential broadband subscription demand: An econometric analysis of Australian choice experiment data», *Applied Economics*, 29, pagg. 1073–1078. doi: 10.1080/000368497326462.

Manski, C. F. e McFadden, D. (1981) *Structural analysis of discrete data with econometric applications*, *Structural analysis of discrete data with econometric applications*. doi: 10.1111/j.1463-1326.2008.front_matter.x.

McFadden, D. (1974) «Conditional logit analysis of qualitative choice behavior», *Frontiers in Econometrics*, 1(2), pagg. 105–142. doi: 10.1108/eb028592.

McLachlan, G. (1992) «Cluster analysis and related techniques in medical research», *Statistical Methods in Medical Research*, 1(1), pagg. 27–48. doi: 10.1177/096228029200100103.

Menard, S. (1995) *Applied logistic regression analysis*, *Sage university paper series on quantitative*

applications in the social sciences.

Minges, M. (2016) «Exploring the Relationship Between Broadband and Economic Growth», *World Development Report*, 1.

Morgan, J. N. e Sonquist, J. A. (1963) «Problems in the Analysis of Survey Data, and a Proposal», *Journal of the American Statistical Association*, 58(302), pagg. 415–434. doi: 10.1080/01621459.1963.10500855.

Nerlove, M. (1978) «Econometric analysis of longitudinal data: approaches, problems and prospects», *Annales de l'insee*, (30/31), pagg. 6–22.

Nisbet, E. C., Stoycheff, E. e Pearce, K. E. (2012) «Internet Use and Democratic Demands: A Multinational, Multilevel Model of Internet Use and Citizen Attitudes About Democracy», *Journal of Communication*, 62(2), pagg. 249–265. doi: 10.1111/j.1460-2466.2012.01627.x.

Oberski, D. (2016) «Mixture models: Latent profile and latent class analysis», *Modern Statistical Methods for HCI*, pagg. 275–287. doi: 10.1007/978-3-319-26633-6_12.

OECD (2016) *PISA 2015 Results in Focus*. Parigi.

Pélissié Du Rausas, M., Manyika, J., Hazan, E., Bughin, J., Chui, M. e Said, R. (2011) *Internet matters : The Net's sweeping impact on growth, jobs, and prosperity*. McKinsey Global Institute.

Presidenza del Consiglio dei Ministri (2015) *Strategia italiana per la banda ultralarga*.

Prieger, J. E. e Hu, W. M. (2008) «The broadband digital divide and the nexus of race, competition, and quality», *Information Economics and Policy*, 20(2), pagg. 150–167. doi: 10.1016/j.infoecopol.2008.01.001.

Quaglione, D., Agovino, M., Di Berardino, C. e Sarra, A. (2017) «Exploring additional determinants of fixed broadband adoption: policy implications for narrowing the broadband demand gap», *Economics of Innovation and New Technology*. Taylor & Francis, 0(0), pagg. 1–21. doi: 10.1080/10438599.2017.1350358.

Quinlan, J. R. (1992) *C4.5: Programs for Machine Learning*, Morgan Kaufmann San Mateo California. doi: 10.1016/S0019-9958(62)90649-6.

Rasch, G. (1961) «Probabilistic Models for Some Intelligence and Attainment Tests», *Information and Control*, 4(4), pag. 382. doi: 10.1016/S0019-9958(61)80061-2.

- Read, D. (2007) «Experienced utility: Utility theory from Jeremy Bentham to Daniel Kahneman», *Thinking and Reasoning*, 13(1), pagg. 45–61. doi: 10.1080/13546780600872627.
- Van Reenen, J., Bloom, N., Draca, M., Kretschmer, T., Sadun, R., Overman, H., Schankerman, M. e M. (2010) *The Economic Impact of ICT, publication*. SMART N. 2007/0020.
- Robbins, L. (1932) *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.
- Robinson, L., Cotten, S. R., Ono, H., Quan-Haase, A., Mesch, G., Chen, W., Schulz, J., Hale, T. M. e Stern, M. J. (2015) «Digital inequalities and why they matter», *Information, Communication & Society*. Routledge, 18(5), pagg. 569–582. doi: 10.1080/1369118X.2015.1012532.
- Rose, K., Eldridge, S. e Chapin, L. (2015) «The Internet of Things: An Overview Understanding the Issues and Challenges of a More Connected World».
- Roycroft, T. R. (2013) «Empirical study of broadband adoption using data from the 2009 Residential Energy Consumption Survey», *Journal of Regulatory Economics*, 43(2), pagg. 214–228. doi: 10.1007/s11149-012-9207-2.
- Ryoo, H. S. e Jang, I. Y. (2009) «MILP approach to pattern generation in logical analysis of data», *Discrete Applied Mathematics*, 157(4), pagg. 749–761. doi: 10.1016/j.dam.2008.07.005.
- Sadowski, B. M. (2017) «Advanced users and the adoption of high speed broadband: Results of a living lab study in the Netherlands», *Technological Forecasting & Social Change*, 115(Complete), pagg. 1–14. doi: 10.1016/j.techfore.2016.09.009.
- Savage, S. J. e Waldman, D. (2005) «Broadband Internet access, awareness, and use: Analysis of United States household data», *Telecommunications Policy*, 29(8), pagg. 615–633. doi: 10.1016/j.telpol.2005.06.001.
- Savage, S. J. e Waldman, D. M. (2009) «Ability, location and household demand for Internet bandwidth», *International Journal of Industrial Organization*. Elsevier B.V., 27(2), pagg. 166–174. doi: 10.1016/j.ijindorg.2008.06.001.
- Schwarz, G. (1978) «Estimating the Dimension of a Model», *The Annals of Statistics*, 6(2), pagg. 461–464. doi: 10.1214/aos/1176344136.
- Sclove, S. L. (1987) «Application of model-selection criteria to some problems in multivariate

analysis», *Psychometrika*, 52(3), pagg. 333–343. doi: 10.1007/BF02294360.

Simon, H. A. (1955) «A Behavioral Model of Rational Choice», *The Quarterly Journal of Economics*, 69(1), pagg. 99–118. doi: 10.2307/1884852.

Simon, H. A. (1959) «Theories of Decision-Making in Economics and Behavioral Science», *The American Economic Review*, 49(3), pagg. 253–283.

Solow, R. M. (1956) «A Contribution to the Theory of Economic Growth», *Source: The Quarterly Journal of Economics*. The MIT Press, 70(1), pagg. 65–94.

Srinuan, C. e Bohlin, E. (2013) «Analysis of fixed broadband access and use in Thailand: Drivers and barriers», *Telecommunications Policy*. Elsevier, 37(8), pagg. 615–625. doi: 10.1016/j.telpol.2013.03.006.

Train, K. E. (2009) «Discrete choice methods with simulation», *Cambridge University Press, New York*, (2nd), pag. 1148. doi: 10.1016/S0898-1221(04)90100-9.

Train, K. E., McFadden, D. L. e Ben-Akiva, M. (1987) «The Demand for Local Telephone Service: A Fully Discrete Model of Residential Calling Patterns and Service Choices», *The RAND Journal of Economics*, 18(1), pagg. 109–123. doi: 10.2307/2555538.

Tsoumakas, G., Katakis, I. e Vlahavas, I. (2010) *Data Mining and Knowledge Discovery Handbook, Journal of Chemical Information and Modeling*. doi: 10.1017/CBO9781107415324.004.

Weiner, M. D., Puniello, O. T., Noland, R. B., Ciemnecki, D. e Turakhia, C. (2012) «Consider the non-adopter: Developing a prediction model for the adoption of household-level broadband access», *Socio-Economic Planning Sciences*. Elsevier Ltd, 46(3), pagg. 183–193. doi: 10.1016/j.seps.2012.04.001.

Whitacre, B. e Rhinesmith, C. (2016) «Broadband un-adopters», *Telecommunications Policy*. Elsevier, 40(1), pagg. 1–13. doi: 10.1016/j.telpol.2015.11.008.

Indice delle figure

| | |
|---|----|
| Figura 1 - Quota di famiglie (con almeno un componente tra 16 e 74 anni) con accesso ad Internet a banda larga fissa, per paese di residenza (anno 2016) - Fonte: Eurostat..... | 23 |
| Figura 2 - Quota di famiglie (con almeno un componente tra 16 e 74 anni) con accesso ad Internet a banda larga mobile, per paese di residenza (anno 2016) - Fonte: Eurostat | 24 |
| Figura 3 – Quota totale di famiglie (con almeno un componente tra 16 e 74 anni) con accesso ad Internet a banda larga, per paese di residenza (anno 2016) - Fonte: Eurostat | 25 |
| Figura 4 - Quota di famiglie (con almeno un componente tra 16 e 74 anni) con accesso ad Internet a banda larga solo mobile, per paese di residenza (anno 2016) - Fonte: Eurostat. | 26 |
| Figura 5 – Quota di individui (tra 16 e 74 anni) che utilizzano Internet, per frequenza di utilizzo e paese di residenza (anno 2016). Fonte: Eurostat | 27 |
| Figura 6 - Struttura per età della popolazione italiana (anni 2005-2014). Fonte: elaborazione su dati microdati Istat AVQ | 40 |
| Figura 7 - Indice di vecchiaia della popolazione italiana (anni 2005-2014). Fonte: elaborazione su dati microdati Istat AVQ | 41 |
| Figura 8- Popolazione italiana, fuori dal circuito scolastico, per titolo di studio (anno 2014). Fonte: elaborazione su dati microdati Istat AVQ | 42 |
| Figura 9 - Popolazione italiana di 15 anni o più per condizione professionale (anno 2014). Fonte: elaborazione su dati microdati Istat AVQ | 43 |
| Figura 10 - Quota di utenti Internet (almeno una volta a settimana) per classe di età, in Italia (anno 2014). Fonte: elaborazione su dati microdati Istat AVQ..... | 44 |
| Figura 11 – Quota di utenti Internet per livello di istruzione (persone fuori dal percorso scolastico, anno 2014). Fonte: elaborazione su microdati Istat AVQ..... | 45 |
| Figura 12 – Quota di utenti Internet in Italia per condizione professionale (persone di 15 anni o più, anno 2014). Fonte: elaborazione su microdati Istat AVQ | 45 |
| Figura 13 – Quota di utenti Internet in Italia per posizione professionale (occupati, anno 2014). Fonte: elaborazione su microdati Istat AVQ..... | 46 |

| | |
|--|-----|
| Figura 14 – Quota di utenti Internet (almeno una volta a settimana) in Italia (popolazione di 6 anni o più, anni 2005-2014). Fonte: elaborazione su microdati Istat AVQ..... | 47 |
| Figura 15 - Utenza Internet per diversi segmenti significativi della popolazione (anni 2005-2014). Fonte: elaborazione su microdati Istat AVQ..... | 49 |
| Figura 16 - Diffusione ICT tra le famiglie italiane (anni 2005-2014). Fonte: elaborazione su dati Istat AVQ..... | 51 |
| Figura 17 - Diffusione delle diverse tecnologie di connessione Internet tra le famiglie italiane (anno 2016). Fonte: Eurostat | 52 |
| Figura 18 - Famiglie italiane per tipologia di connessione ad Internet e classe di età della Persona di Riferimento (anno 2014). Fonte: elaborazione su microdati Istat AVQ..... | 53 |
| Figura 19 - Famiglie italiane per tipologia di connessione ad Internet e titolo di studio della Persona di Riferimento (anno 2014). Fonte: elaborazione su microdati Istat AVQ..... | 54 |
| Figura 20 - Famiglie italiane per tipologia di connessione ad Internet e condizione professionale della Persona di Riferimento (anno 2014). Fonte: elaborazione su microdati Istat AVQ..... | 55 |
| Figura 21 - Famiglie italiane per tipologia di connessione ad Internet e condizione professionale della Persona di Riferimento (solo famiglie con PR occupata, anno 2014). Fonte: elaborazione su microdati Istat AVQ..... | 56 |
| Figura 22 - Quota di famiglie raggiunte da banda larga fissa (anno 2016). Fonte: Commissione Europea, Digital Agenda Scoreboard | 103 |
| Figura 23 – Quota di famiglie raggiunte da tecnologie mobili ad alta velocità (almeno HSPA), per paese (anno 2016). Fonte: Commissione Europea, Digital Agenda Scoreboard..... | 104 |
| Figura 24 -Rappresentazione del digital divide in quattro stadi | 106 |
| Figura 25 - Modello di adozione ipotizzato | 109 |
| Figura 26 – Albero risultante dall'applicazione dell'algorithmo QUEST (anno 2012) | 127 |
| Figura 27 - Evoluzione del tasso di utenza di Internet in ogni nodo della partizione (anni 2005-2012)..... | 130 |

Indice delle tabelle

| | |
|--|----|
| Tabella 1 - Principali variabili di segmentazione della rilevazione "Aspetti della vita quotidiana" di Istat, e relative modalità | 38 |
| Tabella 2 - Principali variabili relative all'uso di Internet e alla connessione, contenute nella rilevazione "Aspetti della vita quotidiana" di Istat, e relative modalità | 39 |
| Tabella 3 - Utenza Internet per diversi segmenti significativi della popolazione (anni 2005-2014). Fonte: elaborazione su microdati Istat AVQ..... | 48 |
| Tabella 4 - Crescita media annua dell'utenza regolare di Internet in Italia, in punti percentuali, per segmenti significativi della popolazione (anni 2005-2014). Fonte: elaborazione su microdati Istat AVQ..... | 49 |
| Tabella 5 - Quota di utenti di Internet (almeno una volta a settimana) all'interno di ciascun segmento individuato dalle quattro variabili principali (anno 2012). Fonte: elaborazione su microdati Istat AVQ..... | 50 |
| Tabella 6 - Motivi della non adozione di una connessione Internet da parte delle famiglie (percentuale di tutte le famiglie che non adottano) (anno 2015). Fonte: Eurostat..... | 57 |
| Tabella 7 - Analisi dei motivi della non adozione di una connessione Internet a casa da parte delle famiglie (anno 2015). Fonte: Eurostat..... | 59 |
| Tabella 8 – Persone che segnalano le motivazioni principali del non uso di Internet da parte degli individui (anno 2014). Fonte: elaborazione su microdati Istat AVQ..... | 61 |
| Tabella 9 - Suddivisione per classi di età delle motivazioni principali al non uso di Internet da parte degli individui (anno 2014). Fonte: elaborazione su microdati Istat AVQ | 61 |
| Tabella 10 - Suddivisione per titolo di studio delle motivazioni principali al non uso di Internet da parte degli individui (anno 2014). Fonte: elaborazione su microdati Istat AVQ | 62 |
| Tabella 11 - Suddivisione per condizione e posizione professionale delle motivazioni principali al non uso di Internet da parte degli individui (anno 2014). Fonte: elaborazione su microdati Istat AVQ..... | 64 |

| | |
|---|-----|
| Tabella 12 – Valori di probabilità dell’uscita, all’interno della regressione logistica, secondo le variazioni di una sola variabile indipendente dicotomica..... | 77 |
| Tabella 13 - Riassunto dei risultati nei diversi test della metodologia SLAD | 119 |
| Tabella 14 - Pattern positivi prodotti nel test 3 dalla Procedura 1 della metodologia SLAD (anno 2012) | 121 |
| Tabella 15 - Pattern negativi prodotti nel test 3 dalla Procedura 1 della metodologia SLAD (anno 2012) | 122 |
| Tabella 16 - Pattern positivi prodotti nel test 4 dalla Procedura 2 della metodologia SLAD (anno 2012) | 124 |
| Tabella 17 - Pattern negativi prodotti nel test 4 dalla Procedura 2 della metodologia SLAD (anno 2012) | 125 |
| Tabella 18 - Partizione risultante dall’aggregazione dei nodi dell’albero (anno 2012) | 128 |
| Tabella 19 - Classi latenti di utenti per frequenza d'uso di Internet (anno 2014) | 135 |
| Tabella 20 - Probabilità stimata di utilizzo dell'applicazione per classe di appartenenza (anno 2014) | 136 |
| Tabella 21 - Numero di attività svolte per gruppo di attività, per ciascuna delle classi latenti individuate (anno 2014)..... | 138 |
| Tabella 22 - Caratteristiche socio-demografiche delle classi latenti di utenti (anno 2014). | 139 |
| Tabella 23 - Probabilità di adozione di una qualsiasi connessione a banda larga nel caso base, in percentuale, per cluster di utenza potenziale e situazione economica (anno 2012)..... | 143 |
| Tabella 24 - Numero di famiglie (in migliaia) adottanti una qualsiasi connessione a banda larga nel caso base, per cluster di utenza potenziale e situazione economica (anno 2012) | 143 |
| Tabella 25 - Probabilità incrementali di adozione di banda larga per lo scenario 1, in punti percentuali, per cluster di utenza potenziale e situazione economica (anno 2012) | 145 |
| Tabella 26 - Probabilità incrementali di adozione di banda larga per lo scenario 2, in punti percentuali, per cluster di utenza potenziale e situazione economica (anno 2012) | 145 |
| Tabella 27 - Probabilità incrementali di adozione di banda larga per lo scenario 3, in punti percentuali, per cluster di utenza potenziale e situazione economica (anno 2012) | 146 |

| | |
|--|-----|
| Tabella 28 - Probabilità incrementali di adozione di banda larga per lo scenario 4, in punti percentuali, per cluster di utenza potenziale e situazione economica (anno 2012)..... | 147 |
| Tabella 29 – Quota delle famiglie italiane all'interno di ciascuna classe sulla base del nodo di appartenenza dei componenti (anni 2005 – 2012)..... | 148 |
| Tabella 30 – Consistenza delle famiglie italiane, secondo la classificazione sulla base delle classi latenti (anno 2014)..... | 150 |
| Tabella 31 - Suddivisione delle famiglie per classe latente dei componenti e connessione ad Internet da casa (anno 2014)..... | 151 |
| Tabella 32 - Suddivisione delle famiglie nelle diverse ripartizioni geografiche di residenza, per classe latente dei componenti (anno 2014)..... | 152 |
| Tabella 33 – Suddivisione degli individui per fasce di età, sulle diverse categorie di famiglie per classi latenti dei componenti (anno 2014)..... | 153 |
| Tabella 34 – Suddivisione delle famiglie all'interno delle diverse categorie per classi latenti, per strumento di connessione ad Internet. Famiglie con connessione fissa a banda larga (anno 2014)..... | 154 |
| Tabella 35 – Suddivisione delle famiglie all'interno delle diverse categorie per classi latenti, per strumento di connessione ad Internet. Famiglie con sola connessione mobile a banda larga (anno 2014)..... | 154 |

Appendice

Tabella A.1 – Stima dei coefficienti di regressione logistica per l'uso di Internet

| | | B | Errore std | Wald | gl | Sig. | Exp(B) | Intervallo di confidenza al 95% per Exp(B) | |
|---------------------------------|---|----------------|------------|----------|----|-------|---------|--|------------------|
| | | | | | | | | Limite inferiore | Limite superiore |
| Intercetta | | -6,713 | ,068 | 9787,876 | 1 | 0,000 | | | |
| Sesso | <i>maschio</i> | ,574 | ,012 | 2443,474 | 1 | 0,000 | 1,775 | 1,735 | 1,816 |
| | <i>femmina</i> | 0 ^a | | | 0 | | | | |
| Classe di età | <i>6-7</i> | 2,563 | ,077 | 1102,494 | 1 | ,000 | 12,976 | 11,154 | 15,095 |
| | <i>8-10</i> | 3,727 | ,068 | 2987,771 | 1 | 0,000 | 41,560 | 36,361 | 47,503 |
| | <i>11-14</i> | 5,070 | ,065 | 6118,126 | 1 | 0,000 | 159,179 | 140,188 | 180,742 |
| | <i>15-17</i> | 3,664 | ,071 | 2655,881 | 1 | 0,000 | 39,014 | 33,940 | 44,848 |
| | <i>18-19</i> | 3,829 | ,072 | 2855,771 | 1 | 0,000 | 45,998 | 39,972 | 52,933 |
| | <i>20-24</i> | 3,544 | ,064 | 3041,446 | 1 | 0,000 | 34,595 | 30,501 | 39,238 |
| | <i>25-34</i> | 3,404 | ,061 | 3114,995 | 1 | 0,000 | 30,091 | 26,701 | 33,913 |
| | <i>35-44</i> | 2,902 | ,060 | 2312,602 | 1 | 0,000 | 18,202 | 16,172 | 20,487 |
| | <i>45-54</i> | 2,231 | ,060 | 1376,459 | 1 | ,000 | 9,308 | 8,273 | 10,472 |
| | <i>55-59</i> | 1,831 | ,060 | 918,067 | 1 | ,000 | 6,239 | 5,542 | 7,024 |
| | <i>60-64</i> | 1,542 | ,060 | 664,173 | 1 | ,000 | 4,674 | 4,156 | 5,255 |
| | <i>65-74</i> | 1,087 | ,060 | 330,168 | 1 | ,000 | 2,964 | 2,637 | 3,333 |
| <i>75+</i> | 0 ^a | | | 0 | | | | | |
| Titolo di studio | <i>Laurea</i> | 2,481 | ,033 | 5696,406 | 1 | 0,000 | 11,949 | 11,204 | 12,744 |
| | <i>Diploma</i> | 2,006 | ,029 | 4751,726 | 1 | 0,000 | 7,432 | 7,020 | 7,868 |
| | <i>Licenza media o Abilitazione professionale</i> | 1,039 | ,028 | 1423,006 | 1 | 0,000 | 2,827 | 2,679 | 2,984 |
| | <i>Licenza elementare o meno</i> | 0 ^a | | | 0 | | | | |
| Condizione Professionale | <i>Occupato</i> | ,691 | ,026 | 725,563 | 1 | ,000 | 1,995 | 1,898 | 2,098 |
| | <i>In cerca di occupazione</i> | ,513 | ,032 | 258,762 | 1 | ,000 | 1,670 | 1,569 | 1,778 |
| | <i>Casalinga</i> | -,418 | ,032 | 172,146 | 1 | ,000 | ,659 | ,619 | ,701 |
| | <i>Studente</i> | 1,635 | ,041 | 1605,085 | 1 | 0,000 | 5,129 | 4,735 | 5,556 |
| | <i>Ritirato dal lavoro, Inabile o altro</i> | 0 ^a | | | 0 | | | | |
| | <i>Minore di 15 anni</i> | 0 ^a | | | 0 | | | | |

| | | B | Errore std | Wald | gl | Sig. | Exp(B) | Intervallo di confidenza al 95% per Exp(B) | |
|-------------------------|---|----------------|------------|-----------|----|-------|--------|--|------------------|
| | | | | | | | | Limite inferiore | Limite superiore |
| | | | | | | | | | |
| Posizione professionale | <i>Dirigente o Quadro</i> | 1,223 | ,034 | 1275,888 | 1 | ,000 | 3,397 | 3,177 | 3,633 |
| | <i>Impiegato</i> | ,881 | ,025 | 1217,664 | 1 | ,000 | 2,412 | 2,296 | 2,535 |
| | <i>Operaio</i> | -,403 | ,025 | 258,344 | 1 | ,000 | ,668 | ,636 | ,702 |
| | <i>Imprenditore</i> | ,706 | ,041 | 298,413 | 1 | ,000 | 2,026 | 1,870 | 2,195 |
| | <i>Libero professionista</i> | 1,144 | ,037 | 957,668 | 1 | ,000 | 3,139 | 2,919 | 3,374 |
| | <i>Lavoratore autonomo, Coadiuvante, Socio cooperativa, Apprendista, Dipendente da casa, Precario</i> | ,075 | ,027 | 7,847 | 1 | ,005 | 1,078 | 1,023 | 1,136 |
| | <i>Senza posizione attuale o precedente</i> | 0 ^a | | | 0 | | | | |
| Connessione a casa | <i>DSL</i> | 2,421 | ,014 | 30329,129 | 1 | 0,000 | 11,262 | 10,959 | 11,573 |
| | <i>Altra BL (compreso 3G)</i> | 2,245 | ,022 | 10710,912 | 1 | 0,000 | 9,442 | 9,049 | 9,852 |
| | <i>Narrowband</i> | 1,946 | ,016 | 15022,125 | 1 | 0,000 | 7,000 | 6,785 | 7,221 |
| | <i>Nessuna</i> | 0 ^a | | | 0 | | | | |
| Anno | <i>2005</i> | -1,177 | ,023 | 2631,679 | 1 | 0,000 | ,308 | ,295 | ,322 |
| | <i>2006</i> | -1,017 | ,023 | 2009,430 | 1 | 0,000 | ,362 | ,346 | ,378 |
| | <i>2007</i> | -,973 | ,022 | 1987,896 | 1 | 0,000 | ,378 | ,362 | ,394 |
| | <i>2008</i> | -,734 | ,022 | 1128,578 | 1 | ,000 | ,480 | ,460 | ,501 |
| | <i>2009</i> | -,473 | ,022 | 478,159 | 1 | ,000 | ,623 | ,597 | ,650 |
| | <i>2010</i> | -,248 | ,021 | 141,160 | 1 | ,000 | ,780 | ,749 | ,813 |
| | <i>2011</i> | -,087 | ,021 | 17,371 | 1 | ,000 | ,917 | ,880 | ,955 |
| | <i>2012</i> | 0 ^a | | | 0 | | | | |
| Ripart. territoriale | <i>Nord Ovest</i> | ,523 | ,020 | 667,292 | 1 | ,000 | 1,686 | 1,621 | 1,755 |
| | <i>Nord Est</i> | ,529 | ,020 | 673,130 | 1 | ,000 | 1,697 | 1,631 | 1,767 |
| | <i>Centro</i> | ,333 | ,021 | 259,332 | 1 | ,000 | 1,395 | 1,339 | 1,452 |
| | <i>Sud</i> | -,145 | ,019 | 57,308 | 1 | ,000 | ,865 | ,833 | ,898 |
| | <i>Isole</i> | 0 ^a | | | 0 | | | | |
| Dimens | <i>comuni centro di area metropolitana</i> | ,013 | ,021 | ,362 | 1 | ,547 | 1,013 | ,972 | 1,055 |

| | B | Errore std | Wald | gl | Sig. | Exp(B) | Intervallo di confidenza al 95% per Exp(B) | |
|--|----------------|------------|--------|----|------|--------|---|------------------|
| | | | | | | | Limite inferiore | Limite superiore |
| | | | | | | | <i>comuni periferia dell'area metropolitana</i> | ,046 |
| <i>comuni fino a 2.000</i> | -,124 | ,023 | 29,206 | 1 | ,000 | ,883 | ,845 | ,924 |
| <i>comuni tra 2.001-10.000</i> | -,134 | ,017 | 65,609 | 1 | ,000 | ,875 | ,847 | ,904 |
| <i>comuni tra 10.001-50.000</i> | -,085 | ,016 | 27,357 | 1 | ,000 | ,918 | ,889 | ,948 |
| <i>comuni con oltre 50.000 abitanti</i> | 0 ^a | | | 0 | | | | |
| Pseudo R quadro (Nagelkerke): 0,649 | | | | | | | | |
| a. Questo parametro è impostato su zero perché è ridondante. | | | | | | | | |

Tabella A.2 - Stima dei parametri di regressione logistica ordinale per il giudizio delle risorse economiche

| | | Stima | Errore std | Wald | gl | Sign. | Intervallo di confidenza 95% | |
|---|---|----------------|------------|----------|----|-------|------------------------------|------------------|
| | | | | | | | Limite inferiore | Limite superiore |
| Giudizio sulle risorse economiche della famiglia | <i>ottime</i> | -8,182 | ,133 | 3785,459 | 1 | 0,000 | -8,442 | -7,921 |
| | <i>adeguate</i> | -2,131 | ,130 | 269,806 | 1 | ,000 | -2,386 | -1,877 |
| | <i>scarse</i> | 1,377 | ,129 | 113,448 | 1 | ,000 | 1,123 | 1,630 |
| Distanza dai servizi h/sett collaboratrice domestica | | ,023 | ,002 | 136,300 | 1 | ,000 | ,019 | ,027 |
| Consumi culturali | | -,008 | ,002 | 18,873 | 1 | ,000 | -,012 | -,005 |
| Anno | <i>2005</i> | -,001 | ,001 | 0,520 | 1 | ,471 | -,002 | ,001 |
| | <i>2006</i> | -,079 | ,025 | 9,929 | 1 | ,002 | -,128 | -,030 |
| | <i>2007</i> | -,064 | ,025 | 6,713 | 1 | ,010 | -,113 | -,016 |
| | <i>2008</i> | -,064 | ,025 | 6,614 | 1 | ,010 | -,112 | -,015 |
| | <i>2009</i> | ,188 | ,024 | 61,635 | 1 | ,000 | ,141 | ,236 |
| | <i>2010</i> | ,050 | ,024 | 4,180 | 1 | ,041 | ,002 | ,097 |
| | <i>2011</i> | -,001 | ,024 | 0,001 | 1 | ,981 | -,048 | ,047 |
| | <i>2012</i> | -,045 | ,024 | 3,391 | 1 | ,066 | -,092 | ,003 |
| | 0 ^a | | | | 0 | | | |
| Dimensione comune di residenza | <i>comuni centro di area metropolitana</i> | ,179 | ,024 | 56,339 | 1 | ,000 | ,132 | ,226 |
| | <i>comuni periferia dell'area metropolitana</i> | ,038 | ,026 | 2,085 | 1 | ,149 | -,013 | ,089 |
| | <i>comuni fino a 2.000</i> | -,089 | ,026 | 11,436 | 1 | ,001 | -,141 | -,037 |
| | <i>comuni tra 2.001-10.000</i> | -,055 | ,020 | 7,807 | 1 | ,005 | -,093 | -,016 |
| | <i>comuni tra 10.001-50.000</i> | -,004 | ,019 | 0,047 | 1 | ,829 | -,041 | ,033 |
| | <i>comuni con oltre 50.000 abitanti</i> | 0 ^a | | | 0 | | | |
| Ripartizione territoriale | <i>Nord Ovest</i> | ,031 | ,025 | 1,587 | 1 | ,208 | -,017 | ,079 |
| | <i>Nord Est</i> | ,145 | ,025 | 34,897 | 1 | ,000 | ,097 | ,194 |

| | | | | | | | | |
|---|---|--------|------|-----------|---|-------|--------|--------|
| | <i>Centro</i> | ,058 | ,024 | 5,684 | 1 | ,017 | ,010 | ,106 |
| | <i>Sud</i> | ,012 | ,022 | 0,318 | 1 | ,573 | -,031 | ,056 |
| | <i>Isole</i> | 0a | | | 0 | | | |
| Tipo di nucleo familiare | <i>Nessun nucleo</i> | -,147 | ,033 | 19,762 | 1 | ,000 | -,212 | -,082 |
| | <i>Coppia con figli</i> | -,043 | ,047 | 0,826 | 1 | ,364 | -,135 | ,050 |
| | <i>Coppia senza figli</i> | ,048 | ,048 | 0,975 | 1 | ,323 | -,047 | ,142 |
| | <i>Monogenitore padre</i> | -,084 | ,062 | 1,811 | 1 | ,178 | -,205 | ,038 |
| | <i>Monogenitore madre</i> | 0a | | | 0 | | | |
| Soddisfazione sulla situazione economica | <i>Molto</i> | -4,447 | ,059 | 5711,565 | 1 | 0,000 | -4,563 | -4,332 |
| | <i>Abbastanza</i> | -2,948 | ,024 | 15166,175 | 1 | 0,000 | -2,994 | -2,901 |
| | <i>Poco</i> | -1,364 | ,022 | 3855,628 | 1 | 0,000 | -1,408 | -1,321 |
| | <i>Per niente</i> | 0a | | | 0 | | | |
| Posizione Professionale della PR maschio | <i>Dirigente o Quadro</i> | -,378 | ,096 | 15,415 | 1 | ,000 | -,567 | -,189 |
| | <i>Impiegato</i> | -,133 | ,092 | 2,077 | 1 | ,150 | -,314 | ,048 |
| | <i>Operaio</i> | -,067 | ,091 | 0,533 | 1 | ,465 | -,246 | ,112 |
| | <i>Imprenditore</i> | -,237 | ,099 | 5,751 | 1 | ,016 | -,431 | -,043 |
| | <i>Libero professionista</i> | -,167 | ,097 | 2,925 | 1 | ,087 | -,358 | ,024 |
| | <i>Lavoratore autonomo, Coadiuvante, Socio cooperativa, Apprendista, Dipendente da casa, Precario</i> | -,074 | ,092 | 0,639 | 1 | ,424 | -,255 | ,107 |
| | <i>Senza posizione attuale o precedente</i> | -,294 | ,101 | 8,515 | 1 | ,004 | -,491 | -,097 |
| | <i>Nessuna Persona di Riferimento maschio</i> | 0a | | | 0 | | | |
| Posizione Professionale della PR femmina | <i>Dirigente o Quadro</i> | -,232 | ,058 | 16,227 | 1 | ,000 | -,345 | -,119 |
| | <i>Impiegato</i> | -,134 | ,040 | 11,409 | 1 | ,001 | -,211 | -,056 |
| | <i>Operaio</i> | -,065 | ,037 | 3,128 | 1 | ,077 | -,137 | ,007 |
| | <i>Imprenditore</i> | -,171 | ,070 | 5,948 | 1 | ,015 | -,309 | -,034 |
| | <i>Libero professionista</i> | -,154 | ,059 | 6,816 | 1 | ,009 | -,269 | -,038 |
| | <i>Lavoratore autonomo, Coadiuvante, Socio cooperativa, Apprendista, Dipendente da casa, Precario</i> | -,076 | ,038 | 3,926 | 1 | ,048 | -,152 | -,001 |
| | <i>Senza posizione attuale o precedente</i> | -,150 | ,038 | 15,845 | 1 | ,000 | -,224 | -,076 |
| | <i>Nessuna Persona di Riferimento femmina</i> | 0a | | | 0 | | | |
| | <i>Occupato</i> | -,139 | ,099 | 1,962 | 1 | ,161 | -,333 | ,055 |

| | | | | | | | | |
|---|---|-------|------|----------|---|-------|-------|-------|
| Condizione Professionale della PR maschio | <i>In cerca di occupazione o Studente</i> | ,354 | ,100 | 12,638 | 1 | ,000 | ,159 | ,549 |
| | <i>Ritirato dal lavoro, Inabile o altro</i> | ,008 | ,097 | 0,007 | 1 | ,934 | -,183 | ,199 |
| | <i>Nessuna Persona di Riferimento maschio</i> | 0a | | | 0 | | | |
| Condizione Professionale della PR femmina | <i>Occupato</i> | ,069 | ,052 | 1,803 | 1 | ,179 | -,032 | ,170 |
| | <i>In cerca di occupazione o Studente</i> | ,328 | ,057 | 32,622 | 1 | ,000 | ,215 | ,440 |
| | <i>Casalinga</i> | ,084 | ,049 | 2,897 | 1 | ,089 | -,013 | ,180 |
| | <i>Ritirato dal lavoro, Inabile o altro</i> | ,102 | ,052 | 3,784 | 1 | ,052 | -,001 | ,204 |
| | <i>Nessuna Persona di Riferimento femmina</i> | 0a | | | 0 | | | |
| Titolo di studio della PR maschio | <i>Laurea</i> | -,357 | ,039 | 84,064 | 1 | ,000 | -,433 | -,280 |
| | <i>Diploma</i> | -,098 | ,027 | 12,789 | 1 | ,000 | -,152 | -,044 |
| | <i>Licenza media o Abilitazione professionale</i> | -,041 | ,021 | 3,565 | 1 | ,059 | -,083 | ,002 |
| | <i>Licenza elementare o meno</i> | 0a | | | 0 | | | |
| | <i>Nessuna Persona di Riferimento maschio</i> | 0a | | | 0 | | | |
| Titolo di studio della PR femmina | <i>Laurea</i> | -,299 | ,035 | 72,087 | 1 | ,000 | -,368 | -,230 |
| | <i>Diploma</i> | -,153 | ,026 | 35,603 | 1 | ,000 | -,203 | -,103 |
| | <i>Licenza media o Abilitazione professionale</i> | -,048 | ,020 | 5,810 | 1 | ,016 | -,087 | -,009 |
| | <i>Licenza elementare o meno</i> | 0a | | | 0 | | | |
| | <i>Nessuna Persona di Riferimento femmina</i> | 0a | | | 0 | | | |
| Numero di vacanze (4 notti o più) | <i>0</i> | 0,932 | ,036 | 684,471 | 1 | ,000 | ,862 | 1,002 |
| | <i>1</i> | ,295 | ,034 | 75,275 | 1 | ,000 | ,228 | ,361 |
| | <i>2</i> | ,119 | ,038 | 9,643 | 1 | ,002 | ,044 | ,195 |
| | <i>3+</i> | 0a | | | 0 | | | |
| Mancata vacanza per motivi economici | <i>no</i> | -,744 | ,017 | 1951,311 | 1 | 0,000 | -,777 | -,711 |
| | <i>sì</i> | 0a | | | 0 | | | |
| Possesso di un'assicurazione sulla salute o infortuni | <i>no</i> | ,105 | ,019 | 31,251 | 1 | ,000 | ,068 | ,141 |
| | <i>sì</i> | 0a | | | 0 | | | |
| Possesso di un'assicurazione sulla vita o per pensione integrativa | <i>no</i> | ,100 | ,019 | 28,585 | 1 | ,000 | ,063 | ,136 |
| | <i>sì</i> | 0a | | | 0 | | | |
| | <i>no</i> | ,088 | ,015 | 34,307 | 1 | ,000 | ,058 | ,117 |

| | | | | | | | | |
|---|---|-------|------|----------|---|------|-------|-------|
| Possesso lavastoviglie | <i>sì</i> | 0a | | | 0 | | | |
| Possesso antenna parabolica | <i>no</i> | ,080 | ,014 | 30,667 | 1 | ,000 | ,052 | ,108 |
| | <i>sì</i> | 0a | | | 0 | | | |
| Possesso di condizionatori | <i>no</i> | ,108 | ,016 | 47,803 | 1 | ,000 | ,078 | ,139 |
| | <i>sì</i> | 0a | | | 0 | | | |
| Numero di possessori di bancomat | 0 | -,089 | ,064 | 1,955 | 1 | ,162 | -,213 | ,036 |
| | 1 | -,130 | ,062 | 4,303 | 1 | ,038 | -,252 | -,007 |
| | 2 | -,109 | ,061 | 3,128 | 1 | ,077 | -,229 | ,012 |
| | 3 | -,153 | ,063 | 5,851 | 1 | ,016 | -,277 | -,029 |
| | 4+ | 0a | | | 0 | | | |
| Numero di possessori di carte di credito | 0 | ,213 | ,059 | 12,927 | 1 | ,000 | ,097 | ,330 |
| | 1 | ,116 | ,059 | 3,834 | 1 | ,050 | ,000 | ,232 |
| | 2 | ,086 | ,060 | 2,031 | 1 | ,154 | -,032 | ,204 |
| | 3+ | 0a | | | 0 | | | |
| Spese abitazione troppo alte | <i>no</i> | -,627 | ,013 | 2244,199 | 1 | ,000 | -,653 | -,601 |
| | <i>sì</i> | 0a | | | 0 | | | |
| Abitazione troppo piccola | <i>no</i> | -,281 | ,020 | 205,429 | 1 | ,000 | -,319 | -,243 |
| | <i>sì</i> | 0a | | | 0 | | | |
| Titolo di godimento dell'abitazione | <i>affitto, subaffitto</i> | ,005 | ,026 | 0,041 | 1 | ,840 | -,045 | ,056 |
| | <i>proprietà</i> | -,227 | ,023 | 100,387 | 1 | ,000 | -,271 | -,182 |
| | <i>usufrutto, titolo gratuito o altro</i> | 0a | | | 0 | | | |
| Assicurazione contro i furti nell'abitazione | <i>no</i> | ,053 | ,022 | 5,700 | 1 | ,017 | ,010 | ,097 |
| | <i>sì</i> | 0a | | | 0 | | | |
| Tipo di abitazione | <i>villa o villino</i> | -,378 | ,040 | 91,306 | 1 | ,000 | -,455 | -,300 |
| | <i>signorile</i> | -,458 | ,041 | 127,684 | 1 | ,000 | -,537 | -,378 |
| | <i>civile</i> | -,252 | ,031 | 67,426 | 1 | ,000 | -,312 | -,192 |
| | <i>economico popolare</i> | -,012 | ,033 | 0,137 | 1 | ,712 | -,077 | ,052 |
| | <i>rurale o abitazione impropria</i> | 0a | | | 0 | | | |
| Numero di auto | 0 | ,322 | ,039 | 68,055 | 1 | ,000 | ,245 | ,398 |
| | 1 | ,159 | ,034 | 22,362 | 1 | ,000 | ,093 | ,225 |
| | 2 | ,035 | ,032 | 1,177 | 1 | ,278 | -,028 | ,097 |
| | 3+ | 0a | | | 0 | | | |
| Numero di stanze che compongono l'abitazione | 1 | ,360 | ,063 | 32,653 | 1 | ,000 | ,237 | ,484 |
| | 2 | ,379 | ,035 | 114,314 | 1 | ,000 | ,309 | ,448 |
| | 3 | ,246 | ,029 | 73,350 | 1 | ,000 | ,190 | ,302 |
| | 4 | ,149 | ,027 | 31,270 | 1 | ,000 | ,097 | ,201 |
| | 5 | ,114 | ,027 | 18,114 | 1 | ,000 | ,062 | ,167 |
| | 6 | ,076 | ,031 | 6,098 | 1 | ,014 | ,016 | ,136 |
| | 7+ | 0a | | | 0 | | | |
| Numero di percettori di reddito | 0 | -,088 | ,072 | 1,484 | 1 | ,223 | -,230 | ,054 |
| | 1 | -,033 | ,055 | 0,352 | 1 | ,553 | -,141 | ,076 |
| | 2 | -,035 | ,051 | ,473 | 1 | ,492 | -,135 | ,065 |
| | 3 | -,050 | ,051 | 0,961 | 1 | ,327 | -,150 | ,050 |
| | 4+ | 0a | | | 0 | | | |
| Numero di mantenuti | 0 | -,714 | ,048 | 217,238 | 1 | ,000 | -,809 | -,619 |
| | 1 | -,534 | ,042 | 164,338 | 1 | ,000 | -,616 | -,452 |
| | 2 | -,347 | ,040 | 76,757 | 1 | ,000 | -,424 | -,269 |
| | 3 | -,205 | ,040 | 26,088 | 1 | ,000 | -,284 | -,127 |
| | 4+ | 0a | | | 0 | | | |
| Età della PR in classi | 18-34 | -,026 | ,032 | 0,688 | 1 | ,407 | -,089 | ,036 |
| | 35-54 | ,078 | ,025 | 9,382 | 1 | ,002 | ,028 | ,128 |
| | 55-64 | ,105 | ,021 | 25,096 | 1 | ,000 | ,064 | ,147 |

| | | | | | | | |
|--|-----|----|--|--|---|--|--|
| | 65+ | 0a | | | 0 | | |
| Pseudo R quadro (Nagelkerke): 0,411 | | | | | | | |
| a. Questo parametro è impostato su zero perché è ridondante. | | | | | | | |

Tabella A.3 - Stima dei parametri di regressione logistica ordinale per la soddisfazione economica

| | | Stima | Errore std | Wald | gl | Sign. | Intervallo di confidenza 95% | |
|---|---|----------------|------------|-----------|----|-------|------------------------------|------------------|
| | | | | | | | Limite inferiore | Limite superiore |
| Soglia soddisfazione economica | <i>Molto</i> | -6,467 | ,121 | 2838,262 | 1 | 0,000 | -6,705 | -6,229 |
| | <i>Abbastanza</i> | -1,969 | ,120 | 268,132 | 1 | ,000 | -2,205 | -1,734 |
| | <i>Poco</i> | 1,095 | ,120 | 83,310 | 1 | ,000 | ,860 | 1,330 |
| Distanza dai servizi | | ,016 | ,002 | 71,814 | 1 | ,000 | ,012 | ,019 |
| h/sett collaboratrice domestica | | -,013 | ,002 | 53,126 | 1 | ,000 | -,016 | -,009 |
| Consumi culturali | | -,005 | ,001 | 31,583 | 1 | ,000 | -,006 | -,003 |
| Anno | <i>2005</i> | -,193 | ,023 | 68,345 | 1 | ,000 | -,239 | -,147 |
| | <i>2006</i> | -,243 | ,023 | 110,481 | 1 | ,000 | -,288 | -,197 |
| | <i>2007</i> | -,309 | ,023 | 179,324 | 1 | ,000 | -,355 | -,264 |
| | <i>2008</i> | -,053 | ,023 | 5,471 | 1 | ,019 | -,097 | -,009 |
| | <i>2009</i> | -,122 | ,023 | 29,188 | 1 | ,000 | -,167 | -,078 |
| | <i>2010</i> | -,157 | ,023 | 48,735 | 1 | ,000 | -,202 | -,113 |
| | <i>2011</i> | -,131 | ,023 | 33,664 | 1 | ,000 | -,175 | -,087 |
| | <i>2012</i> | 0 ^a | | | 0 | | | |
| Dimensione comuni | <i>comuni centro di area metropolitana</i> | ,118 | ,022 | 27,898 | 1 | ,000 | ,074 | ,162 |
| | <i>comuni periferia dell'area metropolitana</i> | ,067 | ,024 | 7,655 | 1 | ,006 | ,020 | ,115 |
| | <i>comuni fino a 2.000</i> | -,269 | ,025 | 119,878 | 1 | ,000 | -,317 | -,221 |
| | <i>comuni tra 2.001-10.000</i> | -,109 | ,018 | 35,746 | 1 | ,000 | -,145 | -,073 |
| | <i>comuni tra 10.001-50.000</i> | -,066 | ,018 | 13,854 | 1 | ,000 | -,101 | -,031 |
| | <i>comuni con oltre 50.000 abitanti</i> | 0 ^a | | | 0 | | | |
| Ripartizione territoriale | <i>Nord Ovest</i> | -,399 | ,023 | 304,629 | 1 | ,000 | -,444 | -,354 |
| | <i>Nord Est</i> | -,602 | ,023 | 684,544 | 1 | ,000 | -,647 | -,557 |
| | <i>Centro</i> | -,348 | ,023 | 232,269 | 1 | ,000 | -,392 | -,303 |
| | <i>Sud</i> | -,248 | ,021 | 143,886 | 1 | ,000 | -,288 | -,207 |
| | <i>Isole</i> | 0 ^a | | | 0 | | | |
| Tipo di nucleo familiare | <i>Nessun nucleo</i> | ,106 | ,031 | 11,630 | 1 | ,001 | ,045 | ,167 |
| | <i>Coppia con figli</i> | -,278 | ,044 | 39,344 | 1 | ,000 | -,364 | -,191 |
| | <i>Coppia senza figli</i> | -,461 | ,045 | 103,746 | 1 | ,000 | -,550 | -,373 |
| | <i>Monogenitore padre</i> | -,066 | ,058 | 1,316 | 1 | ,251 | -,180 | ,047 |
| | <i>Monogenitore madre</i> | 0 ^a | | | 0 | | | |
| Giudizio sulle risorse economiche della famiglia | <i>ottime</i> | -4,966 | ,067 | 5427,725 | 1 | 0,000 | -5,098 | -4,834 |
| | <i>adeguate</i> | -3,067 | ,027 | 13209,582 | 1 | 0,000 | -3,119 | -3,015 |
| | <i>scarse</i> | -1,396 | ,025 | 3152,576 | 1 | 0,000 | -1,445 | -1,347 |
| | <i>assolutamente insufficienti</i> | 0 ^a | | | 0 | | | |
| | <i>Dirigente o Quadro</i> | -,085 | ,091 | ,880 | 1 | ,348 | -,262 | ,093 |

| | Stima | Errore std | Wald | gl | Sign. | Intervallo di confidenza 95% | | |
|--|---|----------------|------|---------|-------|------------------------------|------------------|-------|
| | | | | | | Limite inferiore | Limite superiore | |
| Posizione Professionale della PR maschio | <i>Impiegato</i> | ,073 | ,088 | ,695 | 1 | ,404 | -,099 | ,244 |
| | <i>Operaio</i> | ,153 | ,087 | 3,108 | 1 | ,078 | -,017 | ,323 |
| | <i>Imprenditore</i> | ,035 | ,093 | ,138 | 1 | ,710 | -,148 | ,217 |
| | <i>Libero professionista</i> | ,077 | ,092 | ,697 | 1 | ,404 | -,103 | ,256 |
| | <i>Lavoratore autonomo, Coadiuvante, Socio cooperativa, Apprendista, Dipendente da casa, Precario</i> | ,200 | ,088 | 5,227 | 1 | ,022 | ,029 | ,372 |
| | <i>Senza posizione attuale o precedente</i> | -,025 | ,095 | ,070 | 1 | ,791 | -,212 | ,162 |
| | <i>Nessuna Persona di Riferimento maschio</i> | 0 ^a | | | 0 | | | |
| | | | | | | | | |
| Posizione Professionale della PR femmina | <i>Dirigente o Quadro</i> | -,282 | ,052 | 29,283 | 1 | ,000 | -,384 | -,180 |
| | <i>Impiegato</i> | -,181 | ,037 | 23,628 | 1 | ,000 | -,255 | -,108 |
| | <i>Operaio</i> | -,113 | ,035 | 10,444 | 1 | ,001 | -,181 | -,044 |
| | <i>Imprenditore</i> | -,107 | ,064 | 2,794 | 1 | ,095 | -,233 | ,019 |
| | <i>Libero professionista</i> | ,005 | ,054 | ,008 | 1 | ,927 | -,101 | ,110 |
| | <i>Lavoratore autonomo, Coadiuvante, Socio cooperativa, Apprendista, Dipendente da casa, Precario</i> | -,118 | ,037 | 10,527 | 1 | ,001 | -,190 | -,047 |
| | <i>Senza posizione attuale o precedente</i> | -,166 | ,036 | 21,491 | 1 | ,000 | -,236 | -,096 |
| | <i>Nessuna Persona di Riferimento femmina</i> | 0 ^a | | | 0 | | | |
| Condizione Professionale della PR maschio | <i>Occupato</i> | ,078 | ,094 | ,688 | 1 | ,407 | -,106 | ,262 |
| | <i>In cerca di occupazione o Studente</i> | ,504 | ,095 | 28,426 | 1 | ,000 | ,319 | ,689 |
| | <i>Ritirato dal lavoro, Inabile o altro</i> | ,162 | ,092 | 3,068 | 1 | ,080 | -,019 | ,343 |
| | <i>Nessuna Persona di Riferimento maschio</i> | 0 ^a | | | 0 | | | |
| Condizione Professionale della PR femmina | <i>Occupato</i> | ,346 | ,049 | 50,406 | 1 | ,000 | ,251 | ,442 |
| | <i>In cerca di occupazione o Studente</i> | ,595 | ,054 | 119,412 | 1 | ,000 | ,488 | ,702 |
| | <i>Casalanga</i> | ,320 | ,047 | 47,207 | 1 | ,000 | ,229 | ,412 |

| | Stima | Errore std | Wald | gl | Sign. | Intervallo di confidenza 95% | | |
|---|---|----------------|------|----------|-------|------------------------------|------------------|-------|
| | | | | | | Limite inferiore | Limite superiore | |
| | <i>Ritirato dal lavoro, Inabile o altro</i> | ,323 | ,050 | 42,492 | 1 | ,000 | ,226 | ,420 |
| | <i>Nessuna Persona di Riferimento femmina</i> | 0 ^a | | | 0 | | | |
| TITOLO DI STUDIO DELLA PR MASCHIO | <i>Laurea</i> | -,185 | ,035 | 27,931 | 1 | ,000 | -,253 | -,116 |
| | <i>Diploma</i> | ,002 | ,026 | ,006 | 1 | ,938 | -,048 | ,052 |
| | <i>Licenza media o Abilitazione professionale</i> | -,007 | ,020 | ,108 | 1 | ,742 | -,046 | ,033 |
| | <i>Licenza elementare o meno</i> | 0 ^a | | | 0 | | | |
| | <i>Nessuna Persona di Riferimento maschio</i> | 0 ^a | | | 0 | | | |
| TITOLO DI STUDIO DELLA PR FEMMINA | <i>Laurea</i> | -,109 | ,032 | 11,442 | 1 | ,001 | -,172 | -,046 |
| | <i>Diploma</i> | -,090 | ,024 | 14,086 | 1 | ,000 | -,137 | -,043 |
| | <i>Licenza media o Abilitazione professionale</i> | -,031 | ,019 | 2,667 | 1 | ,102 | -,068 | ,006 |
| | <i>Licenza elementare o meno</i> | 0 ^a | | | 0 | | | |
| | <i>Nessuna Persona di Riferimento femmina</i> | 0 ^a | | | 0 | | | |
| NUMERO DI VACANZE (4 NOTTI O PIÙ) | <i>0</i> | 1,031 | ,032 | 1047,780 | 1 | ,000 | ,968 | 1,093 |
| | <i>1</i> | ,326 | ,030 | 118,765 | 1 | ,000 | ,267 | ,384 |
| | <i>2</i> | ,094 | ,034 | 7,789 | 1 | ,005 | ,028 | ,160 |
| | <i>3+</i> | 0 ^a | | | 0 | | | |
| MANCATA VACANZA PER MOTIVI ECONOMICI | <i>no</i> | -,739 | ,016 | 2122,856 | 1 | 0,000 | -,770 | -,707 |
| | <i>sì</i> | 0 ^a | | | 0 | | | |
| POSSESSO DI UN'ASSICURAZIONE SULLA SALUTE O INFORTUNI | <i>no</i> | ,075 | ,017 | 19,417 | 1 | ,000 | ,042 | ,109 |
| | <i>sì</i> | 0 ^a | | | 0 | | | |
| POSSESSO DI UN'ASSICURAZIONE SULLA VITA O PER PENSIONE INTEGRATIVA | <i>no</i> | ,085 | ,017 | 25,154 | 1 | ,000 | ,052 | ,119 |
| | <i>sì</i> | 0 ^a | | | 0 | | | |
| POSSESSO LAVASTOVIGLIE | <i>no</i> | ,037 | ,014 | 6,889 | 1 | ,009 | ,009 | ,064 |
| | <i>sì</i> | 0 ^a | | | 0 | | | |
| POSSESSO ANTENNA PARABOLICA | <i>no</i> | ,074 | ,013 | 31,034 | 1 | ,000 | ,048 | ,100 |
| | <i>sì</i> | 0 ^a | | | 0 | | | |
| POSSESSO DI CONDIZIONATORI | <i>no</i> | -,004 | ,014 | ,091 | 1 | ,763 | -,033 | ,024 |
| | <i>sì</i> | 0 ^a | | | 0 | | | |
| NUMERO DI POSSESSORI DI BANCOMAT | <i>0</i> | ,119 | ,058 | 4,155 | 1 | ,042 | ,005 | ,233 |
| | <i>1</i> | ,037 | ,057 | ,410 | 1 | ,522 | -,076 | ,149 |
| | <i>2</i> | -,032 | ,056 | ,317 | 1 | ,574 | -,142 | ,079 |
| | <i>3</i> | ,134 | ,057 | 5,476 | 1 | ,019 | ,022 | ,247 |

| | | Stima | Errore std | Wald | gl | Sign. | Intervallo di confidenza 95% | |
|--|---|----------------|------------|----------|----|-------|------------------------------|------------------|
| | | | | | | | Limite inferiore | Limite superiore |
| | 4+ | 0 ^a | | | 0 | | | |
| Numero di possessori di carte di credito | 0 | ,140 | ,052 | 7,139 | 1 | ,008 | ,037 | ,242 |
| | 1 | ,063 | ,052 | 1,487 | 1 | ,223 | -,038 | ,165 |
| | 2 | -,003 | ,053 | ,003 | 1 | ,958 | -,106 | ,101 |
| | 3+ | 0 ^a | | | 0 | | | |
| Spese abitazione troppo alte | no | -,412 | ,012 | 1130,877 | 1 | ,000 | -,436 | -,388 |
| | sì | 0 ^a | | | 0 | | | |
| Abitazione troppo piccola | no | -,195 | ,019 | 108,974 | 1 | ,000 | -,231 | -,158 |
| | sì | 0 ^a | | | 0 | | | |
| Titolo di godimento dell'abitazione | <i>affitto, subaffitto</i> | ,127 | ,025 | 26,681 | 1 | ,000 | ,079 | ,175 |
| | <i>proprietà</i> | -,081 | ,021 | 14,421 | 1 | ,000 | -,123 | -,039 |
| | <i>usufrutto, titolo gratuito o altro</i> | 0 ^a | | | 0 | | | |
| Assicurazione contro i furti nell'abitazione | no | ,039 | ,020 | 3,822 | 1 | ,051 | ,000 | ,079 |
| | sì | 0 ^a | | | 0 | | | |
| Tipo di abitazione | <i>villa o villino</i> | -,211 | ,037 | 32,737 | 1 | ,000 | -,284 | -,139 |
| | <i>signorile</i> | -,299 | ,038 | 63,437 | 1 | ,000 | -,373 | -,226 |
| | <i>civile</i> | -,110 | ,029 | 14,110 | 1 | ,000 | -,168 | -,053 |
| | <i>economico popolare</i> | ,051 | ,032 | 2,585 | 1 | ,108 | -,011 | ,112 |
| | <i>rurale o abitazione impropria</i> | 0 ^a | | | 0 | | | |
| Numero di auto | 0 | ,002 | ,036 | ,004 | 1 | ,950 | -,068 | ,073 |
| | 1 | -,011 | ,031 | ,133 | 1 | ,715 | -,071 | ,049 |
| | 2 | -,001 | ,029 | ,002 | 1 | ,968 | -,058 | ,056 |
| | 3+ | 0 ^a | | | 0 | | | |
| Numero di stanze che compongono l'abitazione | 1 | ,051 | ,059 | ,726 | 1 | ,394 | -,066 | ,167 |
| | 2 | ,126 | ,033 | 14,636 | 1 | ,000 | ,062 | ,191 |
| | 3 | ,050 | ,026 | 3,609 | 1 | ,057 | -,002 | ,102 |
| | 4 | ,028 | ,024 | 1,307 | 1 | ,253 | -,020 | ,075 |
| | 5 | ,033 | ,024 | 1,829 | 1 | ,176 | -,015 | ,081 |
| | 6 | -,005 | ,028 | ,027 | 1 | ,870 | -,059 | ,050 |
| | 7+ | 0 ^a | | | 0 | | | |
| Numero di percettori di reddito | 0 | ,237 | ,068 | 12,124 | 1 | ,000 | ,104 | ,370 |
| | 1 | ,061 | ,051 | 1,408 | 1 | ,235 | -,040 | ,162 |
| | 2 | -,015 | ,047 | ,102 | 1 | ,750 | -,108 | ,078 |
| | 3 | ,117 | ,047 | 6,231 | 1 | ,013 | ,025 | ,209 |
| | 4+ | 0 ^a | | | 0 | | | |
| Numero di mantenuti | 0 | ,030 | ,046 | ,410 | 1 | ,522 | -,061 | ,120 |
| | 1 | ,167 | ,040 | 17,441 | 1 | ,000 | ,088 | ,245 |
| | 2 | ,109 | ,038 | 8,145 | 1 | ,004 | ,034 | ,183 |
| | 3 | ,008 | ,039 | ,047 | 1 | ,829 | -,068 | ,084 |
| | 4+ | 0 ^a | | | 0 | | | |
| Età della PR in classi | 18-34 | ,033 | ,030 | 1,257 | 1 | ,262 | -,025 | ,091 |
| | 35-54 | ,115 | ,024 | 23,582 | 1 | ,000 | ,069 | ,162 |
| | 55-64 | ,180 | ,020 | 83,317 | 1 | ,000 | ,141 | ,218 |
| | 65+ | 0 ^a | | | 0 | | | |

Pseudo R quadro (Nagelkerke): 0,411
a. Questo parametro è impostato su zero perché è ridondante.

Tabella A.4 - Stima dei parametri della regressione logistica per la connessione da casa della famiglia

DSL (rispetto alla categoria base, nessuna connessione)

| | | B | Errore std | Wald | gl | Sig. | Exp(B) | Intervallo di confidenza al 95% per Exp(B) | |
|-----------------------------------|--|---------|------------|----------|----|-------|-----------|--|------------------|
| | | | | | | | | Limite inferiore | Limite superiore |
| Intercetta | | -16,095 | 1,158 | 193,112 | 1 | 0 | | | |
| Situazione economica (completa) | | 10,667 | 1,269 | 70,621 | 1 | 0 | 42916,874 | 3565,891 | 82267,857 |
| Situazione economica (fino a 0,1) | | -8,199 | 1,359 | 36,422 | 1 | 0 | 0 | 1,92E-05 | 0,004 |
| Situazione economica (fino a 0,3) | | -2,052 | 0,448 | 20,948 | 1 | 0 | 0,128 | 0,053 | 0,309 |
| Situazione economica (fino a 0,5) | | -1,008 | 0,391 | 6,647 | 1 | 0,01 | 0,365 | 0,17 | 0,785 |
| Situazione economica (fino a 0,7) | | 3,555 | 0,375 | 90,016 | 1 | 0 | 34,99 | 16,788 | 72,928 |
| Situazione economica (fino a 0,9) | | 6,425 | 1,357 | 22,413 | 1 | 0 | 616,998 | 43,163 | 8819,725 |
| Anno | 2005 | -1,637 | 0,042 | 1521,995 | 1 | 0 | 0,195 | 0,179 | 0,211 |
| | 2006 | -1,443 | 0,04 | 1323,838 | 1 | 0 | 0,236 | 0,219 | 0,255 |
| | 2007 | -0,949 | 0,037 | 641,106 | 1 | 0 | 0,387 | 0,36 | 0,417 |
| | 2008 | -0,521 | 0,035 | 216,395 | 1 | 0 | 0,594 | 0,554 | 0,637 |
| | 2009 | -0,3 | 0,035 | 74,239 | 1 | 0 | 0,741 | 0,692 | 0,793 |
| | 2010 | -0,209 | 0,034 | 37,321 | 1 | 0 | 0,811 | 0,758 | 0,867 |
| | 2011 | -0,246 | 0,035 | 50,654 | 1 | 0 | 0,782 | 0,731 | 0,837 |
| | 2012 | 0b | | | 0 | | | | |
| Ripartizione territoriale | Nord Ovest | -0,086 | 0,035 | 6,164 | 1 | 0,013 | 0,918 | 0,857 | 0,982 |
| | Nord Est | 0,001 | 0,035 | 0 | 1 | 0,987 | 1,001 | 0,934 | 1,072 |
| | Centro | 0,124 | 0,035 | 12,257 | 1 | 0 | 1,132 | 1,056 | 1,213 |
| | Sud | -0,102 | 0,034 | 9,269 | 1 | 0,002 | 0,903 | 0,845 | 0,964 |
| | Isole | 0b | | | 0 | | | | |
| Dimensione comuni | comuni centro di area metropolitana | 0,303 | 0,033 | 82,653 | 1 | 0 | 1,355 | 1,269 | 1,446 |
| | comuni periferia dell'area metropolitana | 0,015 | 0,037 | 0,156 | 1 | 0,693 | 1,015 | 0,944 | 1,091 |
| | comuni fino a 2.000 | -0,928 | 0,041 | 507,374 | 1 | 0 | 0,395 | 0,365 | 0,429 |
| | comuni tra 2.001-10.000 | -0,371 | 0,028 | 178,594 | 1 | 0 | 0,69 | 0,654 | 0,729 |

| | B | Errore std | Wald | gl | Sig. | Exp(B) | Intervallo di confidenza al 95% per Exp(B) | | |
|----------------------------|---|------------|-------|----------|------|--------|--|------------------|--------|
| | | | | | | | Limite inferiore | Limite superiore | |
| | <i>comuni tra 10.001-50.000</i> | -0,227 | 0,027 | 70,796 | 1 | 0 | 0,797 | 0,756 | 0,84 |
| | <i>comuni con oltre 50.000 abitanti</i> | 0b | | | 0 | | | | |
| Età della PR | 18-34 | -0,368 | 0,065 | 32,255 | 1 | 0 | 0,692 | 0,61 | 0,786 |
| | 35-54 | -0,047 | 0,048 | 0,987 | 1 | 0,321 | 0,954 | 0,868 | 1,047 |
| | 55-64 | 0,089 | 0,04 | 5,075 | 1 | 0,024 | 1,093 | 1,012 | 1,181 |
| | 65+ | 0b | | | 0 | | | | |
| Numero componenti famiglia | 1 | 3,672 | 0,114 | 1039,842 | 1 | 0 | 39,349 | 31,477 | 49,189 |
| | 2 | 3,161 | 0,103 | 934,204 | 1 | 0 | 23,6 | 19,269 | 28,903 |
| | 3 | 2,7 | 0,093 | 841,33 | 1 | 0 | 14,877 | 12,396 | 17,855 |
| | 4 | 2,133 | 0,088 | 584,602 | 1 | 0 | 8,443 | 7,102 | 10,037 |
| | 5 | 1,331 | 0,091 | 212,006 | 1 | 0 | 3,785 | 3,164 | 4,528 |
| | 6-11 | 0b | | | 0 | | | | |
| Età minima in famiglia | 0-5 | 3,661 | 0,185 | 391,496 | 1 | 0 | 38,899 | 27,067 | 55,903 |
| | 6-7 | 3,358 | 0,187 | 322,118 | 1 | 0 | 28,739 | 19,916 | 41,472 |
| | 8-10 | 3,107 | 0,184 | 284,503 | 1 | 0 | 22,361 | 15,584 | 32,085 |
| | 11-14 | 2,949 | 0,182 | 261,948 | 1 | 0 | 19,087 | 13,355 | 27,279 |
| | 15-17 | 3,136 | 0,182 | 297,971 | 1 | 0 | 23,022 | 16,124 | 32,87 |
| | 18-19 | 3,265 | 0,182 | 321,983 | 1 | 0 | 26,187 | 18,331 | 37,409 |
| | 20-24 | 3,222 | 0,173 | 346,598 | 1 | 0 | 25,083 | 17,867 | 35,214 |
| | 25-34 | 3,032 | 0,167 | 331,295 | 1 | 0 | 20,74 | 14,963 | 28,748 |
| | 35-44 | 2,554 | 0,159 | 257,255 | 1 | 0 | 12,856 | 9,41 | 17,564 |
| | 45-54 | 2,234 | 0,154 | 209,654 | 1 | 0 | 9,334 | 6,898 | 12,629 |
| | 55-59 | 1,907 | 0,154 | 152,449 | 1 | 0 | 6,735 | 4,976 | 9,117 |
| | 60-64 | 1,64 | 0,15 | 120,025 | 1 | 0 | 5,153 | 3,843 | 6,91 |
| | 65-74 | 1,063 | 0,136 | 61,193 | 1 | 0 | 2,894 | 2,218 | 3,777 |
| 75+ | 0b | | | 0 | | | | | |
| Età media della famiglia | 18-19 | -0,781 | 0,179 | 18,968 | 1 | 0 | 0,458 | 0,322 | 0,651 |
| | 20-24 | -0,94 | 0,17 | 30,416 | 1 | 0 | 0,391 | 0,28 | 0,546 |
| | 25-34 | -0,715 | 0,164 | 19,121 | 1 | 0 | 0,489 | 0,355 | 0,674 |
| | 35-44 | -0,267 | 0,154 | 2,988 | 1 | 0,084 | 0,766 | 0,566 | 1,036 |
| | 45-54 | 0,02 | 0,143 | 0,019 | 1 | 0,892 | 1,02 | 0,77 | 1,351 |
| | 55-59 | 0,096 | 0,138 | 0,482 | 1 | 0,488 | 1,101 | 0,839 | 1,444 |

| | B | Errore std | Wald | gl | Sig. | Exp(B) | Intervallo di confidenza al 95% per Exp(B) | |
|--|-------|------------|----------|----|-------|--------|--|------------------|
| | | | | | | | Limite inferiore | Limite superiore |
| 60-64 | 0,03 | 0,133 | 0,05 | 1 | 0,824 | 1,03 | 0,794 | 1,336 |
| 65-74 | 0,092 | 0,119 | 0,591 | 1 | 0,442 | 1,096 | 0,868 | 1,384 |
| 75+ | 0b | | | 0 | | | | |
| Risorse economiche ottime o adeguate * somma probabilità di utilizzo degli occupati | 1,747 | 0,03 | 3361,433 | 1 | 0 | 5,738 | 5,409 | 6,087 |
| Risorse economiche scarse o inadeguate * somma probabilità di utilizzo degli occupati | 1,981 | 0,031 | 4089,293 | 1 | 0 | 7,253 | 6,826 | 7,707 |
| Risorse economiche ottime o adeguate * somma probabilità di utilizzo dei disoccupati | 1,851 | 0,055 | 1137,741 | 1 | 0 | 6,364 | 5,715 | 7,087 |
| Risorse economiche scarse o inadeguate * somma probabilità di utilizzo dei disoccupati | 2,009 | 0,047 | 1832,437 | 1 | 0 | 7,453 | 6,798 | 8,171 |
| Risorse economiche ottime o adeguate * somma probabilità di utilizzo degli studenti | 2,169 | 0,041 | 2823,467 | 1 | 0 | 8,749 | 8,077 | 9,478 |
| Risorse economiche scarse o inadeguate * somma probabilità di utilizzo degli studenti | 2,159 | 0,039 | 3004,915 | 1 | 0 | 8,667 | 8,023 | 9,362 |
| Risorse economiche ottime o adeguate * somma probabilità di utilizzo degli inattivi | 2,384 | 0,055 | 1888,95 | 1 | 0 | 10,852 | 9,746 | 12,084 |
| Risorse economiche scarse o inadeguate * somma probabilità di utilizzo degli inattivi | 2,523 | 0,067 | 1408,678 | 1 | 0 | 12,469 | 10,93 | 14,225 |

b. Questo parametro è impostato su zero perché è ridondante.

Altro tipo di banda larga (rispetto alla categoria base, nessuna connessione)

| | | B | Errore std | Wald | gl | Sig. | Exp(B) | Intervallo di confidenza al 95% per Exp(B) | |
|-----------------------------------|--|---------|------------|----------|----|-------|----------|--|------------------|
| | | | | | | | | Limite inferiore | Limite superiore |
| Intercetta | | -17,015 | 1,71 | 99,008 | 1 | 0 | | | |
| Situazione economica (completa) | | 10,274 | 1,866 | 30,328 | 1 | 0 | 28962,46 | 747,935 | 1121520,12 |
| Situazione economica (fino a 0,1) | | -9,885 | 2 | 24,438 | 1 | 0 | 5,09E-05 | 1,01E-06 | 0,003 |
| Situazione economica (fino a 0,3) | | -1,045 | 0,688 | 2,309 | 1 | 0,129 | 0,352 | 0,091 | 1,354 |
| Situazione economica (fino a 0,5) | | -0,733 | 0,615 | 1,421 | 1 | 0,233 | 0,48 | 0,144 | 1,604 |
| Situazione economica (fino a 0,7) | | 2,775 | 0,59 | 22,13 | 1 | 0 | 16,043 | 5,048 | 50,987 |
| Situazione economica (fino a 0,9) | | 7 | 2,007 | 12,17 | 1 | 0 | 1096,988 | 21,485 | 56010,21 |
| Anno | 2005 | -4,368 | 0,144 | 919,951 | 1 | 0 | 0,013 | 0,01 | 0,017 |
| | 2006 | -4,169 | 0,128 | 1063,788 | 1 | 0 | 0,015 | 0,012 | 0,02 |
| | 2007 | -2,736 | 0,073 | 1397,194 | 1 | 0 | 0,065 | 0,056 | 0,075 |
| | 2008 | -2,899 | 0,081 | 1288,305 | 1 | 0 | 0,055 | 0,047 | 0,065 |
| | 2009 | -2,061 | 0,06 | 1187,054 | 1 | 0 | 0,127 | 0,113 | 0,143 |
| | 2010 | -0,61 | 0,042 | 212,081 | 1 | 0 | 0,543 | 0,5 | 0,59 |
| | 2011 | -0,224 | 0,04 | 31,881 | 1 | 0 | 0,799 | 0,739 | 0,864 |
| | 2012 | 0b | | | 0 | | | | |
| Ripartizione territoriale | Nord Ovest | -0,083 | 0,053 | 2,397 | 1 | 0,122 | 0,921 | 0,829 | 1,022 |
| | Nord Est | -0,099 | 0,054 | 3,313 | 1 | 0,069 | 0,906 | 0,815 | 1,008 |
| | Centro | -0,029 | 0,055 | 0,272 | 1 | 0,602 | 0,972 | 0,872 | 1,083 |
| | Sud | -0,182 | 0,052 | 12,351 | 1 | 0 | 0,833 | 0,753 | 0,923 |
| | Isole | 0b | | | 0 | | | | |
| Dimensione comuni | comuni centro di area metropolitana | 0,333 | 0,055 | 36,649 | 1 | 0 | 1,396 | 1,253 | 1,555 |
| | comuni periferia dell'area metropolitana | -0,118 | 0,063 | 3,544 | 1 | 0,06 | 0,888 | 0,785 | 1,005 |
| | comuni fino a 2.000 | 0,149 | 0,057 | 6,791 | 1 | 0,009 | 1,16 | 1,038 | 1,297 |
| | comuni tra 2.001-10.000 | 0,066 | 0,044 | 2,211 | 1 | 0,137 | 1,068 | 0,979 | 1,165 |

| | B | Errore std | Wald | gl | Sig. | Exp(B) | Intervallo di confidenza al 95% per Exp(B) | | |
|---|--------|------------|-------|---------|-------|--------|--|------------------|--------|
| | | | | | | | Limite inferiore | Limite superiore | |
| | | | | | | | | | |
| <i>comuni tra 10.001-50.000</i> | -0,036 | 0,044 | 0,674 | 1 | 0,412 | 0,964 | 0,884 | 1,052 | |
| <i>comuni con oltre 50.000 abitanti</i> | 0b | | | 0 | | | | | |
| Età della PR | 18-34 | 0,045 | 0,102 | 0,192 | 1 | 0,661 | 1,046 | 0,857 | 1,276 |
| | 35-54 | 0,098 | 0,079 | 1,558 | 1 | 0,212 | 1,103 | 0,946 | 1,287 |
| | 55-64 | 0,039 | 0,068 | 0,328 | 1 | 0,567 | 1,04 | 0,91 | 1,189 |
| | 65+ | 0b | | | 0 | | | | |
| Numero componenti | 1 | 3,272 | 0,18 | 332,195 | 1 | 0 | 26,354 | 18,537 | 37,465 |
| | 2 | 2,6 | 0,159 | 268,646 | 1 | 0 | 13,47 | 9,87 | 18,383 |
| | 3 | 2,052 | 0,139 | 217,975 | 1 | 0 | 7,787 | 5,93 | 10,225 |
| | 4 | 1,527 | 0,129 | 139,345 | 1 | 0 | 4,603 | 3,572 | 5,931 |
| | 5 | 0,903 | 0,133 | 46,111 | 1 | 0 | 2,466 | 1,9 | 3,2 |
| | 6-11 | 0b | | | 0 | | | | |
| Età minima in famiglia | 0-5 | 3,579 | 0,457 | 61,398 | 1 | 0 | 35,83 | 14,638 | 87,705 |
| | 6-7 | 3,215 | 0,458 | 49,191 | 1 | 0 | 24,894 | 10,138 | 61,127 |
| | 8-10 | 3,06 | 0,455 | 45,207 | 1 | 0 | 21,32 | 8,739 | 52,016 |
| | 11-14 | 2,899 | 0,453 | 40,986 | 1 | 0 | 18,152 | 7,473 | 44,09 |
| | 15-17 | 3,032 | 0,452 | 45,024 | 1 | 0 | 20,745 | 8,556 | 50,3 |
| | 18-19 | 3,231 | 0,452 | 51,204 | 1 | 0 | 25,303 | 10,443 | 61,306 |
| | 20-24 | 3,128 | 0,444 | 49,732 | 1 | 0 | 22,833 | 9,572 | 54,468 |
| | 25-34 | 3,046 | 0,438 | 48,45 | 1 | 0 | 21,029 | 8,919 | 49,579 |
| | 35-44 | 2,757 | 0,43 | 41,093 | 1 | 0 | 15,752 | 6,78 | 36,594 |
| | 45-54 | 2,109 | 0,426 | 24,54 | 1 | 0 | 8,238 | 3,577 | 18,973 |
| | 55-59 | 1,772 | 0,427 | 17,204 | 1 | 0 | 5,882 | 2,546 | 13,589 |
| | 60-64 | 1,347 | 0,423 | 10,13 | 1 | 0,001 | 3,846 | 1,678 | 8,816 |
| | 65-74 | 0,796 | 0,403 | 3,904 | 1 | 0,048 | 2,217 | 1,006 | 4,885 |
| 75+ | 0b | | | 0 | | | | | |
| Età media della famiglia | 18-19 | 0,679 | 0,425 | 2,543 | 1 | 0,111 | 1,971 | 0,856 | 4,538 |
| | 20-24 | 0,569 | 0,418 | 1,852 | 1 | 0,174 | 1,767 | 0,778 | 4,009 |
| | 25-34 | 0,669 | 0,412 | 2,637 | 1 | 0,104 | 1,952 | 0,871 | 4,374 |
| | 35-44 | 0,923 | 0,403 | 5,249 | 1 | 0,022 | 2,517 | 1,143 | 5,546 |
| | 45-54 | 1,066 | 0,392 | 7,398 | 1 | 0,007 | 2,905 | 1,347 | 6,264 |

| | B | Errore std | Wald | gl | Sig. | Exp(B) | Intervallo di confidenza al 95% per Exp(B) | |
|---|-------|------------|----------|----|-------|--------|--|------------------|
| | | | | | | | Limite inferiore | Limite superiore |
| 55-59 | 1,037 | 0,387 | 7,18 | 1 | 0,007 | 2,821 | 1,321 | 6,023 |
| 60-64 | 0,815 | 0,383 | 4,544 | 1 | 0,033 | 2,26 | 1,068 | 4,784 |
| 65-74 | 0,786 | 0,365 | 4,638 | 1 | 0,031 | 2,194 | 1,073 | 4,487 |
| 75+ | 0b | | | 0 | | | | |
| Risorse economiche ottime o adeguate * somma probabilità di utilizzo degli occupati | 1,485 | 0,047 | 1004,792 | 1 | 0 | 4,413 | 4,026 | 4,837 |
| Risorse economiche scarse o inadeguate * somma probabilità di utilizzo degli occupati | 1,658 | 0,047 | 1229,765 | 1 | 0 | 5,249 | 4,785 | 5,759 |
| Risorse economiche ottime o adeguate * somma probabilità di utilizzo dei disoccupati | 1,51 | 0,078 | 375,345 | 1 | 0 | 4,527 | 3,886 | 5,274 |
| Risorse economiche scarse o inadeguate * somma probabilità di utilizzo dei disoccupati | 1,642 | 0,068 | 591,346 | 1 | 0 | 5,164 | 4,524 | 5,895 |
| Risorse economiche ottime o adeguate * somma probabilità di utilizzo degli studenti | 1,74 | 0,058 | 892,85 | 1 | 0 | 5,695 | 5,081 | 6,384 |
| Risorse economiche scarse o inadeguate * somma probabilità di utilizzo degli studenti | 1,739 | 0,058 | 902,724 | 1 | 0 | 5,693 | 5,082 | 6,377 |
| Risorse economiche ottime o adeguate * somma probabilità di utilizzo degli inattivi | 2,064 | 0,082 | 632,869 | 1 | 0 | 7,878 | 6,708 | 9,252 |
| Risorse economiche scarse o inadeguate * somma probabilità di utilizzo degli inattivi | 1,956 | 0,102 | 371,07 | 1 | 0 | 7,07 | 5,794 | 8,627 |

b. Questo parametro è impostato su zero perché è ridondante.

Narrowband (rispetto alla categoria base, nessuna connessione)

| | B | Errore std | Wald | gl | Sig. | Exp(B) | Intervallo di confidenza al 95% per Exp(B) | | |
|-----------------------------------|--|------------|---------|---------|-------|----------|--|------------------|-------|
| | | | | | | | Limite inferiore | Limite superiore | |
| Intercetta | -14,078 | 1,308 | 115,901 | 1 | 0 | | | | |
| Situazione economica (completa) | 6,758 | 1,432 | 22,284 | 1 | 0 | 861,329 | 52,06 | 14250,522 | |
| Situazione economica (fino a 0,1) | -11,619 | 1,913 | 36,882 | 1 | 0 | 8,99E-06 | 2,12E-07 | 0 | |
| Situazione economica (fino a 0,3) | -0,535 | 0,555 | 0,926 | 1 | 0,336 | 0,586 | 0,197 | 1,74 | |
| Situazione economica (fino a 0,5) | -0,74 | 0,468 | 2,503 | 1 | 0,114 | 0,477 | 0,191 | 1,193 | |
| Situazione economica (fino a 0,7) | 2,2 | 0,435 | 25,57 | 1 | 0 | 9,021 | 3,846 | 21,162 | |
| Situazione economica (fino a 0,9) | 3,518 | 1,533 | 5,267 | 1 | 0,022 | 33,711 | 1,671 | 680,014 | |
| Anno | 2005 | 1,229 | 0,053 | 531,657 | 1 | 0 | 3,418 | 3,079 | 3,795 |
| | 2006 | 1,058 | 0,053 | 397,266 | 1 | 0 | 2,88 | 2,595 | 3,196 |
| | 2007 | 1,04 | 0,053 | 379,486 | 1 | 0 | 2,829 | 2,548 | 3,141 |
| | 2008 | 0,787 | 0,054 | 211,71 | 1 | 0 | 2,197 | 1,976 | 2,443 |
| | 2009 | 0,486 | 0,056 | 75,33 | 1 | 0 | 1,626 | 1,457 | 1,815 |
| | 2010 | 0,189 | 0,059 | 10,48 | 1 | 0,001 | 1,209 | 1,078 | 1,355 |
| | 2011 | 0,157 | 0,059 | 7,071 | 1 | 0,008 | 1,17 | 1,042 | 1,314 |
| | 2012 | 0b | | | 0 | | | | |
| Ripartizione territoriale | Nord Ovest | -0,1 | 0,042 | 5,572 | 1 | 0,018 | 0,905 | 0,833 | 0,983 |
| | Nord Est | -0,023 | 0,042 | 0,282 | 1 | 0,595 | 0,978 | 0,9 | 1,062 |
| | Centro | 0,224 | 0,043 | 27,631 | 1 | 0 | 1,251 | 1,151 | 1,361 |
| | Sud | 0,132 | 0,04 | 10,853 | 1 | 0,001 | 1,141 | 1,055 | 1,234 |
| | Isole | 0b | | | 0 | | | | |
| Dimensione comuni | comuni centro di area metropolitana | 0,214 | 0,043 | 24,69 | 1 | 0 | 1,238 | 1,138 | 1,347 |
| | comuni periferia dell'area metropolitana | 0,077 | 0,046 | 2,712 | 1 | 0,1 | 1,08 | 0,986 | 1,183 |
| | comuni fino a 2.000 | 0,44 | 0,042 | 108,522 | 1 | 0 | 1,552 | 1,429 | 1,686 |
| | comuni tra 2.001-10.000 | 0,285 | 0,033 | 72,714 | 1 | 0 | 1,329 | 1,245 | 1,419 |
| | comuni tra 10.001-50.000 | -0,031 | 0,034 | 0,841 | 1 | 0,359 | 0,969 | 0,907 | 1,036 |

| | <i>comuni con oltre 50.000 abitanti</i> | 0b | | | 0 | | | | |
|-------------------------------|---|--------|-------|---------|---|-------|--------|--------|--------|
| Età della PR | 18-34 | -0,42 | 0,075 | 31,75 | 1 | 0 | 0,657 | 0,568 | 0,76 |
| | 35-54 | -0,098 | 0,054 | 3,363 | 1 | 0,067 | 0,906 | 0,816 | 1,007 |
| | 55-64 | 0,004 | 0,044 | 0,006 | 1 | 0,936 | 1,004 | 0,92 | 1,095 |
| | 65+ | 0b | | | 0 | | | | |
| Numero componenti famiglia | 1 | 2,723 | 0,125 | 475,282 | 1 | 0 | 15,224 | 11,918 | 19,446 |
| | 2 | 2,297 | 0,112 | 417,669 | 1 | 0 | 9,94 | 7,975 | 12,389 |
| | 3 | 2,059 | 0,1 | 425,6 | 1 | 0 | 7,838 | 6,446 | 9,532 |
| | 4 | 1,701 | 0,094 | 327,365 | 1 | 0 | 5,482 | 4,559 | 6,591 |
| | 5 | 0,995 | 0,097 | 104,178 | 1 | 0 | 2,704 | 2,234 | 3,273 |
| | 6-11 | 0b | | | 0 | | | | |
| Età minima in famiglia | 0-5 | 3,569 | 0,241 | 219,193 | 1 | 0 | 35,481 | 22,121 | 56,909 |
| | 6-7 | 3,348 | 0,243 | 189,609 | 1 | 0 | 28,435 | 17,657 | 45,792 |
| | 8-10 | 3,173 | 0,24 | 174,85 | 1 | 0 | 23,876 | 14,918 | 38,212 |
| | 11-14 | 3,067 | 0,238 | 166,552 | 1 | 0 | 21,481 | 13,482 | 34,225 |
| | 15-17 | 3,243 | 0,237 | 187,215 | 1 | 0 | 25,601 | 16,089 | 40,736 |
| | 18-19 | 3,336 | 0,237 | 198,34 | 1 | 0 | 28,103 | 17,666 | 44,707 |
| | 20-24 | 3,342 | 0,229 | 213,363 | 1 | 0 | 28,284 | 18,062 | 44,29 |
| | 25-34 | 3,21 | 0,222 | 208,493 | 1 | 0 | 24,768 | 16,021 | 38,291 |
| | 35-44 | 2,761 | 0,214 | 165,802 | 1 | 0 | 15,811 | 10,386 | 24,069 |
| | 45-54 | 2,355 | 0,21 | 126,087 | 1 | 0 | 10,535 | 6,984 | 15,89 |
| | 55-59 | 2,135 | 0,21 | 103,253 | 1 | 0 | 8,456 | 5,602 | 12,764 |
| | 60-64 | 1,813 | 0,205 | 78,216 | 1 | 0 | 6,127 | 4,1 | 9,156 |
| | 65-74 | 1,26 | 0,188 | 45,126 | 1 | 0 | 3,526 | 2,441 | 5,093 |
| 75+ | 0b | | | 0 | | | | | |
| Età media della famiglia | 18-19 | -0,811 | 0,223 | 13,24 | 1 | 0 | 0,444 | 0,287 | 0,688 |
| | 20-24 | -0,93 | 0,213 | 19,077 | 1 | 0 | 0,395 | 0,26 | 0,599 |
| | 25-34 | -0,773 | 0,205 | 14,152 | 1 | 0 | 0,462 | 0,309 | 0,691 |
| | 35-44 | -0,286 | 0,196 | 2,125 | 1 | 0,145 | 0,751 | 0,511 | 1,104 |
| | 45-54 | 0,009 | 0,186 | 0,003 | 1 | 0,96 | 1,009 | 0,701 | 1,453 |
| | 55-59 | 0,038 | 0,181 | 0,045 | 1 | 0,833 | 1,039 | 0,729 | 1,481 |
| | 60-64 | 0,035 | 0,176 | 0,04 | 1 | 0,842 | 1,036 | 0,734 | 1,461 |
| | 65-74 | 0,02 | 0,159 | 0,016 | 1 | 0,898 | 1,021 | 0,747 | 1,394 |

| | 75+ | 0b | | | 0 | | | | |
|--|-------|-------|----------|---|---|-------|-------|--------|--|
| Risorse economiche ottime o adeguate * somma probabilità di utilizzo degli occupati | 1,451 | 0,034 | 1840,229 | 1 | 0 | 4,266 | 3,993 | 4,558 | |
| Risorse economiche scarse o inadeguate * somma probabilità di utilizzo degli occupati | 1,638 | 0,036 | 2124,539 | 1 | 0 | 5,146 | 4,8 | 5,518 | |
| Risorse economiche ottime o adeguate * somma probabilità di utilizzo dei disoccupati | 1,521 | 0,064 | 560,521 | 1 | 0 | 4,576 | 4,035 | 5,19 | |
| Risorse economiche scarse o inadeguate * somma probabilità di utilizzo dei disoccupati | 1,72 | 0,057 | 897,837 | 1 | 0 | 5,585 | 4,991 | 6,25 | |
| Risorse economiche ottime o adeguate * somma probabilità di utilizzo degli studenti | 1,745 | 0,044 | 1579,915 | 1 | 0 | 5,727 | 5,255 | 6,242 | |
| Risorse economiche scarse o inadeguate * somma probabilità di utilizzo degli studenti | 1,857 | 0,043 | 1843,221 | 1 | 0 | 6,404 | 5,883 | 6,97 | |
| Risorse economiche ottime o adeguate * somma probabilità di utilizzo degli inattivi | 2,139 | 0,066 | 1062,204 | 1 | 0 | 8,487 | 7,463 | 9,652 | |
| Risorse economiche scarse o inadeguate * somma probabilità di utilizzo degli inattivi | 2,247 | 0,085 | 697,234 | 1 | 0 | 9,457 | 8,004 | 11,173 | |
| b. Questo parametro è impostato su zero perché è ridondante. | | | | | | | | | |
| Pseudo R quadro (Nagelkerke): 0,567 | | | | | | | | | |

Tabella A.5 - Variabili manifeste dell'analisi di classi latenti per gli utenti di Internet

| |
|---|
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: spedire o ricevere e-mail |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: telefonare via Internet o effettuare videochiamate via webcam |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: uso di servizi di messaggia istantanea |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: inviare messaggi su chat, blog, newsgroup o forum di discussione online |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: partecipare a social network (creare un profilo utente, postare messaggi o altro su Facebook, Twitter, ecc.) |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: esprimere opinioni su temi sociali o politici su web (ad esempio tramite blog, social network, ecc.) |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: partecipare a consultazioni o votazioni online su temi sociali (civici) o politici (pianificazione urbana, firmare una petizione) |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: consultare un wiki per ottenere informazioni (es. wikipedia, altre enciclopedie online) |

| |
|--|
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: partecipare ad un network professionale (es. creare un profilo, postare messaggi o altri contributi su LinkedIn, Xing, ecc.) |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: usare servizi relativi a viaggi o soggiorni |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: cercare lavoro o mandare una richiesta di lavoro |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: prendere un appuntamento con un medico |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: cercare informazioni sanitarie |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: usare servizi bancari |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: vendere merci o servizi |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: cercare informazioni su attività di istruzione o corsi |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: fare un corso online |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: cercare informazioni su merci e servizi |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: scaricare software (diversi da giochi) |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: leggere giornali, informazioni, riviste online |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: leggere o scaricato libri online o e-book |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: ascoltare la radio su web |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: guardare programmi televisivi sul web |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: guardare film in streaming |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: guardare video in streaming |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: scaricare immagini, film, musica |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: giocare o scaricare giochi |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: partecipare a giochi in rete con altri utenti |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: caricare contenuti di propria creazione (testi, fotografie, musica, video, ecc.) |
| Negli ultimi 3 mesi ha usato Internet per le seguenti attività: creare siti web o blog |
| Negli ultimi 3 mesi ha usato servizi di archiviazione su Internet per salvare documenti immagini o altri file |
| Negli ultimi 3 mesi ha usato servizi di archiviazione/condivisione su Internet per condividere documenti, immagini o altri file |
| Negli ultimi 12 mesi ha usato Internet per ottenere informazioni da siti web della P.A. |
| Negli ultimi 12 mesi ha usato Internet per scaricare moduli della P.A. |
| Negli ultimi 12 mesi ha usato Internet per spedire moduli compilati alla P.A. |
| Ha mai comprato o ordinato merci e/o servizi per uso privato su Internet |
| Frequenza di uso di Internet negli ultimi 12 mesi |
| Classi per uso internet da uso in situazioni diverse da casa, studio, lavoro negli ultimi 3 mesi |
| Possiede carta di credito? |
| Frequenza con cui negli ultimi 12 mesi è andato a cinema |
| Frequenza con cui negli ultimi 12 mesi è andato a teatro |
| Frequenza con cui negli ultimi 12 mesi è andato a musei o mostre |
| Frequenza con cui negli ultimi 12 mesi è andato a concerti di musica classica |
| Frequenza con cui negli ultimi 12 mesi è andato ad altri concerti di musica |
| Frequenza con cui negli ultimi 12 mesi è andato a spettacoli sportivi |
| Frequenza con cui negli ultimi 12 mesi è andato in discoteche, balere, night club o altri luoghi dove ballare |
| Frequenza con cui negli ultimi 12 mesi è andato in siti archeologici, monumenti |
| Quanti libri ha letto negli ultimi 12 mesi? (solo libri letti per motivi non strettamente scolastici o professionali) |
| Legge quotidiani almeno una volta alla settimana? |
| Frequenza uso Internet a casa negli ultimi 3 mesi |
| Frequenza uso Internet sul lavoro negli ultimi 3 mesi |
| Numero di attività Internet relativo al video |
| Indice di capacità informatiche (con modello di Rasch) |
| Cluster di skill digitali |
| Utente da mobile |
| Soddisfazione situazione economica negli ultimi 12 mesi |

Tabella A.6 - Variabili manifeste dell'analisi di classi latenti per i non utenti di Internet

| |
|---|
| Motivo non uso Internet: non conosco Internet/non so cosa sia |
| Motivo non uso Internet: non dispongo di alcuno strumento per connettermi |
| Motivo non uso Internet: la connessione a banda larga non è disponibile nella nostra zona |
| Motivo non uso Internet: il costo degli strumenti necessari per connettersi è troppo alto |
| Motivo non uso Internet: il costo del collegamento è troppo alto |
| Motivo non uso Internet: non mi serve/non mi è utile |
| Motivo non uso Internet: non mi piace/non mi interessa |
| Motivo non uso Internet: non so utilizzare Internet/è troppo complicato |
| Motivo non uso Internet: sono diffidente verso le nuove tecnologie |
| Motivo non uso Internet: non mi fido/sono preoccupato per la mia privacy |
| Motivo non uso Internet: ho una disabilità che mi impedisce di utilizzare Internet |
| Motivo non uso Internet: non mi è permesso perché sono minorenne |
| Motivo non uso Internet: altro |
| Quanti libri ha letto negli ultimi 12 mesi? (considerare solo i libri letti per motivi non strettamente scolastici o professionali) |
| Frequenza con cui negli ultimi 12 mesi è andato a cinema |
| Frequenza con cui negli ultimi 12 mesi è andato a teatro |
| Frequenza con cui negli ultimi 12 mesi è andato a musei o mostre |
| Frequenza con cui negli ultimi 12 mesi è andato a concerti di musica classica |
| Frequenza con cui negli ultimi 12 mesi è andato ad altri concerti di musica |
| Frequenza con cui negli ultimi 12 mesi è andato a spettacoli sportivi |
| Frequenza con cui negli ultimi 12 mesi è andato in discoteche, balere, night club o altri luoghi dove ballare |
| Frequenza con cui negli ultimi 12 mesi è andato in siti archeologici, monumenti |
| Possiede carta di credito? |
| Soddisfazione situazione economica negli ultimi 12 mesi |

Tabella A.7 – Stima dei coefficienti delle variabili di input dell'analisi per classi latenti, per gli utenti di Internet

Classe 2 (rispetto alla classe 1, di riferimento)

| | | B | Errore std. | Statistica T | Sig. |
|---------------------------------|-------------------------------|----------------|--------------------|---------------------|-------------|
| Intercetta | | -17,259 | 0,353 | -48,924 | 0,000 |
| Classe di età | 14-24 | -2,741 | 0,294 | -9,334 | 0,000 |
| | 25-34 | -2,614 | 0,274 | -9,530 | 0,000 |
| | 35-44 | -1,910 | 0,268 | -7,124 | 0,000 |
| | 45-54 | -1,477 | 0,267 | -5,538 | 0,000 |
| | 55-64 | -0,720 | 0,260 | -2,764 | 0,006 |
| | 65 o più | 0 ^a | | | |
| Condizione professionale | Occupato | -2,340 | 0,108 | -21,746 | 0,000 |
| | In cerca di nuova occupazione | -2,766 | 0,114 | -24,336 | 0,000 |
| | In cerca di prima occupazione | -3,115 | 0,173 | -17,981 | 0,000 |
| | Casalinga | -1,479 | 0,166 | -8,902 | 0,000 |
| | Studente (15 anni o più) | -2,865 | 0,250 | -11,439 | 0,000 |
| | Ritirato dal lavoro | -1,985 | 0,203 | -9,794 | 0,000 |

| | | B | Errore std. | Statistica T | Sig. |
|--|---|----------------|--------------------|---------------------|-------------|
| | Altra condizione; inabile | -2,709 | 0,254 | -10,652 | 0,000 |
| | Minore di 15 anni | 0 ^a | | | |
| Posizione professionale | Dirigente; autonomo come imprenditore; libero professionista | -1,013 | 0,163 | -6,216 | 0,000 |
| | Direttivo, quadro; impiegato | -0,760 | 0,140 | -5,413 | 0,000 |
| | Capo operaio, operaio subalterno e assimilati; apprendista; lavorante a domicilio per conto d'impresa | -0,281 | 0,138 | -2,033 | 0,042 |
| | Lavoratore in proprio; socio cooperativa Produzione Beni e/o prestazioni di servizio; coadiuvante; collaborazione coordinata e continuativa (con o senza progetto); prestazione d'opera occasionale | -0,579 | 0,149 | -3,901 | 0,000 |
| | Minore di 15 anni | 0 ^a | | | |
| In formazione | | -1,267 | 0,215 | -5,888 | 0,000 |
| Titolo di studio | Laurea | 22,263 | 0,138 | 161,460 | 0,000 |
| | Diploma | 23,317 | 0,129 | 181,153 | 0,000 |
| | Licenza media o abilitazione professionale | 24,361 | 0,134 | 181,696 | 0,000 |
| | Licenza elementare | 25,218 | 0,332 | 76,027 | 0,000 |
| | Nessun titolo | 0 ^a | | | |
| Giudizio sulle risorse econ. della famiglia | Ottime | 0,449 | 0,307 | 1,463 | 0,143 |
| | Adeguate | 0,577 | 0,311 | 1,854 | 0,064 |
| | Scarse | 0,980 | 0,331 | 2,961 | 0,003 |
| | Assolutamente insufficienti | 0 ^a | | | |
| Ripartizione territoriale | Nord-Ovest | -0,298 | 0,110 | -2,714 | 0,007 |
| | Nord-Est | -0,305 | 0,109 | -2,791 | 0,005 |
| | Centro | -0,428 | 0,110 | -3,897 | 0,000 |
| | Sud | 0,233 | 0,104 | 2,236 | 0,025 |
| | Isole | | | | |
| Dimensione comune di residenza | Area metropolitana | -0,298 | 0,077 | -3,889 | 0,000 |
| | Comuni aventi fino a 10.000 abitanti | 0,051 | 0,065 | 0,788 | 0,431 |
| | Comuni con oltre 10.000 abitanti | 0 ^a | | | |

a. Questo parametro è impostato su zero perché è ridondante.

Classe 3 (rispetto alla classe 1, di riferimento)

| | | B | Errore std. | Statistica T | Sig. |
|----------------------|-------|----------|--------------------|---------------------|-------------|
| Intercetta | | 3,582 | 1,039 | 3,449 | 0,001 |
| Classe di età | 14-24 | -1,716 | 0,307 | -5,587 | 0,000 |
| | 25-34 | -1,817 | 0,291 | -6,239 | 0,000 |
| | 35-44 | -1,183 | 0,285 | -4,148 | 0,000 |

| | | B | Errore std. | Statistica T | Sig. |
|---|---|----------------|-------------|--------------|-------|
| | 45-54 | -0,966 | 0,283 | -3,408 | 0,001 |
| | 55-64 | -0,500 | 0,275 | -1,822 | 0,068 |
| | 65 o più | 0 ^a | | | |
| Condizione professionale | Occupato | 0,334 | 0,180 | 1,855 | 0,064 |
| | In cerca di nuova occupazione | 0,102 | 0,184 | 0,557 | 0,578 |
| | In cerca di prima occupazione | 0,231 | 0,218 | 1,057 | 0,290 |
| | Casalinga | 1,340 | 0,209 | 6,397 | 0,000 |
| | Studente (15 anni o più) | 0,098 | 0,256 | 0,384 | 0,701 |
| | Ritirato dal lavoro | 1,255 | 0,251 | 4,990 | 0,000 |
| | Altra condizione; inabile | 0,223 | 0,300 | 0,743 | 0,458 |
| | Minore di 15 anni | 0 ^a | | | |
| Posizione professionale | Dirigente; autonomo come imprenditore; libero professionista | -1,176 | 0,173 | -6,804 | 0,000 |
| | Direttivo, quadro; impiegato | -0,601 | 0,143 | -4,215 | 0,000 |
| | Capo operaio, operaio subalterno e assimilati; apprendista; lavorante a domicilio per conto d'impresa | 0,063 | 0,139 | 0,456 | 0,649 |
| | Lavoratore in proprio; socio cooperativa Produzione Beni e/o prestazioni di servizio; coadiuvante; collaborazione coordinata e continuativa (con o senza progetto); prestazione d'opera occasionale | -0,524 | 0,150 | -3,492 | 0,000 |
| | Minore di 15 anni | 0 ^a | | | |
| In formazione | | -0,729 | 0,179 | -4,064 | 0,000 |
| Titolo di studio | Laurea | -2,446 | 1,100 | -2,224 | 0,026 |
| | Diploma | -1,856 | 1,097 | -1,692 | 0,091 |
| | Licenza media o abilitazione professionale | -1,030 | 1,096 | -0,939 | 0,348 |
| | Licenza elementare | -0,632 | 1,173 | -0,538 | 0,590 |
| | Nessun titolo | 0 ^a | | | |
| Giudizio sulle risorse econ. della famiglia | Ottime | 0,279 | 0,300 | 0,930 | 0,352 |
| | Adeguate | 0,441 | 0,304 | 1,451 | 0,147 |
| | Scarse | 0,733 | 0,324 | 2,261 | 0,024 |
| | Absolutamente insufficienti | 0 ^a | | | |
| Ripartizione territoriale | Nord-Ovest | -0,222 | 0,113 | -1,971 | 0,049 |
| | Nord-Est | -0,116 | 0,112 | -1,038 | 0,299 |
| | Centro | -0,272 | 0,112 | -2,421 | 0,015 |
| | Sud | 0,079 | 0,109 | 0,730 | 0,466 |
| | Isole | 0 ^a | | | |
| Dimensione | Area metropolitana | -0,139 | 0,078 | -1,780 | 0,075 |

| | | B | Errore std. | Statistica T | Sig. |
|--|--------------------------------------|----------------|--------------------|---------------------|-------------|
| | Comuni aventi fino a 10.000 abitanti | 0,205 | 0,067 | 3,082 | 0,002 |
| | Comuni con oltre 10.000 abitanti | 0 ^a | | | |

a. Questo parametro è impostato su zero perché è ridondante.

Classe 4 (rispetto alla classe 1, di riferimento)

| | | B | Errore std. | Statistica T | Sig. |
|---------------------------------|---|----------------|--------------------|---------------------|-------------|
| | Intercetta | 5,346 | 1,003 | 5,331 | 0,000 |
| Classe di età | 14-24 | -3,460 | 0,299 | -11,570 | 0,000 |
| | 25-34 | -3,213 | 0,275 | -11,689 | 0,000 |
| | 35-44 | -2,203 | 0,267 | -8,258 | 0,000 |
| | 45-54 | -1,707 | 0,265 | -6,438 | 0,000 |
| | 55-64 | -0,963 | 0,258 | -3,741 | 0,000 |
| | 65 o più | 0 ^a | | | |
| Condizione professionale | Occupato | 0,669 | 0,174 | 3,857 | 0,000 |
| | In cerca di nuova occupazione | 0,273 | 0,178 | 1,539 | 0,124 |
| | In cerca di prima occupazione | 0,404 | 0,220 | 1,837 | 0,066 |
| | Casalinga | 1,996 | 0,199 | 10,039 | 0,000 |
| | Studente (15 anni o più) | 0,231 | 0,325 | 0,712 | 0,477 |
| | Ritirato dal lavoro | 1,376 | 0,242 | 5,683 | 0,000 |
| | Altra condizione; inabile | 0,395 | 0,288 | 1,372 | 0,170 |
| | Minore di 15 anni | 0 ^a | | | |
| Posizione professionale | Dirigente; autonomo come imprenditore; libero professionista | -1,437 | 0,173 | -8,289 | 0,000 |
| | Direttivo, quadro; impiegato | -0,846 | 0,143 | -5,921 | 0,000 |
| | Capo operaio, operaio subalterno e assimilati; apprendista; lavorante a domicilio per conto d'impresa | 0,101 | 0,141 | 0,720 | 0,472 |
| | Lavoratore in proprio; socio cooperativa Produzione Beni e/o prestazioni di servizio; coadiuvante; collaborazione coordinata e continuativa (con o senza progetto); prestazione d'opera occasionale | -0,443 | 0,151 | -2,942 | 0,003 |
| | Minore di 15 anni | 0 ^a | | | |
| | In formazione | -0,729 | -1,512 | 0,255 | -5,929 |
| Titolo di studio | Laurea | -3,094 | 1,033 | -2,995 | 0,003 |
| | Diploma | -1,943 | 1,029 | -1,889 | 0,059 |
| | Licenza media o abilitazione professionale | -0,767 | 1,029 | -0,746 | 0,456 |
| | Licenza elementare | 0,122 | 1,101 | 0,111 | 0,912 |
| | Nessun titolo | 0 ^a | | | |

| | | B | Errore std. | Statistica T | Sig. |
|---|--------------------------------------|----------------|-------------|--------------|-------|
| Giudizio sulle risorse econ. della famiglia | Ottime | 0,454 | 0,316 | 1,438 | 0,150 |
| | Adeguate | 0,690 | 0,320 | 2,157 | 0,031 |
| | Scarse | 0,660 | 0,342 | 1,932 | 0,053 |
| | Assolutamente insufficienti | 0 ^a | | | |
| Ripartizione territoriale | Nord-Ovest | -0,430 | 0,111 | -3,858 | 0,000 |
| | Nord-Est | -0,376 | 0,111 | -3,374 | 0,001 |
| | Centro | -0,614 | 0,112 | -5,472 | 0,000 |
| | Sud | 0,093 | 0,107 | 0,869 | 0,385 |
| | Isole | 0 ^a | | | |
| Dimensione comune di residenza | Area metropolitana | -0,244 | 0,078 | -3,113 | 0,002 |
| | Comuni aventi fino a 10.000 abitanti | 0,153 | 0,067 | 2,285 | 0,022 |
| | Comuni con oltre 10.000 abitanti | 0 ^a | | | |

a. Questo parametro è impostato su zero perché è ridondante.

Classe 5 (rispetto alla classe 1, di riferimento)

| | | B | Errore std. | Statistica T | Sig. |
|--------------------------|---|----------------|-------------|--------------|-------|
| Intercetta | | 2,344 | 1,270 | 1,846 | 0,065 |
| Classe di età | 14-24 | -4,081 | 0,326 | -12,519 | 0,000 |
| | 25-34 | -2,598 | 0,268 | -9,695 | 0,000 |
| | 35-44 | -1,446 | 0,260 | -5,556 | 0,000 |
| | 45-54 | -0,962 | 0,259 | -3,710 | 0,000 |
| | 55-64 | -0,478 | 0,255 | -1,874 | 0,061 |
| | 65 o più | 0 ^a | | | |
| Condizione professionale | Occupato | 0,738 | 0,211 | 3,496 | 0,000 |
| | In cerca di nuova occupazione | -0,047 | 0,218 | -0,215 | 0,829 |
| | In cerca di prima occupazione | -1,102 | 0,372 | -2,961 | 0,003 |
| | Casalinga | 0,868 | 0,246 | 3,533 | 0,000 |
| | Studente (15 anni o più) | 0,350 | 0,285 | 1,225 | 0,221 |
| | Ritirato dal lavoro | 1,068 | 0,273 | 3,909 | 0,000 |
| | Altra condizione; inabile | 0,469 | 0,329 | 1,422 | 0,155 |
| | Minore di 15 anni | 0 ^a | | | |
| Posizione professionale | Dirigente; autonomo come imprenditore; libero professionista | -0,349 | 0,171 | -2,049 | 0,040 |
| | Direttivo, quadro; impiegato | -0,200 | 0,157 | -1,277 | 0,202 |
| | Capo operaio, operaio subalterno e assimilati; apprendista; lavorante a domicilio per conto d'impresa | -0,615 | 0,164 | -3,759 | 0,000 |

| | | B | Errore std. | Statistica T | Sig. |
|--|---|----------------|-------------|--------------|--------|
| | Lavoratore in proprio; socio cooperativa Produzione Beni e/o prestazioni di servizio; coadiuvante; collaborazione coordinata e continuativa (con o senza progetto); prestazione d'opera occasionale | -0,395 | 0,164 | -2,405 | 0,016 |
| | Minore di 15 anni | 0 ^a | | | |
| In formazione | | -0,729 | -0,266 | 0,153 | -1,740 |
| Titolo di studio | Laurea | -0,596 | 1,398 | -0,426 | 0,670 |
| | Diploma | -0,387 | 1,396 | -0,277 | 0,781 |
| | Licenza media o abilitazione professionale | -0,282 | 1,397 | -0,202 | 0,840 |
| | Licenza elementare | -0,093 | 1,474 | -0,063 | 0,950 |
| | Nessun titolo | 0 ^a | | | |
| Giudizio sulle risorse econ. della famiglia | Ottime | -0,097 | 0,244 | -0,398 | 0,690 |
| | Adeguate | -0,194 | 0,249 | -0,777 | 0,437 |
| | Scarse | -0,537 | 0,293 | -1,831 | 0,067 |
| | Assolutamente insufficienti | 0 ^a | | | |
| Ripartizione territoriale | Nord-Ovest | 0,124 | 0,113 | 1,096 | 0,273 |
| | Nord-Est | 0,347 | 0,112 | 3,094 | 0,002 |
| | Centro | -0,083 | 0,113 | -0,738 | 0,460 |
| | Sud | -0,094 | 0,115 | -0,823 | 0,411 |
| | Isole | 0 ^a | | | |
| Dimensione comune di residenza | Area metropolitana | -0,099 | 0,072 | -1,369 | 0,171 |
| | Comuni aventi fino a 10.000 abitanti | 0,122 | 0,065 | 1,865 | 0,062 |
| | Comuni con oltre 10.000 abitanti | 0 ^a | | | |

a. Questo parametro è impostato su zero perché è ridondante.

Classe 6 (rispetto alla classe 1, di riferimento)

| | | B | Errore std. | Statistica T | Sig. |
|----------------------|----------|----------------|-------------|--------------|-------|
| Intercetta | | 1,410 | 1,150 | 1,226 | 0,220 |
| Classe di età | 14-24 | 1,215 | 0,510 | 2,380 | 0,017 |
| | 25-34 | 0,649 | 0,505 | 1,284 | 0,199 |
| | 35-44 | 0,474 | 0,503 | 0,943 | 0,346 |
| | 45-54 | 0,092 | 0,505 | 0,181 | 0,856 |
| | 55-64 | -0,102 | 0,478 | -0,213 | 0,832 |
| | 65 o più | 0 ^a | | | |
| Condizio | Occupato | 0,056 | 0,207 | 0,273 | 0,785 |

| | | B | Errore std. | Statistica T | Sig. |
|--|---|----------------|------------------------|-------------------------|-------------|
| | In cerca di nuova occupazione | 0,101 | 0,209 | 0,482 | 0,630 |
| | In cerca di prima occupazione | -0,247 | 0,228 | -1,080 | 0,280 |
| | Casalinga | 0,825 | 0,242 | 3,413 | 0,001 |
| | Studente (15 anni o più) | 0,011 | 0,266 | 0,043 | 0,966 |
| | Ritirato dal lavoro | 0,866 | 0,342 | 2,530 | 0,011 |
| | Altra condizione; inabile | -0,203 | 0,327 | -0,620 | 0,535 |
| | Minore di 15 anni | 0 ^a | | | |
| Posizione professionale | Dirigente; autonomo come imprenditore; libero professionista | -1,644 | 0,204 | -8,044 | 0,000 |
| | Direttivo, quadro; impiegato | -1,042 | 0,133 | -7,865 | 0,000 |
| | Capo operaio, operaio subalterno e assimilati; apprendista; lavorante a domicilio per conto d'impresa | -0,073 | 0,122 | -0,600 | 0,548 |
| | Lavoratore in proprio; socio cooperativa Produzione Beni e/o prestazioni di servizio; coadiuvante; collaborazione coordinata e continuativa (con o senza progetto); prestazione d'opera occasionale | -0,656 | 0,135 | -4,863 | 0,000 |
| | Minore di 15 anni | 0 ^a | | | |
| | In formazione | -0,729 | -0,840 | 0,175 | -4,804 |
| Titolo di studio | Laurea | -1,871 | 1,193 | -1,569 | 0,117 |
| | Diploma | -0,980 | 1,190 | -0,824 | 0,410 |
| | Licenza media o abilitazione professionale | -0,050 | 1,190 | -0,042 | 0,966 |
| | Licenza elementare | 0,530 | 1,265 | 0,419 | 0,675 |
| | Nessun titolo | 0 ^a | | | |
| Giudizio sulle risorse econ. della famiglia | Ottime | 0,297 | 0,320 | 0,929 | 0,353 |
| | Adeguate | 0,470 | 0,323 | 1,453 | 0,146 |
| | Scarse | 0,657 | 0,341 | 1,924 | 0,054 |
| | Assolutamente insufficienti | 0 ^a | | | |
| Ripartizione territoriale | Nord-Ovest | -0,337 | 0,111 | -3,043 | 0,002 |
| | Nord-Est | -0,424 | 0,111 | -3,808 | 0,000 |
| | Centro | -0,419 | 0,111 | -3,788 | 0,000 |
| | Sud | 0,127 | 0,103 | 1,232 | 0,218 |
| | Isole | 0 ^a | | | |
| Dimensione comune di residenza | Area metropolitana | -0,071 | 0,079 | -0,906 | 0,365 |
| | Comuni aventi fino a 10.000 abitanti | 0,019 | 0,067 | 0,278 | 0,781 |
| | Comuni con oltre 10.000 abitanti | 0 ^a | | | |

a. Questo parametro è impostato su zero perché è ridondante.

Tabella A.8 - Stima dei coefficienti delle variabili di input dell'analisi per classi latenti, per i non utenti

Classe 2 (rispetto alla classe 1, di riferimento)

| | | B | Errore std. | Statistica T | Sig. |
|--------------------------------------|---|----------------|-------------|--------------|-------|
| Intercetta | | 3,188 | 0,218 | 14,622 | 0,000 |
| Classe di età | 14-24 | -2,000 | 0,304 | -6,585 | 0,000 |
| | 25-34 | -0,026 | 0,204 | -0,126 | 0,899 |
| | 35-44 | 0,138 | 0,134 | 1,028 | 0,304 |
| | 45-54 | 0,155 | 0,112 | 1,376 | 0,169 |
| | 55-64 | 0,110 | 0,086 | 1,286 | 0,198 |
| | 65 o più | 0 ^a | | | |
| Condizione e posizione professionale | Occupato <i>white collar</i> | -1,316 | 0,193 | -6,811 | 0,000 |
| | Occupato <i>blue collar</i> | -0,432 | 0,181 | -2,391 | 0,017 |
| | Pensionato <i>white collar</i> | -0,434 | 0,185 | -2,340 | 0,019 |
| | Pensionato <i>blue collar</i> | -1,177 | 0,189 | -6,222 | 0,000 |
| | In cerca di nuova occupazione | -0,032 | 0,211 | -0,149 | 0,881 |
| | In cerca di prima occupazione | -0,004 | 0,347 | -0,011 | 0,991 |
| | Casalinga | -0,409 | 0,177 | -2,307 | 0,021 |
| | Studente (14 anni o più) | 0 ^a | | | |
| Titolo di studio e formazione | Diploma (non in formazione) o licenza media (in formazione) | -1,841 | 0,102 | -18,130 | 0,000 |
| | Laurea (in formazione o non) o diploma (in formazione) | -2,963 | 0,161 | -18,421 | 0,000 |
| | Licenza media (non in formazione) | -0,940 | 0,092 | -10,183 | 0,000 |
| | Licenza elementare o meno | 0 ^a | | | |
| Ripartizione territoriale | Nord-Ovest | -0,631 | 0,126 | -5,029 | 0,000 |
| | Nord-Est | -0,997 | 0,126 | -7,921 | 0,000 |
| | Centro | -0,348 | 0,130 | -2,666 | 0,008 |
| | Sud | 0,222 | 0,122 | 1,817 | 0,069 |
| | Isole | 0 ^a | | | |
| Dimensione comune di residenza | Area metropolitana | -0,179 | 0,088 | -2,033 | 0,042 |
| | Comuni aventi fino a 10.000 abitanti | 0,003 | 0,069 | 0,042 | 0,966 |
| | Comuni con oltre 10.000 abitanti | 0 ^a | | | |

a. Questo parametro è impostato su zero perché è ridondante.

Classe 3 (rispetto alla classe 1, di riferimento)

| | | B | Errore std. | Statistica T | Sig. |
|---------------|-------|--------|-------------|--------------|-------|
| Intercetta | | -0,252 | 0,322 | -0,781 | 0,435 |
| Classe di età | 14-24 | 2,209 | 0,288 | 7,667 | 0,000 |

| | | B | Errore std. | Statistica T | Sig. |
|--------------------------------------|---|----------------|-------------|--------------|-------|
| | 25-34 | 3,020 | 0,261 | 11,550 | 0,000 |
| | 35-44 | 2,168 | 0,222 | 9,748 | 0,000 |
| | 45-54 | 1,347 | 0,215 | 6,279 | 0,000 |
| | 55-64 | 0,686 | 0,191 | 3,595 | 0,000 |
| | 65 o più | 0 ^a | | | |
| Condizione e posizione professionale | Occupato <i>white collar</i> | -0,220 | 0,264 | -0,835 | 0,404 |
| | Occupato <i>blue collar</i> | 1,067 | 0,233 | 4,589 | 0,000 |
| | Pensionato <i>white collar</i> | 0,200 | 0,291 | 0,687 | 0,492 |
| | Pensionato <i>blue collar</i> | -0,641 | 0,366 | -1,750 | 0,080 |
| | In cerca di nuova occupazione | 1,733 | 0,250 | 6,928 | 0,000 |
| | In cerca di prima occupazione | 1,286 | 0,338 | 3,806 | 0,000 |
| | Casalinga | -0,066 | 0,244 | -0,272 | 0,786 |
| | Studente (14 anni o più) | 0 ^a | | | |
| Titolo di studio e formazione | Diploma (non in formazione) o licenza media (in formazione) | -1,186 | 0,163 | -7,253 | 0,000 |
| | Laurea (in formazione o non) o diploma (in formazione) | -2,362 | 0,251 | -9,400 | 0,000 |
| | Licenza media (non in formazione) | -0,469 | 0,147 | -3,193 | 0,001 |
| | Licenza elementare o meno | 0 ^a | | | |
| Ripartizione territoriale | Nord-Ovest | -1,115 | 0,178 | -6,280 | 0,000 |
| | Nord-Est | -1,581 | 0,184 | -8,607 | 0,000 |
| | Centro | -0,704 | 0,174 | -4,036 | 0,000 |
| | Sud | -0,111 | 0,153 | -0,725 | 0,468 |
| | Isole | 0 ^a | | | |
| Dimensione comune di residenza | Area metropolitana | 0,167 | 0,125 | 1,340 | 0,180 |
| | Comuni aventi fino a 10.000 abitanti | -0,361 | 0,108 | -3,347 | 0,001 |
| | Comuni con oltre 10.000 abitanti | 0 ^a | | | |

a. Questo parametro è impostato su zero perché è ridondante.